Current genomic deep learning architectures generalize across grass species but not alleles

Travis Wrightsman[1], Taylor H. Ferebee[2], M. Cinta Romay[3], Taylor AuBuchon-Elder[4], Alyssa R. Phillips[5], Michael Syring[6], Elizabeth A. Kellogg[4], Edward S. Buckler[1,3,7]

[1]Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA 14853 [2]Department of Computational Biology, Cornell University, Ithaca, NY, USA 14853 [3]Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA 14853 [4]Donald Danforth Plant Science Center, St. Louis, MO, USA 63132 [5]Department of Evolution and Ecology, University of California, Davis, CA, USA 95616 [6]Iowa State University, Ames, IA, USA 50011 [7]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, USA 14853

Abstract

Non-coding regions of the genome are just as important as coding regions for understanding the mapping from genotype to phenotype. Interpreting deep learning models trained on RNA-seq is an emerging method to highlight functional sites within non-coding regions. Most of the work on RNA abundance models has been done within humans and mice, with little attention paid to plants. Here, we benchmark four genomic deep learning model architectures with genomes and RNA-seq data from 18 species closely related to maize and sorghum within the Andropogoneae. The Andropogoneae are a tribe of C4 grasses that have adapted to a wide range of environments worldwide since diverging 18 million years ago. Hundreds of millions of years of evolution across these species has produced a large, diverse pool of training alleles across species sharing a common physiology. As model input, we extracted 1,026 base pairs upstream of each gene's translation start site. We held out maize as our test set and two closely related species as our validation set, training each architecture on the remaining Andropogoneae genomes. Within a panel of 26 maize lines, all architectures predict expression across genes moderately well but poorly across alleles. DanQ consistently ranked highest or second highest among all architectures yet performance was generally very similar across architectures despite orders of magnitude differences in size. This suggests that state-of-the-art supervised genomic deep learning models are able to generalize moderately well across related species but not sensitively separate alleles within species, the latter of which agrees with recent work within humans. We are releasing the preprocessed data and code for this work as a community benchmark to evaluate new architectures on our across-species and across-allele tasks.

Introduction

Non-coding regions of the genome are well-known to be as important as coding regions for understanding how genotype determines phenotype (Finucane et al., 2015; Rodgers-Melnick et al., 2016). Though tools like AlphaFold2 (Jumper et al., 2021) have dramatically improved our ability to study coding sequence, similarly performing tools do not yet exist for non-coding regions. Nevertheless, over the last decade deep learning models have rapidly improved performance in predicting non-coding genomic features such as chromatin accessibility (Kelley, 2020; Wrightsman et al., 2022), transcription factor binding (Žiga Avsec, Weilert, et al., 2021; Mejía-Guerra & Buckler, 2019), and RNA abundance (Žiga Avsec, Agarwal, et al., 2021; Linder et al., 2023) directly from DNA sequence. These models can then be queried to highlight functional non-coding sites, which can be useful for filtering large sets of variants down to promising

genome editing targets. Further, since most of the modeling work has been done on human and mouse data, there is a need to benchmark their performance in plants.

Models that predict RNA abundance from sequence are particularly attractive due to the relatively cheap cost and standardized protocols of RNA-seq. However, there is room for improvement in these models across a number of areas. While RNA abundance models have shown high performance across genes, recent work in humans (Huang et al., 2023) has highlighted their lack of sensitivity across individuals. Some expression model architectures (Žiga Avsec, Agarwal, et al., 2021; Linder et al., 2023) include coding sequence in the input, which is known to lead to overfitting on gene family instead of true regulatory sequence differences (Washburn et al., 2019). There is also a tendency to maximize data when training these models, without actually measuring the rate of diminishing returns for each additional observation. Finally, while multiple species have been included in some training sets, it is common to test on a set of held-out chromosomes within the training species, rather than testing on a completely held-out species.

Deep learning models benefit from large and diverse training sets of different tissues and genotypes, which are rarely available outside model species. To train RNA expression models on larger sample sizes, we leveraged new long-read genomes and RNA-seq data from 15 wild species of the Andropogoneae tribe. Diverging around 17.5 million years ago (Welker et al., 2020), the Andropogoneae includes globally staple crop plants such as maize, sorghum, and sugarcane. Millions of years of evolution within the tribe has provided a large, diverse pool of training alleles. Sorghum and maize diverged around 12 million years ago (Mya), on the order of the human-chimpanzee split (6–10 Mya), but have a 10-fold higher rate of nucleotide divergence (Chimpanzee Sequencing and Analysis Consortium, 2005; Zhang et al., 2017).

We tested four sequenced-based genomic deep learning architectures, DanQ (Quang & Xie, 2016), HyenaDNA (Nguyen et al., 2023), FNetCompression (Pipoli et al., 2023), and a smaller version of Enformer (Žiga Avsec, Agarwal, et al., 2021), on their ability to predict across species and alleles. DanQ is one of the earliest genomic deep learning architectures, leveraging a long short-term memory recurrent layer to learn the syntax and grammar of motifs detected by a convolutional layer. Enformer is a massive transformer architecture with a context size near 100 kilobases that is among the best performing models for human expression prediction. HyenaDNA is a novel architecture capable of handling long context windows of up to a million base pairs. FNetCompression combines a fast Fourier transform with multi-head attention to efficiently learn from sequences of up to 10 kilobases with a few orders of magnitude less parameters than the other architectures.

