FloraBERT: cross-species transfer learning withattention-based neural networks for geneexpression prediction

Benjamin Levy
Harvard University
Zihao Xu
Harvard University
Liyang Zhao
Harvard University
Karl Kremling
Inari Agriculture (United States)
Ross Altman
Harvard University
Phoebe Wong
Harvard University
Chris Tanner (✉ cwt@mit.edu)
Harvard University

Article

Keywords:

Posted Date: August 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1927200/v1

FloraBERT: cross-species transfer learning with attention-based neural networks for gene expression prediction

Benjamin Levy[1], Zihao Xu[1], Shuying Ni[1], Liyang Zhao[1], Karl Kremling[2], Ross Altman[2], Phoebe Wong[1], and Chris Tanner[1,*]

[1] Institute for Applied Computational Sciences, Harvard University, Cambridge, MA
[2] Inari Agriculture Inc., Cambridge, MA
[*] cwt@mit.edu

ABSTRACT

Recent work in applying deep learning models has demonstrated that endophenotypes, such as RNA transcript abundance, can be predicted from an organism's regulatory DNA. However, due to the vast amount of labelled data required to train previous types of deep learning models, this work has been constrained to species with large amounts of data labelled for a particular task. Here, we present FloraBERT, a transfer-learning based deep learning model that is able to improve predictions of gene expression in a single target species, and it does so by exploiting cross-species genomic information in the form of genome assemblies from all of *plantae*. FloraBERT significantly outperforms simple bag-of-k-*mers* baseline models and achieves comparable performance to prior work that concerns less complex species. Furthermore, investigation of the learned parameters of FloraBERT reveals that the training process encodes biologically salient information, such as taxonomic similarity between species and positional relevance of nucleotides within a promoter. To facilitate future research, we have made the source code and model weights publicly available on GitHub at `https://github.com/benlevyx/florabert`.

Introduction

Understanding associations between phenotype and genetic sequences is critical for advancing research of both agricultural species and genetics-based medicine. Additionally, with the advent of targeted genetic modification techniques like CRISPR-Cas9-based genome editing, knowledge of causal effects, rather than linked-variant effects, is of increasing practical use. Inferring which variant effects are causal stands to benefit from the application of deep learning methods[1–4] borrowed from other fields such as natural language processing (NLP), speech recognition, and image processing. This benefit is especially clear with transfer learning models from NLP and image processing, as these models are able to train on vast repositories of unlabeled sequences from multiple species. After training on these tasks, the models can be trained further to make predictions on narrower tasks for which only sparsely annotated data is available. Recently, this has been demonstrated in biology for protein structure and function prediction[5–7].

While performing genomic prediction of phenotypes using linked single nucleotide polymorphisms (SNP) or insertion/deletion (indel) markers is now standard in plant and livestock breeding programs, these genomic prediction or polygenic risk score models still lack high degrees of generalizability. For example, they are unable to exploit multiple reference genomes within or across species, and predictions are typically made using linked rather than causal loci. Furthermore, genomic prediction models are particularly sensitive to differences between the training data and testing data, which is an active area of development[8]. For example, one approach is to selectively incorporate putatively functional variants or features that are more likely to be causal, rather than including features that are based on biological domain knowledge or just physically linked. However, improvements to prediction accuracy from these approaches have been limited[9, 10]. Additionally, predictions using SNP and small indel markers cannot capture large structural variations, nor can they readily exploit the increasing number of available reference genomes and annotations in one or more species[11–13]. These shortcomings afford the opportunity to better exploit genetic sequence data both within and across species, and to do so by using models that extend beyond making predictions from SNPs called against a single reference genome.

Recent techniques derived from deep learning, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been used to make predictions of endophenotypes directly from underlying genome sequences, instead of relying on specific markers such as SNPs and indels[14–16]. However, these techniques currently rely on having large amounts of task-specific labeled training data. This is highly limiting because such labeled data is scarce and is typically only available for a small fraction of species in a kingdom. Therefore, there exists a need for computational techniques which can learn from increasingly abundant reference genomes and structural diversity within and between multiple species, and to do so without requiring alignment. Such techniques would enable researchers to make endophenotypic predictions beyond variants called against a single reference genome.

