Supplementary Information

Code availability

All code used in the development of FloraBERT is available in the accompanying GitHub repository at https://github.com/benlevyx/florabert . Any scripts or data files referenced below can be found in this repository.

Preparation of unlabelled 1kb upstream extracts for FloraBERT self-supervised pre-training

Table S1. Statistics of processed sequences from each database

	Ensembl	RefSeq	MaizeGDB
Number of Species	93	126	24*
Number of Processed Sequences	3,747,175	4,145,256	944,474
Average Length of Sequences	914	887	934

*Number of maize cultivars, consisting of the 25 NAM genomes except for *Mo18W*

Code:* *02-download-process-db-data.py

`samtools`[?] (v1.11) and `bedtools`[?] (v2.29.1) were used to extract the 1kb upstream promoter sequences from the raw FASTA files downloaded from the RefSeq, Ensembl, or MaizeGDB databases. Briefly, the raw FASTA files (`.fa`) and annotation files (`.gff3`) were downloaded and saved from the public FTP endpoints for each database. Then, sequences were converted to `.fai` format using the `samtools faidx` command. Next, the `bedtools flank` command was used to extract the 1kb sequence upstream of the TSS. Finally, the `bedtools subtract` command was used to subtract any coding regions of adjacent genes from the resulting 1kb promoter sequences and only the contiguous sequence closest to the TSS was retained.

Preparation of *Zea mays* gene expression data for FloraBERT supervised fine-tuning

Table S2. NAM parent reference genomes and annotations downloaded from MaizeGDB[?] on January 21, 2021

Reference name and version	Annotation name and version
Zm-B73-REFERENCE-NAM-5.0	Zm00001eb.1
Zm-B97-REFERENCE-NAM-1.0	Zm00018ab.1
Zm-CML52-REFERENCE-NAM-1.0	Zm00019ab.1
Zm-CML69-REFERENCE-NAM-1.0	Zm00020ab.1
Zm-CML103-REFERENCE-NAM-1.0	Zm00021ab.1
Zm-CML228-REFERENCE-NAM-1.0	Zm00022ab.1
Zm-CML247-REFERENCE-NAM-1.0	Zm00023ab.1
Zm-CML277-REFERENCE-NAM-1.0	Zm00024ab.1
Zm-CML322-REFERENCE-NAM-1.0	Zm00025ab.1
Zm-CML333-REFERENCE-NAM-1.0	Zm00026ab.1
Zm-HP301-REFERENCE-NAM-1.0	Zm00027ab.1
Zm-Il14H-REFERENCE-NAM-1.0	Zm00028ab.1
Zm-Ki3-REFERENCE-NAM-1.0	Zm00029ab.1
Zm-Ki11-REFERENCE-NAM-1.0	Zm00030ab.1
Zm-Ky21-REFERENCE-NAM-1.0	Zm00031ab.1
Zm-M37W-REFERENCE-NAM-1.0	Zm00032ab.1
Zm-M162W-REFERENCE-NAM-1.0	Zm00033ab.1
Zm-Mo18W-REFERENCE-NAM-1.0	Zm00034ab.1
Zm-Ms71-REFERENCE-NAM-1.0	Zm00035ab.1
Zm-NC350-REFERENCE-NAM-1.0	Zm00036ab.1
Zm-NC358-REFERENCE-NAM-1.0	Zm00037ab.1
Zm-Oh7B-REFERENCE-NAM-1.0	Zm00038ab.1
Zm-Oh43-REFERENCE-NAM-1.0	Zm00039ab.1
Zm-P39-REFERENCE-NAM-1.0	Zm00040ab.1
Zm-Tx303-REFERENCE-NAM-1.0	Zm00041ab.1
Zm-Tzi8-REFERENCE-NAM-1.0	Zm00042ab.1

Step 1: ***`cmds_to_dl_NAM_and_B73v5_genomes_and_create_HiSat_indices.txt`***

Gene expression data for 9 tissues of B73 and the 25 founders of the Nested Association Mapping panel in maize[?,?] were aligned against their respective reference genomes and annotations as detailed below. Data were first retrieved from NCBI SRA BioProjects PRJEB35943 and PRJEB36014. The `.gff` annotations were converted to `.gtf` using `gffread`[?] (v0.12.3) and then reference genome indices were created using `hisat2`[?] (v2.2.1) with the `hisat2-build` command after running the included `hisat2_extract_splice_sites.py` and `hisat2_extract_exons.py` scripts on the `.gtf` annotations.

Step 2: ***`alignment_script.sh`***

Reads were then trimmed using `trimmomatic`[?] (v0.39), aligned against their respective indices using `hisat2`[?] (v2.2.1), and then counted using `featureCounts`[?] (v2.0.1).