We aimed to investigate, from a plant perspective, two major open questions in expression modeling from sequence: 1) How well do current sequence-based deep learning architectures generalize across species? and 2) How sensitive are these models across individuals?

Results

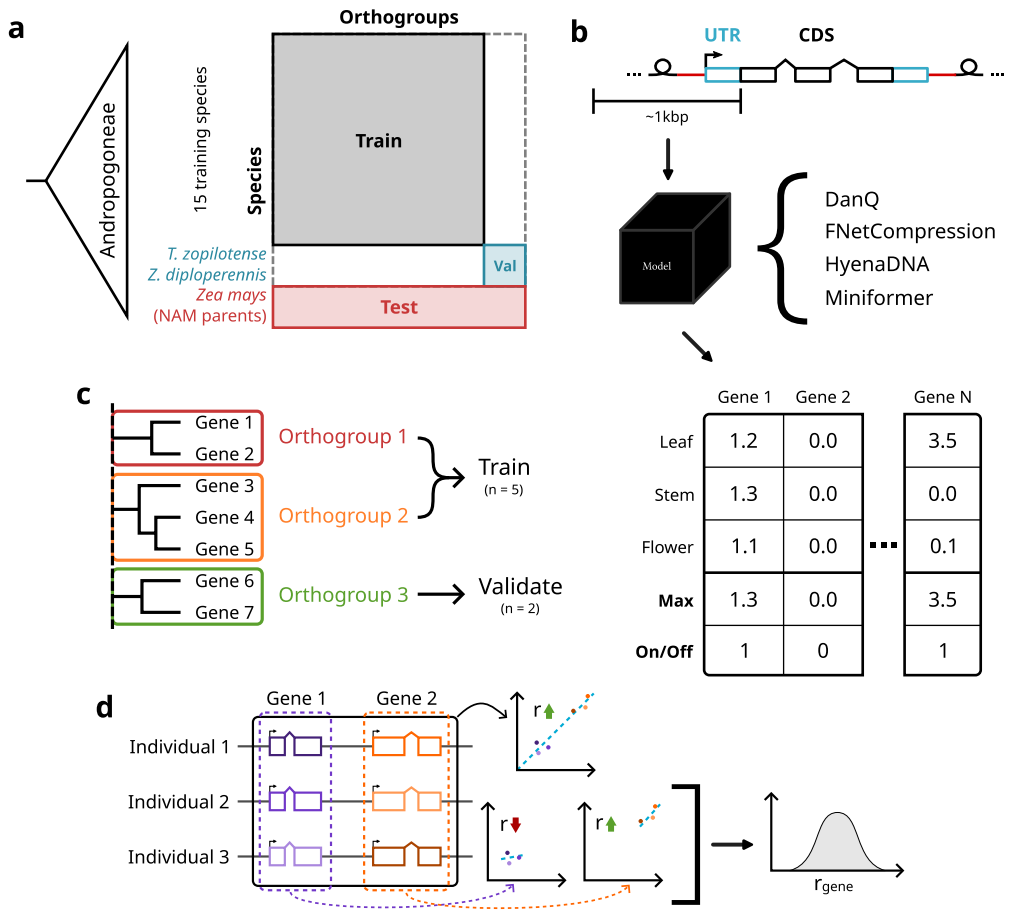


Figure 1: Methods overview: a) data splits; b) models, features, and targets; c) orthogroup-guided splitting; d) metrics (across- vs. within-gene performance)

Current genomic deep learning architectures generalize across species

We trained all four architectures on genomic sequence and RNA-seq data from 15 species within the Andropogoneae clade (Figure 1a). Our validation set consisted of the two sampled species closest to *Zea mays*, *Tripsacum zopilotense* and *Zea diploperennis*, all three of which fall within the Tripsacinae subtribe that diverged a few (0.6–4) million years ago (Chen et al., 2022; Welker et al., 2020). Our test set was the 26 inbred parents of the maize NAM population (Hufford et al., 2021; Yu et al., 2008), held out until hyperparameters were frozen. As input, we extracted 1,026 base pairs upstream of the translation start site to match HyenaDNA's "tiny" configuration (Figure 1b). We trained and evaluated all architectures on two regression tasks, maximum expression across tissues and absolute expression in leaf, as well as two classification tasks, expressed in any tissue and expressed in leaf (Figure 1b).