The application of models that can leverage information from multiple reference genomes within and across species is especially important in *planta* relative to other taxonomic groups. For example, while the structural diversity of plant genomes can prevent functional relationships between lines from being accurately captured by SNPs and indels alone, the relationships can be captured across reference genomes. As well, in the case of deep learning models, these rearrangements can serve as a form of evolutionary data augmentation during training. These structural variants are often functionally significant, especially in gene regulatory contexts[12, 13]. Further, the sample sizes represented in genotypic datasets from individual plant species are more limited than in human datasets[17], meaning that genomic prediction and polygenic risk score calculation is less accurate. Although plant datasets that concern within-species information are more limiting in their sample sizes, the abundance of species in the kingdom can be exploited to make predictions. This is increasingly possible due to the number of structurally diverse genomes that have been assembled, including those from closely-related species. This stands in contrast to the number of observations within a species. Therefore, our proposed model is of greatest relative utility in species for which data are limiting.

In this paper, we propose and implement a model, FloraBERT, that implicitly learns regulatory motifs from promoters in nearly every publicly available sequenced plant genome. This information can then be transferred to predict an endophenotype, such as gene expression, in any particular species of interest. Instead of being constrained to markers called against a single reference genome in one species, our model enables prediction of phenotype from genotype while simultaneously leveraging nearly all publicly available plant reference genomes. This approach is better suited to the pangenome era than one in which training data is limited to SNPs called against a single reference genome.

Specifically, FloraBERT is a BERT-based[18, 19] transformer neural network built on the concept of *self-attention*, which is derived from the fields of natural language processing and machine learning. This model is pre-trained using a corpus of reference sequences and annotations from 93 plant species. Using the transfer learning paradigm, the model is then adapted for a secondary task of predicting gene expression associated with gene-proximal sequences from twenty-five recently released annotated maize reference genomes. We draw inspiration from DNABERT, a transformer model trained on a huge corpus of human genomic sequences[16]. However, unlike DNABERT, we use multiple species in the pre-training step due to smaller available amounts of genomic data in any single plant species.

We find that FloraBERT consistently outperforms simpler bag-of-words (BoW) benchmark models on the task of predicting gene expression from upstream promoter sequence across 9 distinct Zea mays tissues. Furthermore, we find that FloraBERT performs comparably to a CNN-based model trained by Zrimec et al.[15] on predicting gene expression from upstream promoter sequence in a simpler plant species (Arabidopsis thaliana).

Methods

Data collection

We acquired plant DNA sequences from three databases: Ensembl Plants[20], RefSeq[21], and MaizeGDB[22]. The first two databases contain whole genome assemblies for several hundred species of plants, while the third database is specific to maize. We then extracted a 1kb sequence immediately upstream from the transcription start site (TSS) of each gene included in the GFF annotation for that reference genome (see Supplementary Information for detailed methods). Descriptive statistics of the sequences are shown in Supplementary Table **??**.

For the downstream gene expression prediction task, we obtained previously published[13] pairs of promoter sequences and expression levels across 9 different maize tissues of their proximal genes. Gene expression values for B73 and each of the Nested Association Mapping (NAM) founder lines were calculated from reads accessed on the NCBI Sequence Read Archive (details in Supplementary Information). The distribution of gene expression values (expressed in transcripts per million, or TPM) is highly right-skewed. Thus, we transform the expression levels using the natural logarithm with a small offset of 0.001 to avoid taking the logarithm of 0. The log-transformed gene expression level was then used as the target variable for both training and evaluating the model.

Train-test split

Due to shared evolutionary history, the promoter sequences of many genes in the dataset were extremely similar. This phenomenon – known as paralogy – is a common outcome of genome duplications. Failing to take paralogy into account when preparing a held-out test dataset could give an overly optimistic picture of model performance, since paralogous sequences could appear in both training and testing sets. To prevent the model from "memorizing" specific genes' expression patterns based on common regulatory sequences, each group of paralogs was placed either all in the training set or all in the testing set.

