Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep learning the *cis*-regulatory code for gene expression in selected model plants

Fritz Forbang Peleke
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) https://orcid.org/0000-0001-7394-5994

Simon Maria Zumkeller
zumkellersimon@gmail.com

Cluster of Excellence on Plant Sciences (CEPLAS), Forschungszentrum Jülich https://orcid.org/0000-0002-1630-2439

Mehmet Gültas
Faculty of Agriculture, South Westphalia University of Applied Sciences

Armin Schmitt
Center of Integrated Breeding Research (CiBreed), Göttingen

Jędrzej Jakub Szymański
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) https://orcid.org/0000-0003-1086-0920

Research Article

Keywords: machine learning, deep learning, gene regulation, regulatory networks, genetic variation, genome annotation, transcriptomics, RNA sequencing, gene model, Solanum lycopersicum

Posted Date: May 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2873437/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Nature Communications on April 25th, 2024. See the published version at https://doi.org/10.1038/s41467-024-47744-0.

Abstract

Elucidating the relationship between the sequences of non-coding regulatory elements and their target genes is key to understanding gene regulation and its variation between plant species and ecotypes. In this study, we developed deep learning models that link gene sequence data with mRNA copy number for the plant species *Arabidopsis thaliana, Sorghum bicolor, Solanum lycopersicum* and *Zea mays*, and predicted the regulatory effect of gene sequence variation. Our models achieved over 80% accuracy in the species-specific and multi-species prediction tasks and enabled predictive feature selection within the input regulatory sequences. Saliency scores of the model highlighted a set of expression-predictive sequence features and the profound importance of the UTR regions in determining the level of gene expression. Identified sequence features exhibited remarkable conservation across plant species and achieved more than 70% accuracy in cross-species expression prediction. We demonstrated the application of our model on 14 newly assembled tomato genomes, where the effect of structural genetic variation on gene expression is annotated. Finally, we showed that by providing an accurate prediction of differences in the expression of biosynthetic enzymes and their individual homologs, the model highlights known metabolic differences between related genotypes. This was demonstrated for biosynthetic pathways of stress-related compounds in *Solanum lycopersicum* and its wild drought-resistant relative *Solanum pennellii*.

Introduction

Regulation of gene expression relies on the complex interaction of proteins and DNA[1]. One of the key mechanisms is the control of transcription by *cis*-regulatory elements (CRE), which are short DNA sequence motifs within the gene's proximal genomic region recognised by transcription factors (TF) [2]. Additionally, RNA processing determines the turnover of transcripts. This includes intron-splicing [3], stabilisation of RNAs by mRNA-capping [4] and poly(A)-tailing at the RNA 3' end [5]. Accordingly, the regulation of gene expression is mediated by a nucleotide sequence code specifically bound by protein factors. However, experimental molecular biology techniques and enrichment-based methods reveal only a portion of that nucleotide code, as they are reductionistic by default.

There are different experimental methods available to study the interactions of proteins and nucleic acids. These require profound characterisation of, *e.g.*, TFs and establishment of specific methods for each new organism. Chromatin immunoprecipitation (ChIP) sequencing has been most used to determine protein-DNA interactions of TFs [6]. The identification of putative CREs is limited in ChIP-sequencing, however, because chromatin structure and protein-protein interactions may prevent TFs from binding and, consequently, limit the identification of CREs [7–9]. Other methods like *in vitro* microarray binding assays can determine TF-DNA interactions specifically, but their application is technically restricted to well-characterised model species [10]. If established for the organism of interest, such methods are well suited to demonstrate the distinct interactions of TFs and CREs. Linking these interaction studies to gene expression, however, requires further experimentation.

With the increasing availability of genomic data, machine-learning algorithms can be employed to annotate and functionally characterise CREs [11,12]. To reach this goal new integrative approaches capable of systematic investigation of the sequence-to-regulation relationships across multiple plant species and regulation domains are needed [13]. Among these, deep learning (DL) represents the most versatile and increasingly powerful framework that recently enabled significant breakthroughs in *e.g.*, computer vision [14], natural language processing [15] and protein structure prediction [16]. Among DL models, convolutional neural networks (CNN) are particularly effective in the classification, segmentation and feature extraction of image data and have already been successfully implemented on genetic sequence data. For instance, the Enformer model combined convolutional with multi-head attention layers in a neural network for improved integration of the effects of the distal and proximal regulatory elements in human genomics sequences and efficient gene expression prediction [17]. In *Zea mays*, the expression of genes was modelled as a binary classification task, pointing out the potential overfitting of models to gene family-specific features [18].