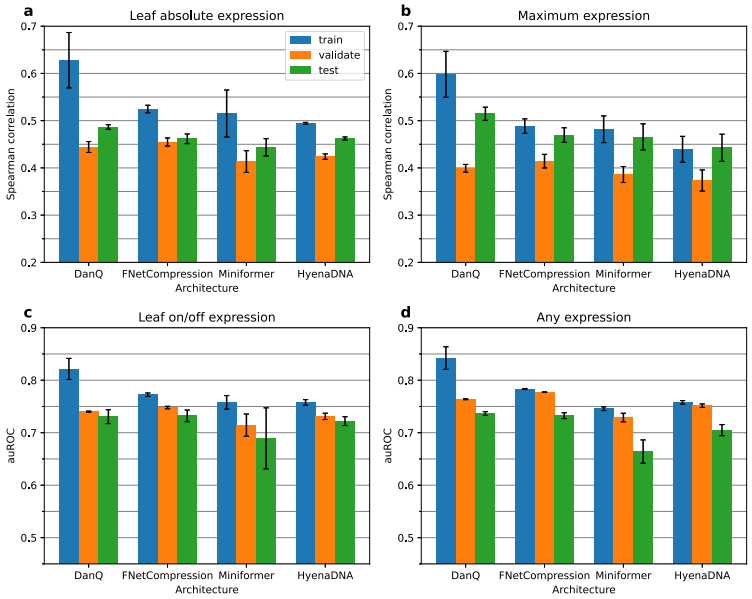


Figure 2: a–d) Model performance across all genes and data splits. Each subfigure shows the performance of all architectures on a single task. Error bars represent one standard deviation from the mean in each direction.

Though benchmarking of sequence-based models has been done within humans (Karollus et al., 2023) and across species in the training set (Kelley, 2020; Levy et al., 2022), there has been little evaluation on entirely held out species. To establish a baseline in plants, we measured performance of all architectures, tasks, genes, and data splits (Figure 2a–d). Rankings by Spearman correlation on the test set are inconsistent, except that DanQ performed the best or tied closely with the best across all tasks. Remarkably, DanQ performs only slightly worse (Figure 2a; $\Delta r = 0.09$) than Enformer in a recent within-human single tissue benchmark (Huang et al., 2023) despite predicting on an unobserved species. Despite having moderate Spearman and Pearson correlation (Supplemental Figure 1), DanQ's predictions on the test set are still underwhelming (Supplemental Figure 2). We observed test set auROC scores in the any expression task slightly lower than previous results (Washburn et al., 2019) on promoter expression classification models trained and tested only within maize. Taken together, these results support modern genomic deep learning architectures are capable of generalizing almost as well across closely related species as they do within species.

Data quantity matters more than composition for modeling RNA abundance across species

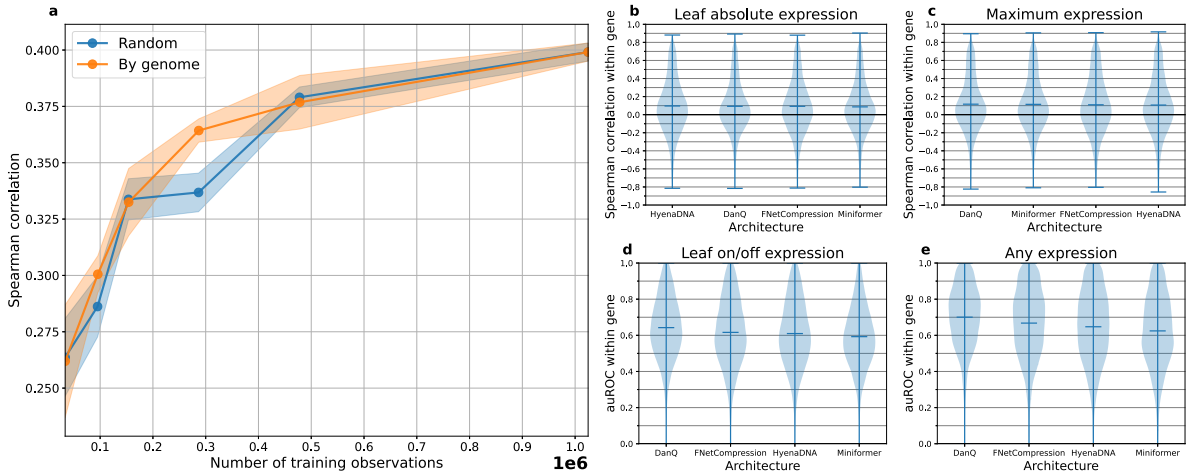


Figure 3: a) Validation set performance of DanQ on the maximum expression task across varying training set sizes and compositions. Points on the lines are mean Pearson correlation across replicate training runs. The standard deviation across replicates is shaded around the mean line. The exponent scale of the bottom axis is denoted in bold on the right. b–e) Distributions of model performance within orthogroups for each task. Architectures are sorted from highest (left) average within-orthogroup performance to lowest (right). Bars within the violins represent the mean of the distribution.

Despite the growing number of plant genomes with transcriptomic data (Sreedasyam et al., 2023), each genome added to the training set increases training time and may give diminishing returns. We measured changes in DanQ's performance on progressively larger fractions of the training data, iteratively adding sequences from a set of genomes or randomly from across all training genomes. Pearson correlation on the validation set rises until approximately 200,000 observations when it begins to show diminishing returns for larger training set sizes (Figure 3a). However, the slope remains positive between the half size and full size data points, suggesting room for improvement with further observations. Comparing iteratively adding whole genomes to randomly sampling an equivalent number of alleles from the entire training set, there are only substantial differences when using less than 8 genomes, with random performing worse than 4 whole genomes. The ablation results clearly support the use of further data to achieve higher performance across genes, which can come from sequencing additional related species.