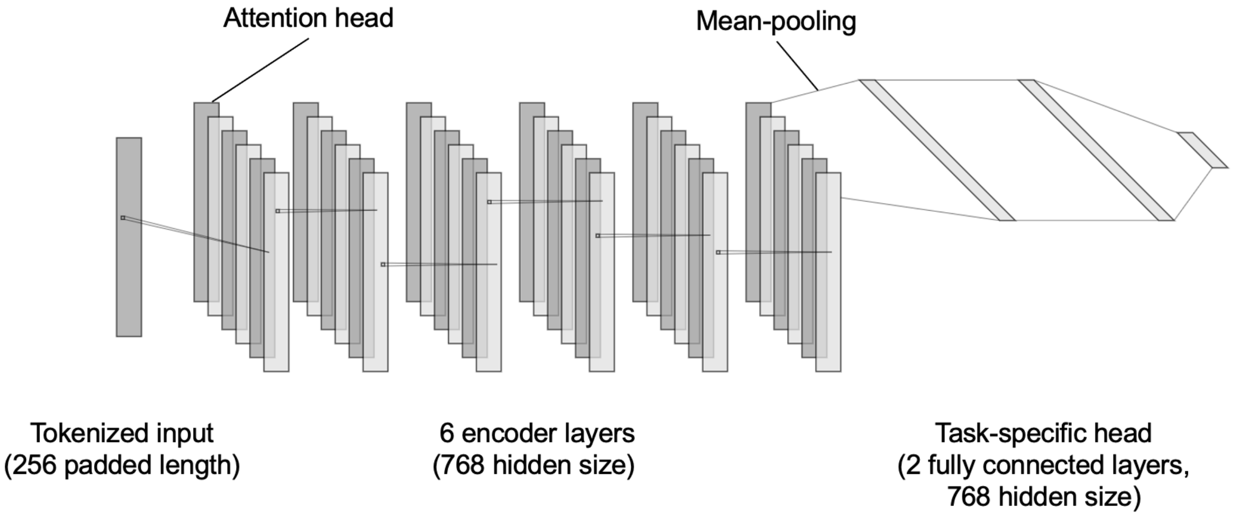

Figure 1. Schematic of pre-trained FloraBERT with fine-tuning head. The tokenized input is fed into 6 successive transformer encoder layers, each of which is comprised of 6 independent self-attention heads. During fine-tuning, the final encoder layer's embedding vectors for all tokens (not including padding, start, or end tokens) are element-wise averaged and the resulting sequence-level embedding vector is passed into a feed-forward neural network, which predicts a gene expression value for each of the 9 tissues.[29]

With this constraint in mind, 70% of the available sequences were allocated to the training set and the remaining 30% to the testing set. Paralogy information was determined from a table compiled by Schnable et al.[23–27]. For the 25 NAM genomes that did not have available paralogy information, each gene was cross-referenced to its respective ortholog in the reference B73 genome. The paralogy information for each NAM genome was then inferred from the cross-referenced B73 genes. Similar to the pre-training dataset (containing all *plantae*), paralogous genes were kept together in either the train or test sets.

Model architecture

Transformers are a modern architecture[18] that have yielded state-of-the-art results in nearly every major Natural Language Processing (NLP) task. Technically, there are Transformer Encoders and Transformer Decoders. The objective of the *Transformer Encoder* is to simply learn a rich, condensed representations (i.e., *embeddings*, which are just vectors of floating-point values) for each distinct input token (e.g., each word). These learned embeddings are context-sensitive, meaning the order of the inputs matters. The objective of *Transformer Decoders* is to generate new sequences of tokens (e.g., words). Colloquially, *Transformer Encoders* are simply referred to as "Transformers", as they are used more frequently than Decoders.

Transformers learn rich representations largely due to multiple, successive *self-attention heads*, which are mechanisms that determine, for each input token, how much focus/attention to place on every *other* token within the provided input sequence. This procedure informs the learned representations, and successive layers of such are intended to yield increasingly rich, abstract representations (i.e., embeddings). These embeddings are then useful or downstream tasks. There are many slight variants of the Transformers model, depending on the exact architecture and data that was used to pre-train it (e.g., BERT[19], RoBERTa[28]).

The FloraBERT architecture is based on the the RoBERTa transformer model. Briefly, it consists of an input-embedding layer, 6 encoder layers, and finally a task-specific layer (i.e., "head"), the structure of which changes based on whether the model is being pre-trained or fine-tuned. A schematic of the model is displayed in Figure 1.

The embedding layer's input is a padded sequence of at most 256 tokens. The inputs originate from the 1kb promoter sequences and are transformed into byte-pair encodings (BPE) – a technique from natural language processing whereby commonly occurring motifs of progressively longer strings of nucleotides are discovered through an iterative process and then grouped into discrete tokens[30]. We fix the vocabulary size at 5,000 and run the tokenization algorithm until 5,000 such motifs are identified. Empirically, we found that after processing with BPE and adding start and end tokens (`<s>` and `</s>`, respectively), no sequence exceeded 256 tokens. To reflect the fact that sequences were obtained by starting at the TSS and proceeding in the 5' direction along the DNA strand, special padding tokens (`<pad>`) are added on the left (5') side to ensure that each sequence is 256 tokens long. This also ensures that all sequences are right-aligned at the TSS.