Despite these methodological advances, the genome-scale identification and annotation of *cis*-regulatory sequence features across multiple plant species is still an uncharted area. Therefore, in this study, we aim at the systematic identification of gene regulatory sequences and annotation of their function in terms of their effect on gene expression. We investigate the conservation of regulatory sequences and their function across the four model plant species *Zea mays*, *Arabidopsis thaliana*, *Solanum lycopersicum* and *Sorghum bicolor*. We classify gene expression states using a large resource of RNA-seq data with a) convolutional neural networks that scan raw sequences as automated motif extractors and b) random forest classifiers trained on generic features generated based on the reference gene models. Finally, we demonstrate the application of our models for functional annotation of genetic variation and metabolic pathway activity prediction in domesticated and wild tomato accessions.

Results

Data resource

In the first step, we have generated a data resource serving as an input for training the model. For that purpose, we used the genome assemblies and annotations of *Arabidopsis thaliana*, *Solanum lycopersicum*, *Zea mays* and *Sorghum bicolor* from the Ensembl Plants database (v52). For each gene of each species 1000 nt upstream and 500 nt downstream of the transcription start site (TSS) and 500 nt upstream and 1000 nt downstream of the transcription termination site (TTS) were extracted (Fig. 1a). This provided 3000 nt flanking sequence per gene that included the close promoter, UTRs and the gene downstream region. These sequences were subsequently one-hot encoded to serve as input to the CNN model. Accordingly, for each species we selected and remapped a set of publicly available RNA-seq profiles obtained for leaf tissue in multiple independent experiments **(Supplementary table 1)**. Estimated gene expression levels were then classified as low, medium or high based on the lower and upper quartiles of the distribution of the log-transformed transcript per million values (logMaxTPM) **(Fig. 1b)**.

Model architecture and training strategy

Our approach builds on the CNN model architecture proposed previously for the pseudogene model [18]. Besides the modified training strategy (see Materials and Methods), a notable difference concerned the use of 1D convolutional and Maxpool layers in contrast to the 2D convolutional layers used by Washburn et al. (2019; **Fig. 1c**). The three convolutional blocks, each composed of two convolutional layers, have proven to efficiently capture sequence features of different scales and complexity within the 3000 nt flanking sequences. The information extracted by the first convolutional layer is integrated by subsequent convolutional layers into more complex representations that are used for predictions in the final output layer. To train the CNN model, we focused our analysis on genes below and above the lower and upper quartiles of the logMaxTPM distribution, respectively. Accordingly, the CNNs were trained as binary classifiers to predict genes as either lowly (below the lower quartile) or highly (above the upper quartile) expressed. Dividing the data into more classes as well as training a regression model resulted in considerably lower accuracies. CNNs were trained using chromosomal level cross-validation, in which for each iteration genes located on one of the chromosomes were used as validation set and the rest for training. To prevent overestimation of the model performance due to overfitting to evolutionary relatedness, genes in the validation set with homologs in the training set were omitted during the estimation of model performances.

Predicting plant gene expression from the sequence of gene flanking regions – model performance