Current architectures poorly generalize across individuals of an inbred maize panel

Recent work (Huang et al., 2023) has shown that current models poorly explain expression variation across individuals. Since our test set is a collection of maize alleles with an order of magnitude more diversity than humans (Chia et al., 2012), we looked at the distribution of test set performance within each orthogroup and expected to see similarly low or even lower performance. We only considered orthogroups that had at least 20 orthologs to have sufficient sample size for calculating correlation or auROC. We saw much lower average within-orthogroup Spearman and Pearson correlations as well as auROC compared to the global across-gene metrics, except for the any expression task (Figure 3b–e; Supplemental Figure 3), which also shows clear differences between architectures. The average within-orthogroup Spearman correlation in the single tissue regression task is double ($r = 0.092$) than what was observed with Enformer ($r = 0.045$) in humans.

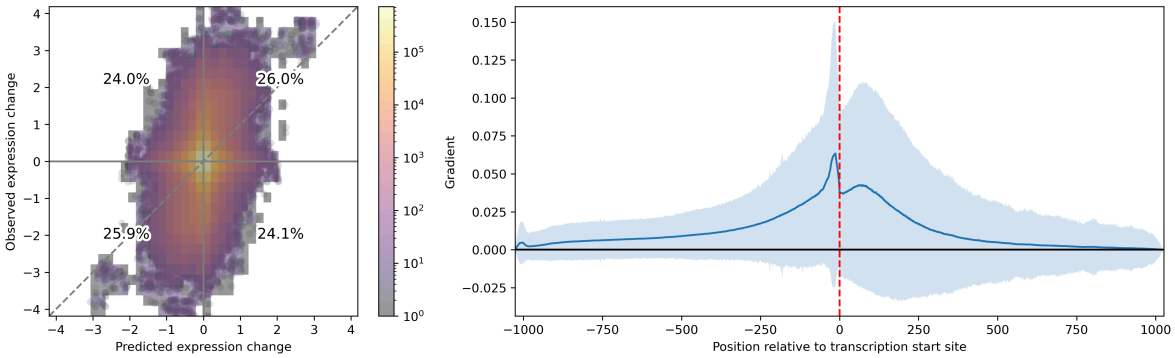


Figure 4: Left: Predicted versus observed $\log_{10}$ expression change in leaf between all NAM ortholog promoter allele pairs within an orthogroup. Percentages in the middle of each quadrant display the proportion of non-zero data points in that quadrant. Right: Saliency map for DanQ trained on maximum expression task. The mean across all B73 genes is plotted as a line, with a single standard deviation shaded above and below.

As an alternative allele-level comparison, we also looked at how well DanQ predicted expression differences between all pairs of maize ortholog promoter alleles within a orthogroup. We observed a general positive relationship between the two, but there is still quite a bit of noise in the predictions (Figure 4, left). The Pearson and Spearman correlation coefficients between the observed and predicted fold changes were only 0.22 and 0.08, respectively. Strikingly, pairs of orthogroups that are two orders of magnitude apart in expression level are still sometimes predicted with the incorrect direction. Despite current architectures showing promising across gene performance in unobserved species, they still struggle across shorter evolutionary timescales, similar to what was seen in humans (Huang et al., 2023).

The maximum expression regression model focuses on the core promoter region

Based on theory and prior interpretation work on expression models (Mendoza-Revilla et al., 2023), we hypothesized our expression models would also pay most attention to the region surrounding the transcription start site. Looking at the average saliency map for DanQ across all B73 genes on the maximum expression task we see that DanQ indeed focuses on the core promoter region and the 5′ UTR (Figure 4, right). There is relatively high variance in the gradient around the transcription start site, taping off with increasing distance, though decaying slower in the UTR than promoter. This hyperfocus on the core promoter region could be why DanQ and other architectures struggle to distinguish expression differences in highly related sequences, since functional mutations are less likely to accumulate in this highly constrained region.

Discussion

Here we have shown that four genomic deep learning architectures are capable of generalizing across species, though they also show the same lack of allelic sensitivity seen in humans (Huang et al., 2023). FNetCompression's performance is particularly remarkable because it has several orders of magnitude fewer parameters than DanQ (57k versus 1.6m, respectively). Large foundation models such as AgroNT (Mendoza-Revilla et al., 2023) show promising results within the training species, but FNetCompression suggests smaller, more efficient models, perhaps also utilizing a fast Fourier transform, are worth further exploration. Since the Pearson correlations we observed are still far from perfect, it is worthwhile to note

that we do not expect *cis* sequence-based models to ever reach perfect correlation, as *cis* factors explain only a third of the genetic variation in expression in maize (Giri et al., 2021). The fact that our models show across individual performance in maize ($r_s = 0.092$) double that observed in humans (Huang et al., 2023) is puzzling. Population genetics has shown that maize has an order of magnitude more genetic variance than humans (Chia et al., 2012), yet our models are generalizing across maize individuals better than what was observed across human individuals. Unlike our validation set, our maize test set includes orthologs of sequences in our training set, which may result in slightly inflated performance estimates. However, this inflation is expected to be less than when coding sequences are included in the model (Washburn et al., 2019), as was the case for the human benchmark. More work will be needed to investigate this and, more generally, where the remaining errors are being made in these models.