Each unique token in the 256-length input is mapped to a 768-dimensional embedding vector, referred to here as a token embedding. Separately, each of the 256 positions in the sequence is also mapped to a 768-dimensional embedding, referred to as a positional embedding. The token embeddings and positional embeddings are summed element-wise (since both are of equal dimensionality) and the output vectors are assembled into a matrix of size 256×768.

The transformer contains 6 encoder layers, each of which consist of 6 "heads" of the same self-attention mechanism, followed by a concatenation of the attention heads and 2 subsequent feed-forward neural network layers. The self-attention and feed-forward layers are connected via residual (skip) connections, as in the original implementation of the transformer[18].

The final output of the encoder layers is a 768-dimensional vector for each input (a matrix of size 256×768). This vector is then used as the input to a task-specific head block. For the pre-training step, we use the masked language modeling task[19]: 15% of input tokens are replaced either with one of: a special `<mask>` token (80% of selected tokens), a randomly chosen token (10% of selected tokens), or left unchanged (10% of selected tokens). The outputs of the model are then fed into a feed-forward neural network and the final outputs for the masked positions are taken to be the log-probabilities for each of the possible tokens.

The downstream task is to predict the level of gene expression in 9 distinct maize tissues: endosperm, tassel inflorescence, leaf base, anther, leaf, ear inflorescence, shoot, root, and leaf tip. Therefore, we formulate it as a multi-label regression problem, where the input is the tokenized gene promoter sequence and the output is a 9-dimensional vector with gene expression values in each tissue. The task-specific head consists of two feed-forward neural network layers, with an intervening layer that has a dropout probability of 0.2. Often, for classification or regression tasks, the final hidden state of the start token (`<s>`) is taken to be a representation of the entire sequence, and so that token alone is fed into subsequent layers[31]. However, we found that model performance on the downstream gene expression prediction task was poor when using this token alone as the representation of the sequence. Instead, we average the final hidden representation of all tokens in the sequence, excluding padding, start, and end tokens. This results in a single 768-dimensional vector, which can then be used to predict gene expression simultaneously in all 9 tissues. We hypothesize that simultaneously predicting all the tissues at once has an implicit regularization effect on the model parameters, penalizing activations that overfit to a particular tissue.

Pre-training

Beginning with randomly initialized weights, we train the language model in a self-supervised fashion on all plant promoter sequences aggregated from the RefSeq[21] and Ensembl[20] databases (approx. 7.9 M). Since the downstream task of predicting gene expression is specific to maize, we perform an additional round of pre-training on only maize promoter sequences obtained from the MaizeGDB[22] database. This is to encourage the model to prioritize salient features for maize-specific promoters. We use the LAMB optimizer[32], a layer-wise optimization algorithm designed for the fast convergence of transformer models. For the LAMB optimizer, we selected the hyper-parameters $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$. The learning rate is linearly warmed up over 200 iterations to a maximum learning rate of 10^{-3} and then linearly decayed to 0 over the course of training for a total of 3 epochs. Training was done using a batch size of 84 promoter sequences on 4 NVIDIA® Tesla V100 Tensor Core GPUs for $\sim$ 24 hours for both pre-training runs.

Fine-tuning

For task-specific fine-tuning we employ the same strategies as in the pre-training phase (LAMB optimizer, linear warm-up) with the exception that the learning rate is not linearly cooled down, but held constant at a lower maximum learning rate of 5^{-5} for a total of 25 epochs. Fine-tuning took roughly 6 hours on 1 NVIDIA® Tesla V100 Tensor Core GPU.

All transformer models were built in Python v3.9 using the HuggingFace[33] Transformers library (v3.4.0) and PyTorch (v1.7.1). Code is available on GitHub at the following link: https://github.com/benlevyx/florabert.

Baseline models

We developed two baseline models for comparison:

1. *Constant mean*: The average gene expression levels for each tissue within the training data are used as the prediction for all the test sequences. Results for this model were not shown because the R^2 for constant predictions is undefined.
2. *Bag-of-kmers*: Raw sequences are first converted into a bag-of-words representation of k-mers (lengths 2-5 were tried) and then fed to a Linear Regression model to predict the expression levels.