In the first step, we trained the CNN models per species with n = #chromosomes cross-validation rounds. The use of chromosomes for splitting the validation set ensured that the set was not overlapping with any of the training sequences and was similar in terms of chromatin accessibility. These single-species reference (SSR) models achieved average accuracies of 78.39% (auROC = 0.85) for *Zea mays*, 83.79% (auROC = 0.91) for *Arabidopsis thaliana*, 77.96% (auROC = 0.84) for *Sorghum bicolor*, and 78.39% (auROC = 0.86) for *Solanum lycopersicum* (Fig. 2a). This result showed the extent of flanking sequence-determined gene expression in leaves that could be learned by a relatively simple CNN and thus likely represents rather an underestimation of the real value. The varying size of the training sets between SSR models did not directly reflect the model performance, indicating that it is not a major limiting factor. Next, we trained multi-species reference (MSR) models using a leave-one-out procedure; each time training the model on three of the species and evaluating it on the fourth. The MSRs showed a similar performance as the species-specific SSRs, showing that the predictive sequence features are fairly conserved (Fig. 2a). The shuffled-sequence controls (SSC) retaining the nucleotide composition exhibited close to random classification performance for all species, highlighting the existence and the role of localised sequence features as gene expression predictors. Finally, we tested the SSR models in terms of their prediction accuracy in species they were not trained on, in what we called a cross-species prediction (Fig. 2b). The best cross-species prediction accuracy of 78.86% was observed for the *Sorghum bicolor* SSR model using as input *Zea mays* sequences. The worst performance of 65.81% was

observed for the *Sorghum bicolor* SSR model predictions on *Solanum lycopersicum* sequences **(Supplementary Table 2).**

The significant drop in cross-species prediction performance in comparison to the SSR and MSR models indicated that the gene expression in each species is generally determined by two classes of sequence features: a) species-specific, and b) conserved across multiple species. While the SSR models learn the combination of them and are hardly generalizable across species, the MSR models seem to capture the conserved features and generalise well.

Identification and characterization of predictive sequence features

Identification of the location and specific sequence features associated with gene expression requires interpretation of the trained models. Here, we used the Deep Learning Important Features (DeepLIFT) algorithm, which enables the interpretation of convolutional neural networks trained on genetic sequences by providing nucleotide-resolution contribution scores across input sequences [19]. The scores are calculated by backpropagating the contribution of each neuron of the trained model to each nucleotide of the input and thus take the positive, negative, or zero values depending on the impact of the nucleotide on the predicted values: high expression, low expression, and irrelevant, respectively. Averaging obtained scores across the input sequences revealed conserved patterns. We observed that for all species the most salient regions of the flanking sequences were those proximal to the TSS and TTS containing the 5'UTR and 3'UTR, respectively (Fig. 3a). The salient regions received predominantly positive scores indicating the presence of expression-associated features, rather than those with deleterious effects. While the general distribution of DeepLIFT scores was conserved across the species, some differences also occurred. *Solanum lycopersicum* exhibited a much higher importance of the TSS upstream region than the other species. *Zea mays* on the other hand was the only species exhibiting low saliency of the sequences downstream of the TTS. In contrast, *Sorghum bicolor* exhibited the importance of the TTS downstream sequence features for predicting low gene expression.

The DeepLIFT saliency scores were further used to extract expression-predictive sequence motifs using the Transcription Factor Motif Discovery from the Importance Scores algorithm (TF-MoDISco) [20]. Extracted high-scoring subsequences (seqlets) and their summary contribution scores were averaged into contribution weight matrices (CWMs). The CWM motif encoding is analogous to the position weight matrix (PWM), however it represents the importance of nucleotides for the expression prediction rather than their frequency. Thus, we will refer to CWMs identified with the TF-MoDISco as Expression-Predictive Motifs (EPM) in the following sections. We observed that the distribution of seqlets used for the generation of EPMs showed a general specificity towards the 5'UTR and 3'UTR, with individual EPMs showing EPMs-to-region specificity (Fig. 3b).

Identification and characterization of predictive gene model features

While the interpretation of the models revealed a conserved pattern of associated saliency scores and the presence of specific EPMs, the contribution of general features of gene models, such as GC content, CpG content and lengths of UTRs is difficult to infer from the CNN model. Therefore, we generated ten generic features per gene and trained random forest (RF) classifiers for the SSR and MSR training strategy. Remarkably, our RF-SSR classifiers achieved accuracy values of 81.99% in *Arabidopsis thaliana*, 86.82% in *Solanum lycopersicum*, 76.21% in *Sorghum bicolor*, and 79.66% in *Zea mays* **(Fig. 4a)**. We observed a lower performance for the RF-MSR models with accuracy values of 67.77%, 64.91%, 72.84% and 78.48%, respectively **(Fig. 4a)**. The impurity-based feature contributions calculated for the RF-SSR model indicated that the UTR-based features were the most important for classification with the lengths of the 3' UTRs being the most important feature across all species **(Fig. 4b)**. 5'UTR length showed higher importance for genes of *Arabidopsis thaliana*, while 3'UTR GC content was highly predictive in genes of *Solanum lycopersicum*. The impurity-based contributions for the RF-MSR model also showed the importance of the UTR-based features, with the length of the 3'UTR being the most important feature **(Fig. 4c)**.