This stringent benchmark, both across species and across individuals in a held-out species, is something that all expression models should be continually evaluated against to get a better sense of generalizability than within species testing. Our ablation results show that we are not yet saturated in terms of training data, meaning there is a need for further benchmarks on larger sets of species. Training across species presents the opportunity to not only leverage larger data sets but to learn the general rules of eukaryotic gene expression. Future work should consider training across many distantly-related species to learn general rules, then successively fine-tuning within clades to learn lineage-specific patterns. Consideration of data balance may be necessary, as prevalent polyploidy within plants (Song et al., 2023) leads to vastly different gene counts across species and complicates transcript quantification. However, scaling to bigger data will come with metadata challenges, exacerbated by the plethora of standards across databases (McQuilton et al., 2016). While considering better architectures with higher data needs, it will be increasingly important to better organize expression databases. Lastly, with the rising utilization of foundation models trained across massive numbers of genomes, it will also be critical to maintain true hold-out species for fair model evaluation. A CASP-like competition for RNA abundance modeling may be useful for this, as new sequence-based models of non-coding biology are developed.

Materials and Methods

The companion Zenodo repository (Wrightsman et al., 2024) contains the source code required to reproduce this manuscript. Pandas (McKinney, 2010) and Polars (Vink et al., 2023) were used to process tabular data. Matplotlib (Hunter, 2007) was used for plotting figures. This manuscript is written in and rendered using Typst (Haug, 2022; Mädje, 2022).

Software Environments

Software environments were managed with pixi (Arts et al., n.d.). Packages were downloaded from the conda-forge (Conda-Forge Community, 2015) and Bioconda (Grüning et al., 2018) Conda channels. The exact software versions used in this work are defined in configuration files within the manuscript's companion repository.

Data preprocessing

All genome assemblies and annotations were downloaded from MaizeGDB (Hufford et al., 2021; Portwood et al., 2018). Version 5 of the B73 assembly and annotation was used. For PanAnd, version 1 of the assemblies and version 2 of the annotations were used. B73 and other NAM parent RNA-seq data were downloaded from ArrayExpress accessions E-MTAB-8628 and E-MTAB-8633, respectively. Other Andropogoneae RNA-seq data were downloaded from NCBI accession PRJNA1098707. Transcript quantifications were obtained using quantify-RNA-pipeline (Wrightsman, 2023). RNA-seq samples with less

than 5 million mapped reads were dropped from further analysis. eggnog-mapper (Cantalapiedra et al., 2021) was used to assign proteins to Poales orthogroups.

Zea mays genes were assigned to the test set. 90% of orthogroups were randomly chosen as training orthogroups, with the other 10% used for validation. *Zea diploperennis* and *Tripsacum zopliotense* genes in the validation orthogroups were used as the validation set. Genes in the training orthogroups in all remaining Andropogoneae genomes were assigned to the training set.

Annotations were processed using gffutils (Dale, 2023). For each gene, the highest expressed transcript across all tissues was selected as a representative gene model. TPM values from other transcripts of the same gene that share the same transcription start site were added to the chosen transcript's TPM. For the purposes of computing max expression, only leaf, shoot, and floral tissues were used as only those tissues had sufficient sampling across all species. The any tissue and leaf on/off expression classification task targets were binarized from the max expression and leaf absolute expression regression task targets (TPM). Specifically, TPM values of zero were kept as zero (unexpressed) and TPM values greater than zero were set to one (expressed).

Model architectures

Exact hyperparameter settings for each architecture is specified in configuration files within the companion repository. DanQ (Quang & Xie, 2016) and FNetCompression (Pipoli et al., 2023) were both converted to PyTorch (Paszke et al., 2019), keeping all hyperparameters identical. Miniformer is a scaled-down version of the Enformer (Žiga Avsec, Agarwal, et al., 2021) architecture, with lower model dimensions and fewer layers. HyenaDNA (Nguyen et al., 2023) was used in the "tiny" configuration. Classification architectures were identical to their regression counterparts except that the final activation function was changed to sigmoid.

Training

PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019) and Hydra (Yadan, 2019) were used to orchestrate the training process and provide an interface for the data loader. MLFlow (Zaharia et al., 2018) was used to track experiment parameters and metrics as well as store model artifacts. As input, 1,026 base pairs upstream from the translation start site were extracted. The sequence was reverse complemented if the transcript was on the negative strand. If the model used 1-hot encoded sequence as input, the sequence was 1-hot encoded and then padded or trimmed as needed to be exactly 1,026 base pairs in length. If the model used tokens as input, the input sequence was tokenized to a max length of 1,026, padding as needed. If the task was regression (max or absolute leaf expression), TPM values were $\log_{10}(\mathrm{TPM} + 1)$ transformed. Training continued until the validation loss failed to decrease after three epochs. The model checkpoint at the end of the epoch with the lowest validation loss was kept. Each combination of architecture and task was run three times with different initial seeds to estimate model robustness. Any runs that failed to converge were restarted with a different seed value until a total of three converged runs were obtained.

Ablation

"By genome" ablation was performed by filtering to one, two, three, four, or eight training genomes, in order of increasing phylogenetic distance from the test set. The first eight genomes, in order, were *Tripsacum dactyloides* "FL_9056069_6", "McKain_334-5", *Elionurus tripsacoides*, *Hemarthria compressa*, *Thelepogon elegans*, *Sorghastrum nutans*, *Ischaemum rugosum*, and *Pogonatherum paniceum*. "Random" ablation randomly sampled an equivalent number of observations from the total set of observations from all 15

training species. For example, if the "By genome" ablation had two genomes with 30,000 and 35,000 observations, then the corresponding "random" ablation experiment would randomly sample 65,000 observations from the total training set. Each ablation run was repeated six times to measure robustness.