Results

FloraBERT's pre-trained weights can distinguish between plant types

We used principal component analysis (PCA) to determine whether the pre-trained weights of FloraBERT contained information that could be used to distinguish and relate different species based on their evolutionary similarity. Although we observed some clustering of species within certain genera, as well as a clear split between monocots and eudicots, these results were not unique to the embeddings from FloraBERT; similar clustering was observed with a simple 1-mer embedding model (see Supplementary Information for details).

Positional importance of motifs in promoter for model predictions

To investigate the most important features involved in the model's predictions after performing transfer learning for the gene expression level prediction task, we first conducted a positional importance analysis by randomly substituting k-mers of nucleotides of varying lengths at different positions along the sequence (e.g., changing AATG to CACC). Then, we observed how much the predicted expression levels changed. Positional importance was calculated as:

$$\text{mean}\left(\left|\log \text{TPM}' - \log \text{TPM}\right|\right)$$

Figure 2 shows that the nucleotide positions with the greatest influence on model predictions are those much closer to the transcription start site (TSS). Nucleotides within 100 nt upstream of the TSS had the greatest influence, which significantly attenuated as the distance from the TSS increased. Substituting larger blocks of nucleotides (5-mers and 6-mers) led to much more noticeable increases in the mean absolute difference of model predictions (measured in log-TPM).

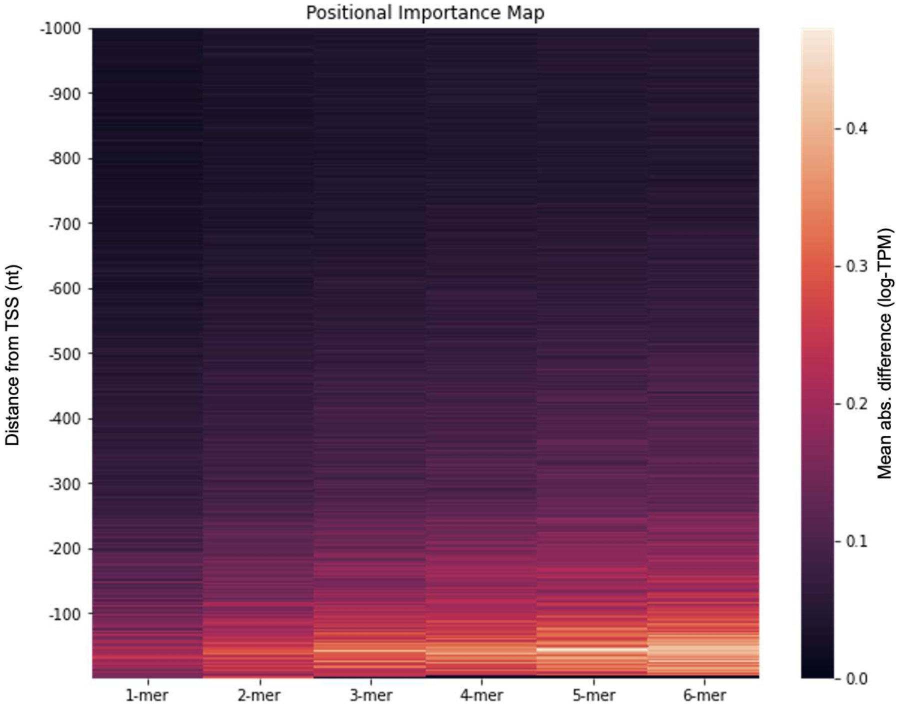

Figure 2. Permutation-based positional importance, averaged over all tissues. Band brightness is proportional to the mean absolute difference (in log-TPM) between the original model predictions and the predicted gene expression levels after a k-mer is randomly permuted at the given location upstream of the transcription start site (TSS).

Predicting mRNA expression levels from promoter sequences

Table 1 shows the predictive performance of the fine-tuned model for predicting mRNA log-transcripts per million (log-TPM) compared to several baseline models. Within each individual tissue, FloraBERT outperforms the Bag of Words (BoW) models considerably, with the highest overall performance on prediction of ear inflorescence (19.2% R^2). When R^2 is calculated simultaneously for all tissues, BoW models have a negligible R^2 of close to 0%, whereas FloraBERT is able to achieve an R^2 of 15.3%.

Discussion

In this paper, we demonstrate the utility of FloraBERT – a transformer-based model adapted from the field of natural language processing – for endophenotypic prediction tasks in plants. FloraBERT is distinct from previous plant-specific models due to its use of transfer learning to aggregate gene regulatory knowledge from across all assembled and annotated plant genomes for

Table 1. R^2 scores for FloraBERT and baseline models; overall and disaggregated by tissue.