Expression predictive motifs (EPMs) exhibit sequence and positional conservation

In total, we have identified 261 EPMs in total (522 with reverse complement versions). This includes 35 EPMs in *Arabidopsis thaliana*, 29 EPMs in *Sorghum bicolor*, 26 EPMs in *Solanum lycopersicum* and 46 EPMs in *Zea mays* for the SSR models and 28, 34, 30 and 33 for the MSR models, respectively **(Supplementary Table 3)**. The EPMs were divided into two meta-clusters associated with the prediction of either high or low rates of gene expression and labelled as 'p0' (136 EPMs) and 'p1' (125 EPMs), respectively. Hierarchical clustering of the 522 EPM using the Smith-Waterman algorithm identified 19 multi-species clusters of homologous motifs **(Fig. 5a)**. These 19 clusters included 474 of 522 EPMs, while 48 EPMs remain unique **(Fig. 5a; Supplementary Fig. 1)**. The smallest and largest clusters include 8 and 80 different EPMs, respectively. In the following sections, we name these clusters after their consensus sequence according to the IUPAC nucleotide code with the minimum number of repetitions as a prefix. For example, cluster 5W+ (WWWWW) features EPMs with at least five or more (+) nucleotides with weak complementary binding (A or T); or cluster 3SSN + is named after commonly three times repetitive base triplet composed of two strong (S) binding nucleotides, like G or C followed by an ambiguous nucleotide (N) **(Fig. 5b)**. Seven EPM clusters had multiple significant matches to previously characterised TF binding sites (e-value < 0.5) according to the JASPAR database of *cis*-regulatory elements [21] **(Fig. 5a, Supplementary Table 3)**. Other EPM exhibit similarity to known structural gene elements, such as the U and A-rich regions involved in the selection of mRNA poly-adenylation sites [22] and hexanucleotide U repetitions acting as terminators for the nuclear RNA Polymerase III [23].

The EMP cluster 2CWY + is most similar to the TF binding site of A. thaliana AGL42

All four MSR models identified EPMs belonging to the 2CWY + cluster as highly contributing to the prediction of low gene expression. In all four analysed species, 2CWY + EPMs exhibit a strict positional

preference for the transcription start site (TSS and TSS-downstream) and the transcription termination site (TTS and upstream-TTS) **(Fig. 5c)**. The high conservation of the sequence, the positional preference and uniform prediction towards low rates of transcription suggest an evolutionary conservation and regulatory function of 2CWY-EPMs. In addition, epm*Zema*-M047-p1m09−513 and epm*Arth*-M045-p1m03−868 display two variants of the 2CWY + cluster, where after two CAT repetitions, a guanine, instead of cytosine is found prevalently. As an exception, the SSR model in *S. lycopersicum*, 2CWY + type epm*Soly*-S027-p1m00−7754 was identified with positional preference to both down- and upstream of the TSS and TTS, respectively. However, 2CWY + EPMs in *S. lycopersicum* were uniformly predictive for low gene expression indicating the function of that EPM is not necessarily position-specific. In total, 10 of 11 EPMs of the identified 2CYW + cluster exhibited such conserved position-independent effects on the gene expression **(Supplementary Fig. 1)**.

The EPMs of the 2CWY + clusters, *e.g.*, epm*Arth*-S043-p1m01−3124 (CATCAT), significantly match the TF binding site (CAYCAT) of AGL42 (also FYF, FOREVER YOUNG FLOWER). AGL42/FYF is a MADS-box transcription factor of the evolutionary eukaryotic old MIKC class closely related to the SUPPRESSOR OF OVEREXPRESSION OC1 (SOC1) which is a gene central to the development of the flower, particularly mediated by gibberellin signalling [24,25]. AGL42 (FYF) is interacting with SOC1 and is expressed in the shoot apical meristem during the vegetative phase and during the floral transition where it promotes flowering at the shoot apical and axillary meristems [25]. Here, AGL42 (FYF) actively controls floral senescence/abscission by repressing ethylene responses together with a regulatory network of *FYF*-like genes [26–28].