Ortholog contrast

The first DanQ training run model was used to predict the expression for each transcript in the test set. All possible pairs of orthologs within each orthogroup were generated for the contrast. Orthogrups were filtered to those that contained between 20 and 35 members, to avoid private genes and retroelements and have sufficient sample sizes to calculate correlation and auROC.

Saliency map

Captum (Kokhlikyan et al., 2020) was used to compute saliency. For each position, the absolute value of saliency for each channel was summed. The mean and standard deviation of this sum was computed across all B73 genes.

Acknowledgements

This work was funded by the USDA-ARS and NSF PanAnd grant (Award 1822330). T.H.F. was supported by a USDA NIFA AFRI predoctoral fellowship (TF: 2022-67011-36564). Some compute resources were provided by the Cornell University BRC Bioinformatics Core Facility (RRID:SCR_021757). This research used resources provided by the SCINet project and/or the AI Center of Excellence of the USDA Agricultural Research Service, ARS project numbers 0201-88888-003-000D and 0201-88888-002-000D. Aaron Gokaslan helped design and debug the data loader.

Author Contributions

In terms of the Contributor Roles Taxonomy (CRediT): All authors were responsible for Writing – review & editing. T.W. was additionally responsible for Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, and Writing – original draft. T.H.F. for Conceptualization, Data curation, and Methodology. M.C.R. for Data curation and Project administration. T.A., A.R.P., and M.S. for Resources. E.A.K. for Data curation and Resources. E.S.B. for Conceptualization, Funding acquisition, Methodology, Project administration, and Supervision.

References

Arts, R., Zalmstra, B., Vollprecht, W., & Jager, T. de. *pixi*. https://github.com/prefix-dev/pixi/releases/tag/v0.16.1

Avsec, Žiga, Agarwal, et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, *18*(10), 1196. https://doi.org/10.1038/s41592-021-01252-x

Avsec, Žiga, Weilert, et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, *53*(3), 354. https://doi.org/10.1038/s41588-021-00782-6

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Chen, L., Luo, J., Jin, M., Yang, N., Liu, X., Peng, Y., Li, W., Phillips, A., Cameron, B., Bernal, J. S., Rellán-Álvarez, R., Sawers, R. J. H., Liu, Q., Yin, Y., Ye, X., Yan, J., Zhang, Q., Zhang, X., Wu, S., … Yan, J. (2022).

Genome sequencing reveals evidence of adaptive variation in the genus Zea. *Nature Genetics*, *54*(11), 1736. https://doi.org/10.1038/s41588-022-01184-y

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., Leon, N. de, Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., Gore, M., Guill, K. E., Holland, J., Hufford, M. B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., … Ware, D. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, *44*(7), 803. https://doi.org/10.1038/ng.2313

Chimpanzee Sequencing and Analysis Consortium, T. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69. https://doi.org/10.1038/nature04072

Conda-Forge Community. (2015). *The conda-forge Project: Community-based Software Distribution Built on the conda Package Format and Ecosystem*. Zenodo. https://doi.org/10.5281/ZENODO.4774216

Dale, R. (2023). *gffutils*. GitHub.

Falcon, W., & The PyTorch Lightning team. (2019, March). *PyTorch Lightning*. https://doi.org/10.5281/zenodo.3828935

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., … Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, *47*(11), 1228. https://doi.org/10.1038/ng.3404

Giri, A., Khaipho-Burch, M., Buckler, E. S., & Ramstein, G. P. (2021). Haplotype associated RNA expression (HARE) improves prediction of complex traits in maize. *PLOS Genetics*, *17*(10), e1009568. https://doi.org/10.1371/journal.pgen.1009568

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475. https://doi.org/10.1038/s41592-018-0046-7

Haug, M. E. (2022). *Fast Typesetting with Incremental Compilation* [Technical University of Berlin]. https://www.user.tu-berlin.de/mhaug/fast-typesetting-incremental-compilation.pdf

Huang, C., Shuai, R. W., Baokar, P., Chung, R., Rastogi, R., Kathail, P., & Ioannidis, N. M. (2023). Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, *55*(12), 2056. https://doi.org/10.1038/s41588-023-01574-w

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., … Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373*(6555), 655. https://doi.org/10.1126/science.abg5289

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90. https://doi.org/10.1109/mcse.2007.55

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583. https://doi.org/10.1038/s41586-021-03819-2

Karollus, A., Mauermeier, T., & Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, *24*(1). https://doi.org/10.1186/s13059-023-02899-9

Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. *PLOS Computational Biology*, *16*(7), e1008050. https://doi.org/10.1371/journal.pcbi.1008050

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch*.