Tissue	All	Anther	Ear Infl.	Endosp.	Leaf	Leaf Base	Leaf Tip	Root	Shoot	Tassel Infl.
Model name										
BOW 2mer	0.000	0.026	0.031	0.036	0.030	0.024	0.038	0.021	0.021	0.029
BOW 3mer	0.000	0.037	0.054	0.054	0.043	0.038	0.052	0.034	0.037	0.042
BOW 4mer	0.000	0.048	0.085	0.085	0.062	0.060	0.070	0.055	0.056	0.059
BOW 5mer	0.000	0.056	0.116	0.117	0.077	0.081	0.086	0.075	0.075	0.075
FloraBERT	**0.153**	**0.092**	**0.192**	**0.169**	**0.129**	**0.137**	**0.136**	**0.101**	**0.126**	**0.123**

enhanced predictive power on more targeted tasks within individual species. During pre-training, embeddings that capture the underlying structure of plant regulatory sequence are learned and can be applied to downstream tasks including prediction of gene expression from promoter sequence, as we have demonstrated here. This is of particular interest to those seeking to design regulatory sequences that give rise to novel expression patterns and altered phenotypes in plants[34, 35].

Pre-trained weights from multiple plant species can be used for a species-specific downstream task

FloraBERT represents a significant innovation by first training across all assembled genomes in the kingdom *plantae* prior to performing species-specific phenotypic prediction via transfer learning. This means that the phylogenetically-independent evolutionary trajectories taken by regulatory sequences across species can be implicitly characterized by a model during training, before being applied to specific tasks within a given species. Regulatory elements contain more diversity across species than within species. Given the fact that conserved non-coding sequences are often shared across species, it is also reasonable to expect that rules of regulatory sequences are partially shared as well – and thus, this kind of cross-species learning may boost the signal of subtle regulatory patterns that would otherwise go undetected in a single-species model[36]. In addition to being detectable in naturally occurring sequences, the applicability of regulatory sequences between species is demonstrated experimentally by the frequent reuse of promoters from one species of plant in other species for genetic engineering[37].

While cross-species training may be applicable across life, we propose this aspect of FloraBERT is of particular practical use when seeking to predict regulation of gene expression in plants. Because each individual of a plant species is less agriculturally valuable than individual animals, species in *plantae* have fewer sequenced, assembled, and annotated individuals than more data-rich animal species including humans. This is demonstrated by the fact that, in maize, only a single reference genome was publicly available until the release of the draft PH207 assembly, which supplements the B73 assembly from 2016[38]. Thus, because there is less representation of assembled individuals within a species, predictions stand to benefit more from borrowing information across species in *plantae*. Furthermore, the abundant rearrangement and shuffling of genetic elements by transposons can be exploited as a form of data augmentation to train models. This is similar to the practice of computationally permuting or shifting the orientation of an image before training a deep learning model[39]; in both instances, a model is able to recognize and leverage common patterns that show up in different contexts. In the case of genetic transposition, an element may appear across or within genomes with alterations such as inversions, insertions, or deletions.

Our results and accuracy are most comparable to the research by Zrimec et al.,[15], who used a convolutional neural network (CNN) to predict gene expression levels based on gene promoters in various species. However, our methods represent an improvement to the CNN-based approaches, as they cannot easily exploit unlabelled assembled genomes across species during the training step and are thus limited to training and prediction within a single species. When testing their model on the promoters of *Arabidopsis thaliana* (thale cress), the authors achieved similar performance to the work described here ($R^2 = 27\%$). However, those results were obtained on a plant species whose genome is simpler than that of Maize.

Similar to the work performed by Zrimec et al., Washburn et al.[14] used a CNN to predict the more- and less-expressed copy of orthologous gene pairs in *Zea mays*, using both upstream and downstream gene-proximal sequences. While not seeking to predict expression quantitatively from *cis*-sequences, their model's ability to predict which copy of a gene in the genome is more highly expressed indicates that it can infer the consequences of upstream regulatory sequence for gene expression. However, as with Zrimec et al. mentioned above, this model stands in contrast to FloraBERT in that it cannot benefit from cross-species learning.