The EPM clusters 2GCB + and 2CT + are likely TF binding sites but predict low and high gene expression dependent on their position in relation to the TSS and TTS.

In contrast to 2CWY+, clusters such as 2GCB+, 2CT + and YACA2 feature position-dependent effects **(Fig. 5d)**. For example, eight of thirteen 2GCB + EPMs found in *A. thaliana, S. bicolor* and *Z. mays* that are located within the transcript window consistently predict high transcript levels **(Fig. 5d, Supplementary Fig. 1)**. Conversely, five 2GCB + EPMs of *S. bicolor* and *Z. mays* that are prevalent upstream of the TSS and downstream of TTS, robustly predict low rates of gene expression. The 2GCB + EPM cluster has neither been identified by the SSR nor MSR models in *S. lycopersicum* **(Fig. 5d).** The EPMs of 2GCB + show significant sequence similarity to the GCC box TF binding sites of the ERF/DREB class of ABA-responsive WRKY-TFs in *Zea mays* [29]. In tomato, the ERF/DREB TFs binding to GCC boxes is a key element of fruit development and ripening signalling [30]. Thus, our results are not indicative of the lack of the functional motif in the *S. lycopersicum* genome, but rather of its functional specificity. Another EPM cluster with a position-dependent effect is 2CT+, a motif exclusively identified by the *A. thaliana* SSR model and homologous to the GA/CT-repeat binding motifs of the BASIC PENTACYSTEINE (BPC5/BBR/BPC) transcription factor family [31]. The 2CT + motifs identified in our study are either localised upstream of the TSS and predict high expression or are localised downstream of the TSS and predict low expression.

CNN models identify perturbations in gene expression associated with structural variance among sub-species and varieties of tomato

To check if our model predicts differences in expression of gene variants, we used a large collection of genomic structural variants (SVs) for wild and cultivated *Solanum spp.* [32]. Due to the application of long-read nanopore sequencing the dataset contained variation data ranging from single nucleotide polymorphisms to large transposon insertions across 15 *Solanum spp*. genotypes.

We used the MSR models trained with *Solanum lycopersicum* as the validation set to generate predictions for a subset of diverse varieties and their genotypes of *Solanum pimpinellifolium* (PAS014479, BGV006775), *S. lycopersicum* var. cerasiforme (BGV006865, BGV007931, BGV007989), *S. lycopersium* processing (M82), *S. lycopersicum* fresh (EA00371, Fla.8924, LYC1410),*S. lycopersicum* vintage (PI69588, Brandywine, EA00990, PI303721) and *S. lycorpersicum* ITAG 3.0 for reference **(Fig. 6a).** Neighbour joining-based clustering of the input regulatory sequences for the 14 genotypes showed identical topology with the hierarchical tree of the predicted expression profiles and remained largely in agreement with the SV-based phylogeny from Alonge and colleagues **(Fig. 6a).** We further investigated orthologous genes that were likely to vary in expression across the genotypes according to predictions of the MSR models and where differential gene expression was shown by Alonge and colleagues **(Supplementary Table 4)**.

We chose exemplarily genes of *Solyc02g08170.4* encoding a putative phosphatidylinositol glycan, class N protein and *Solyc02g080300.3* encoding a glycosyl hydrolase ***β***-glucosidase 45-related proteins. Both genes feature SVs in their TSS and TTS regions accompanied by differential gene expression **(Supplementary Table 4)**. Further, probabilities of the predicted expression patterns of those genes were independent of their phylogenetic relationship between the cultivars **(Fig. 6b, e)**. To characterise the SV-associated change in gene expression, we mapped motifs from the MSR and SSR models of *S. lycopersicum* back to the two genes using BLAMM [33] **(Supplementary Table 6)**. There were 5 and 4 EPMs, respectively, significantly matching (e-value < 0.0001) the TSS regions of the *Solyc02g087170* and the *Solyc02g080300* **(Fig. 6c, f)**. However, only 2 and 1 EPMs, for each gene respectively, were identified in their preferred ranges inferred from the CNN. Any other mapped EPM lying outside their preferred range appeared to not influence the probability scores.