Levy, B., Xu, Z., Zhao, L., Kremling, K., Altman, R., Wong, P., & Tanner, C. (2022). *FloraBERT: cross-species transfer learning with attention-based neural networks for gene expression prediction*. https://doi.org/10.21203/rs.3.rs-1927200/v1

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., & Kelley, D. R. (2023). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Biorxiv*. https://doi.org/10.1101/2023.08.30.555582

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-00a

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., & Sansone, S.-A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*, *2016*, baw75. https://doi.org/10.1093/database/baw075

Mejía-Guerra, M. K., & Buckler, E. S. (2019). A k-mer grammar analysis to uncover maize regulatory architecture. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1693-2

Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., Almeida, B. P. de, Richard, G., Caton, J., Lopez Carranza, N., Skwark, M., Laterre, A., Beguir, K., Pierrot, T., & Lopez, M. (2023). *A Foundational Large Language Model for Edible Plant Genomes*. https://doi.org/10.1101/2023.10.24.563624

Mädje, L. (2022). *A Programmable Markup Language for Typesetting* [Technical University of Berlin]. https://www.user.tu-berlin.de/laurmaedje/programmable-markup-language-for-typesetting.pdf

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., Ermon, S., Ré, C., & Baccus, S. (2023). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 43177–43201). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/86ab6927ee4ae9bde4247793c46797c7-Paper-Conference.pdf

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pipoli, V., Attanasio, G., Lovino, M., & Ficarra, E. (2023). *Squeeze and Learn: Compressing Long Sequences with Fourier Transformers for Gene Expression Prediction*.

Portwood, I., John L, Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., Braun, B. L., Dietze, M., Dunfee, B., Elsik, C. G., Manchanda, N., Coe, E., Sachs, M., Stinard, P., Tolbert, J., … Andorf, C. M. (2018). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Research*, *47*(D1), D1146-D1154. https://doi.org/10.1093/nar/gky1046

Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, *44*(11), e107–e107. https://doi.org/10.1093/nar/gkw226

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., & Buckler, E. S. (2016). Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences*, *113*(22). https://doi.org/10.1073/pnas.1525244113

Song, B., Buckler, E. S., & Stitzer, M. C. (2023). New whole-genome alignment tools are needed for tapping into plant diversity. *Trends in Plant Science*. https://doi.org/10.1016/j.tplants.2023.08.013

Sreedasyam, A., Plott, C., Hossain, M. S., Lovell, J. T., Grimwood, J., Jenkins, J. W., Daum, C., Barry, K., Carlson, J., Shu, S., Phillips, J., Amirebrahimi, M., Zane, M., Wang, M., Goodstein, D., Haas, F. B., Hiss, M., Perroud, P.-F., Jawdy, S. S., … Schmutz, J. (2023). JGI Plant Gene Atlas: an updateable transcriptome resource to improve functional gene descriptions across the plant kingdom. *Nucleic Acids Research*, *51*(16), 8383–8401. https://doi.org/10.1093/nar/gkad616

Vink, Gooijer, Beedie, Gorelli, Zundert, Guo, Hulselmans, universalmind303, Peters, Marshall, chielP, nameexhaustion, Santamaria, Heres, Magarick, ibENPC, Wilksch, Leitao, Haag, … cmdlineluser. (2023). *pola-rs/polars: Python Polars 0.20.3-rc.2*. Zenodo. https://doi.org/10.5281/ZENODO.7697217

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., & Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*, *116*(12), 5542. https://doi.org/10.1073/pnas.1814551116

Welker, C. A. D., McKain, M. R., Estep, M. C., Pasquet, R. S., Chipabika, G., Pallangyo, B., & Kellogg, E. A. (2020). Phylogenomics enables biogeographic analysis and a new subtribal classification of Andropogoneae (Poaceae—Panicoideae). *Journal of Systematics and Evolution*, *58*(6), 1003. https://doi.org/10.1111/jse.12691

Wrightsman, T. (2023, November). *quantify-RNA-pipeline*. Zenodo. https://doi.org/10.5281/zenodo.10201978

Wrightsman, T., Ferebee, T. H., Romay, M. C., AuBuchon-Elder, T., Phillips, A. R., Michael, S., Kellogg, E. A., & Buckler, E. S. (2024). *Current genomic deep learning architectures generalize across grass species but not alleles*. https://doi.org/10.5281/zenodo.10723668

Wrightsman, T., Marand, A. P., Crisp, P. A., Springer, N. M., & Buckler, E. S. (2022). Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks. *The Plant Genome*, *15*(3). https://doi.org/10.1002/tpg2.20249

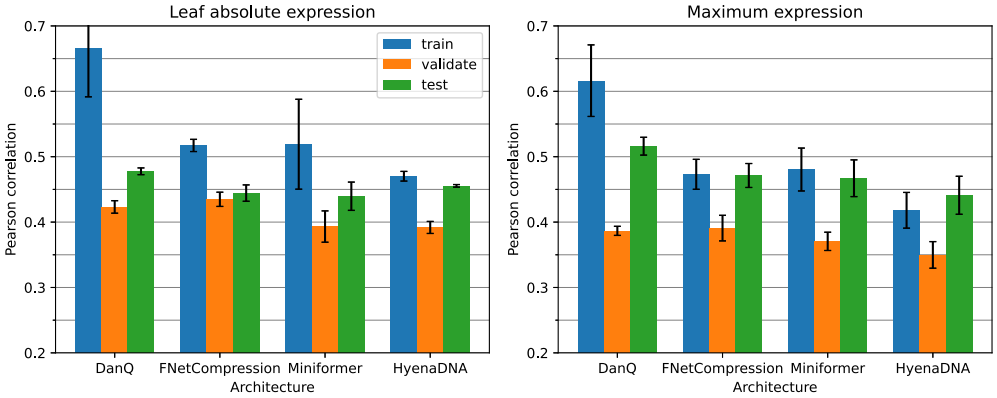
Yadan, O. (2019). *Hydra - A framework for elegantly configuring complex applications*. GitHub.