Evolutionary and genomic information is encoded in pre-trained and fine-tuned model weights

Previous work using NLP-inspired models applied to biological sequences have demonstrated a capacity to learn semantically-relevant features. For instance, Vig et al.[40] demonstrated that when a BERT-based Transformer model is trained on protein sequences, the self-attention matrices can capture meaningful information about three-dimensional protein structure. Likewise, Rives et al.[41] showed the capacity of similar models to directly predict protein function. Through our analysis of positional importance, we have shown that FloraBERT places the most weight on nucleotides that are closest to the TSS, aligning with prior biological knowledge that the most significant promoter bases tend to be closer to the TSS. Additionally, by visualizing the embedding space of FloraBERT's pre-trained weights (e.g., Figure **??**), we demonstrated that the geometry of the high-dimensional embedding space aligned roughly with intuitions about species similarity based on known evolutionary relationships. However, further analysis showed that this geometry largely reflected the relative frequencies of nucleotides in the promoter sequences, since the visualized embeddings appeared to show similar clustering in a simple 1-mer model.

In short, FloraBERT represents a novel method to combine information from genome assemblies across species to detect and interpret patterns in regulatory sequence and predict endophenotypic annotations. While the number of assemblies within species and across kingdoms are increasing[12, 13], it is not always possible to create alignments between the assemblies, which can lead to a loss of valuable information for modeling. Alignment-free methods such as FloraBERT can provide a path to exploit and learn from otherwise untapped genetic diversity.

References

1. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.* **36**, 442–455, DOI: https://doi.org/10.1016/j.tig.2020.03.005 (2020).

2. Liu, J., Li, J., Wang, H. & Yan, J. Application of deep learning in genomics. *Sci. China Life Sci.* **63**, 1860–1878, DOI: 10.1007/s11427-020-1804-5 (2020).

3. Wang, Y. *et al.* Synthetic promoter design in Escherichia coli based on a deep generative network. *Nucleic Acids Res.* **48**, 6403–6412, DOI: 10.1093/nar/gkaa325 (2020).

4. Xu, C. & Jackson, S. A. Machine learning and complex biological data. *Genome Biol.* **20**, 76, DOI: 10.1186/s13059-019-1689-0 (2019).

5. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322, DOI: 10.1038/s41592-019-0598-1 (2019).

6. Rao, R. *et al.* Evaluating protein transfer learning with tape, DOI: 10.48550/ARXIV.1906.08230 (2019).

7. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589, DOI: 10.1038/s41586-021-03819-2 (2021).

8. Isidro, J. *et al.* Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* **128**, 145–158, DOI: 10.1007/s00122-014-2418-4 (2015).

9. MacLeod, I. M. *et al.* Exploiting biological priors and sequence variants enhances qtl discovery and genomic prediction of complex traits. *BMC Genomics* **17**, 144, DOI: 10.1186/s12864-016-2443-6 (2016).

10. Lozano, R. *et al.* Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nat. Plants* **7**, 17–24, DOI: 10.1038/s41477-020-00834-5 (2021).

11. Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051, DOI: 10.1038/s41588-019-0410-2 (2019).

12. Alonge, M. *et al.* Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23, DOI: https://doi.org/10.1016/j.cell.2020.05.021 (2020).

13. Hufford, M. B. *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662, DOI: 10.1126/science.abg5289 (2021).

14. Washburn, J. D. *et al.* Evolutionarily informed deep learning methods for predicting relative transcript abundance from dna sequence. *Proc. Natl. Acad. Sci.* **116**, 5542–5549, DOI: 10.1073/pnas.1814551116 (2019).

15. Zrimec, J. *et al.* Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141, DOI: 10.1038/s41467-020-19921-4 (2020).

16. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120, DOI: 10.1093/bioinformatics/btab083 (2021).

17. Bukowski, R. *et al.* Construction of the third-generation Zea mays haplotype map. *GigaScience* **7**, DOI: 10.1093/gigascience/gix134 (2017). Gix134.

18. Vaswani, A. *et al.* Attention is all you need, DOI: 10.48550/ARXIV.1706.03762 (2017).

19. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, DOI: 10.48550/arxiv.1810.04805 (2018).

20. Howe, K. L. *et al.* Ensembl genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res* **48**, D689–D695, DOI: 10.1093/nar/gkz890 (2020).

21. O'Leary, N. A. *et al.* Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–45, DOI: 10.1093/nar/gkv1189 (2016).

22. Portwood, I., John L *et al.* MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* **47**, D1146–D1154, DOI: 10.1093/nar/gky1046 (2018).