Step 3: ***`feature_counts_to_TPMs.R`***

The resulting counts were then standardized to transcripts per million (TPM)[?].

PCA decomposition of species-level embeddings.

We compared the embeddings produced by two models to determine whether relationships between species were captured better by one or the other:

1. FloraBERT: We extracted the output of the final layer of the model and averaged these outputs on a per-species basis to obtain species-level embeddings. To obtain a gene promoter embedding, the final-layer attention weights for all tokens (excluding padding tokens that are used to denote the start and end of sequences) are averaged into a single 768-dimensional vector. The gene embeddings for a species are then averaged element-wise to obtain a single 768-dimensional centroid for the species.

2. 1-mer: We counted up the number of occurrences of each nucleotide (A, T, C, G, or N, for unknown) and computed the relative frequencies of each nucleotide across all the promoters in the entire species

We then used PCA to reduce these species-level embeddings into just two components, as illustrated in Figure S1. In both models, we can observe a clear separation between the two main groups in land plants: monocots and eudicots. In addition, species in the same genus, such as *Triticum* and *Oryza*, are tightly clustered together, reflecting the fact that their upstream regulatory sequences are mapped to similar regions in high-dimensional embedding space.

This demonstrates that, when averaged at the species level and linearly reduced through PCA, the relationships between the embedded promoter sequences largely reflect relative frequencies of nucleotides. In other words, we do not find evidence that FloraBERT's embeddings encode any evolutionarily salient relationships between species/promoters over and above the information contained in the relative frequencies of nucleotides in those promoters.

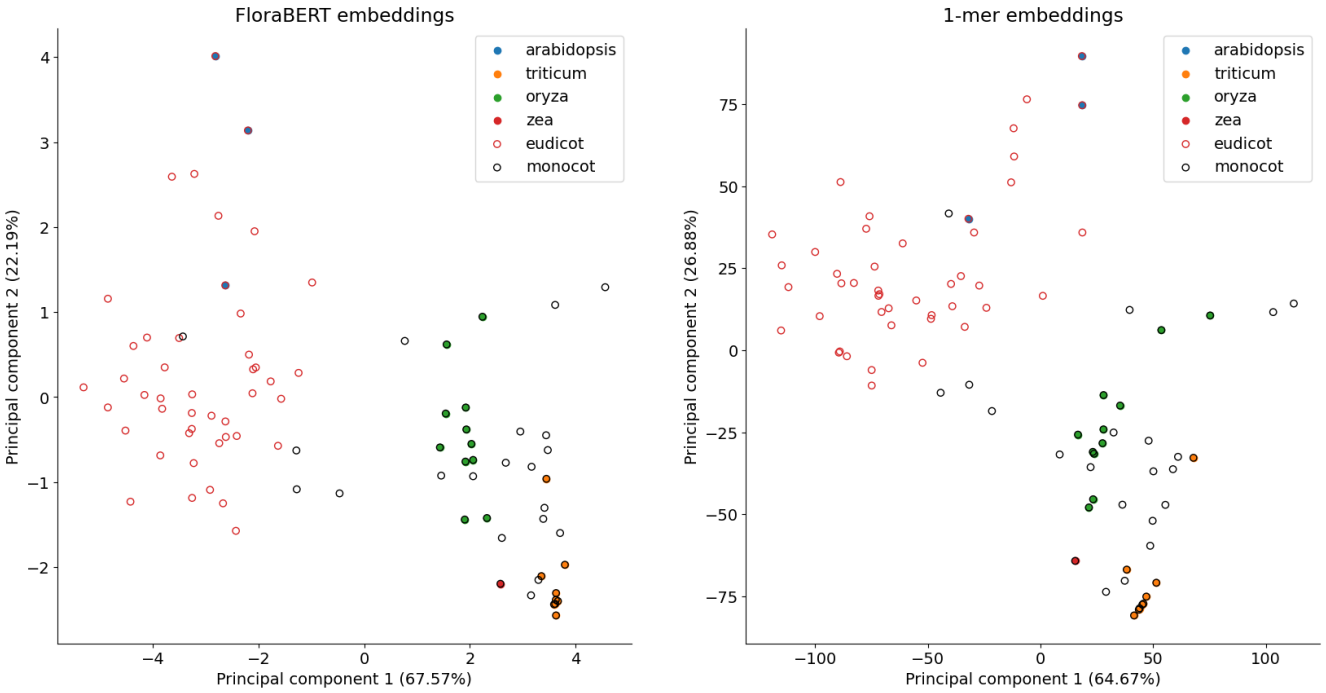

Figure S1. First two PCA components of FloraBERT species-averaged final-layer attention embeddings (left) and 1-mer bag-of-kmers model (right). and species-level embeddings. Selected genera are highlighted in both panels. In this figure, *Zea* corresponds only to the B73 cultivar, shown for illustrative purposes.