Yu, J., Holland, J. B., McMullen, M. D., & Buckler, E. S. (2008). Genetic Design and Statistical Power of Nested Association Mapping in Maize. *Genetics*, *178*(1), 539. https://doi.org/10.1534/genetics.107.074245

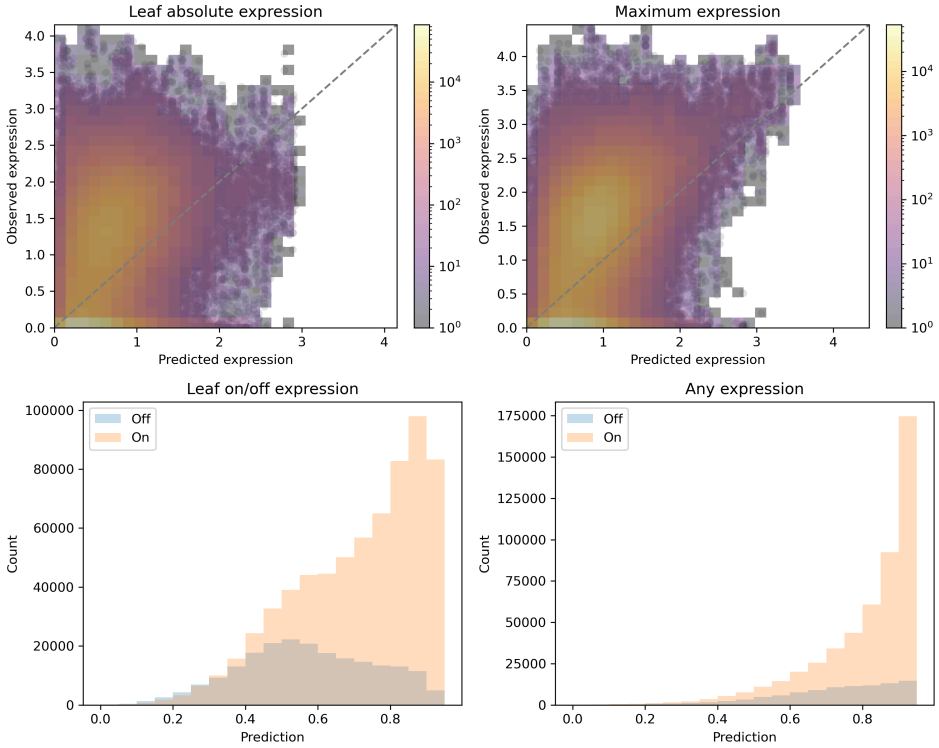
Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., & others. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull.*, *41*(4), 39–45.

Zhang, Y., Ngu, D. W., Carvalho, D., Liang, Z., Qiu, Y., Roston, R. L., & Schnable, J. C. (2017). Differentially Regulated Orthologs in Sorghum and the Subgenomes of Maize. *The Plant Cell*, *29*(8), 1938. https://doi.org/10.1105/tpc.17.00354

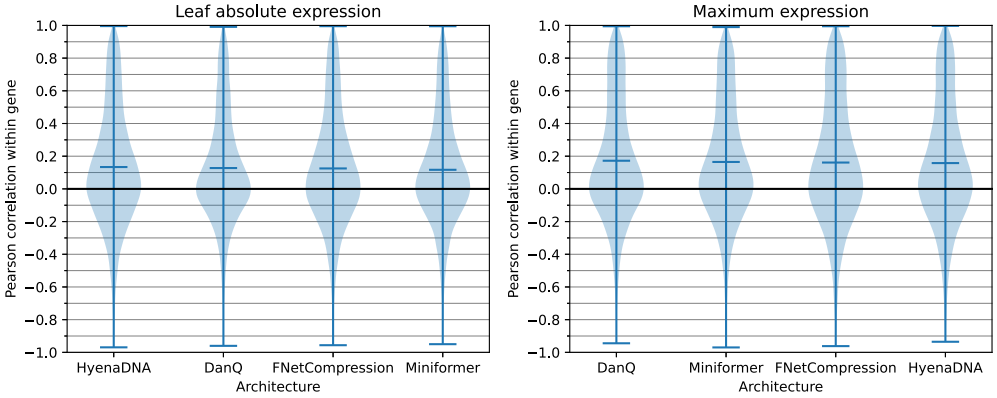
Supplemental Material



Supplemental Figure 1: Pearson correlation in the regression tasks across all genes and data splits. Each plot shows the performance of all architectures on a single task. Error bars represent one standard deviation from the mean in each direction.



Supplemental Figure 2: DanQ predictions on the test set across all tasks. The model from the first training run was used for predictions. Regression tasks (top) are on the log scale. Color in the regression task histogram scatterplots represents the number of observations within that area.



Supplemental Figure 3: Distributions of Pearson correlation within orthogroups for each task. Architectures are sorted from highest (left) average within-orthogroup Pearson correlation to lowest (right). Bars within the violins represent the mean of the distribution.