23. Schnable, J. Grass syntenic gene list sorghum v3 maize v3/4 with teff and oropetium v2. *figshare* https://doi.org/10.6084/m9.figshare.7926674.v1, DOI: 10.6084/m9.figshare.7926674.v1 (2019).

24. Zhang, Y. *et al.* Differentially Regulated Orthologs in Sorghum and the Subgenomes of Maize. *The Plant Cell* **29**, 1938–1951, DOI: 10.1105/tpc.17.00354 (2017).

25. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1101368108 (2011).

26. VanBuren, R., Wai, C. M., Keilwagen, J. & Pardo, J. A chromosome-scale assembly of the model desiccation tolerant grass oropetium thomaeum. *Plant Direct* **2**, e00096, DOI: https://doi.org/10.1002/pld3.96 (2018).

27. VanBuren, R. *et al.* Exceptional subgenome stability and functional divergence in the allotetraploid ethiopian cereal teff. *Nat. Commun.* **11**, 884, DOI: 10.1038/s41467-020-14724-z (2020).

28. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach, DOI: 10.48550/arxiv.1907.11692 (2019).

29. LeNail, A. Nn-svg: Publication-ready neural network architecture schematics. *J. Open Source Softw.* **4**, 747, DOI: 10.21105/joss.00747 (2019).

30. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725, DOI: 10.18653/v1/P16-1162 (Association for Computational Linguistics, Berlin, Germany, 2016).

31. Nambiar, A. *et al.* Transforming the language of life: Transformer neural networks for protein prediction tasks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, DOI: 10.1145/3388440.3412467 (Association for Computing Machinery, New York, NY, USA, 2020).

32. You, Y. *et al.* Large batch optimization for deep learning: Training bert in 76 minutes, DOI: 10.48550/ARXIV.1904.00962 (2019).

33. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45, DOI: 10.18653/v1/2020.emnlp-demos.6 (Association for Computational Linguistics, Online, 2020).

34. Rodriguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E. & Lippman, Z. B. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell* **171**, 470–480.e8, DOI: 10.1016/j.cell.2017.08.030 (2017).

35. Liu, L. *et al.* Enhancing grain-yield-related traits by CRISPR–Cas9 promoter editing of maize CLE genes. *Nat. Plants* **7**, DOI: 10.1038/s41477-021-00858-5 (2021).

36. Van de Velde, J., Van Bel, M., Vaneechoutte, D. & Vandepoele, K. A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. *Plant Physiol.* **171**, 2586–2598, DOI: 10.1104/pp.16.00821 (2016).

37. Kummari, D. *et al.* An update and perspectives on the use of promoters in plant genetic engineering. *J. Biosci.* **45**, 119, DOI: 10.1007/s12038-020-00087-6 (2020).

38. Hirsch, C. N. *et al.* Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. *The Plant Cell* **28**, 2700–2714, DOI: 10.1105/tpc.16.00353 (2016).

39. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **6**, 60, DOI: 10.1186/s40537-019-0197-0 (2019).

40. Vig, J. *et al.* Bertology meets biology: Interpreting attention in protein language models, DOI: 10.48550/arxiv.2006.15222 (2020).

41. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, DOI: 10.1073/pnas.2016239118 (2021). Publisher: National Academy of Sciences Section: Biological Sciences.

Acknowledgements

We thank Peter Koo and Michael Chappuis for critical review and comments on this manuscript. We also acknowledge the support of the Harvard University Institute for Applied Computational Science.

Author contributions statement

B.L. led development of code and tested models with contributions from Z.X., L.Z., and S.N.. R.A. and K.K. conceived of the study and provided technical input. B.L., R.A., and K.K. wrote the manuscript. C.T. oversaw the study, secured resources to support the work, and provided technical input and editing of the manuscript. P.W. provided technical guidance and contributed to editing of the manuscript.

Competing interests

R.A. and K.K. are both employed by and hold equity in Inari Agriculture, Inc. B.L., Z.X., S.N., L.Z., P.W., and C.T. declare no potential conflict of interest.

Additional information

B.L. and Z.X. are employed by McKinsey & Company, R.A. and K.K. are employed by Inari Agriculture Inc. C.T. is currently employed by MIT and Kensho Technologies Inc., but this work was conducted while at Harvard University. P.W. and L.Z. are employed by LinkedIn. S.N. is employed by Microsoft.

Supplementary Files

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

- FloraBERTScientificReportssupplementary.pdf