According to the probability of gene expression, the gene *Solyc02g08170.4* is predicted to exhibit high (0.8) and moderate (0.5) gene expression in two distinct groups of *Solanum* genotypes. On the sequence level this difference coincides with a shared 178 bp indel mutation 38 bp downstream of the TSS among 5 of the 15 *Solanum* genotypes, all predicted for high rates of expression (PAS0144795, BGV007989, BGV006865, EA00990 and EA00371) **(Fig. 6d)**. Interestingly, disrupting epm*Soly*-S008-p0m03F via the SV entails the creation of epm*Soly*-S015- p0m07 now positioned within its preferences range 226 bp downstream of the TSS within *Solyc02g08170.1* 5'UTR **(Fig. 6d)**. This indicates that the CNN predictions depend on the identification of EPMs at their positional optimum and that these are weighted differently.

In the case of *Solyc02g080300* an SV indel mutation occurred in the first intron and the 3'UTR of the genotypes PAS014479, BGV06775 and BGV006865 sharing high probability scores compared to the other species **(Fig. 6e, f)**. Twenty nucleotides upstream of the SV within the first intron two point mutations were found in PAS014479, BGV06775 and BGV006865, too, hereby matching epm*Soly*-S007-p0m03 **(Fig. 6g)**. This EPM predicts high rates of gene expression and is located downstream of the TSS. Accordingly, its presence explains the high prediction of gene expression for *Solyc02g080300* along the 3 of the 15 genotypes. The probability for higher rates of gene expression is similar to the one in *Solyc02g087170* where epm*Soly*-S007 could also be identified in its reverse complement orientation.

The prediction of the CNN can be interpreted using DeepLift, MoDisco and BLAMM. As a result, we retrieved EPMs that can potentially be used for the prediction of rates of gene expression and characterisation of promoter or terminator regions. The CNN, as interpreted using the EPMs, is sensitive to point and indel mutations. In addition, the EPMs' predictive power is likely weighted differently, as the probabilities differ depending on the identified EPMs.

MSR model highlights known phenotypic and metabolic differences between S. lycopersicum **and** S. pennellii

In the next step, we have evaluated if the MSR model highlights interpretable differences between plant species that were not used to train it. For this purpose, we selected the cultivated tomato *S. lycopersicum* and its wild relative *S. pennellii* as examples of highly related species with remarkable differences in phenotype and metabolism. The model, trained on a combined data set from *A. thaliana*, *Z. mays* and *S. bicolor* achieved accuracy values of about 78% and 83% for *S. lycopersicum* and *S. pennellii*, respectively. The model performance was different for the negative and positive labels. For *S. lycopersicum* a true positive rate (TPR) of 0.92 and a true negative rate of 0.7 (TNR) indicated higher confidence for the positive predictions. This is also reflected in the distribution of the expression predictions for genes of the "middle" second and third quartiles (Fig. 7a; yellow distributions). The same was observed for *S. pennellii*, with a TPR of 0.85 and a TNR of 0.82 (Fig. 7b).

In order to estimate if these MSR predictions are indicative of physiological and metabolic differences between the two species, we performed Mercator4 MapMan annotation of functional gene ontology on the *S. lycopersicum* and *S. pennellii* proteome [34]. Then, we expressed each functional category in terms of the observed (true labels) and predicted expression (prediction probabilities) of genes classified into each functional category. In total, 90 MapMan categories exhibited a significant change of the predicted expression between *S. lycopersicum* and *S. pennellii* (Wilcoxon rank-sum test; FDR < 0.05; **Fig. 7c**; **Supplementary table 5**). These included general terms such as "enzyme" (3856 transcripts) and "transcriptional regulation" (1829 transcripts), as well as very specific ones, such as "ATP synthase complex" (14 transcripts) or "circadian clock system" (36 transcripts). The observed and predicted expression differences between *S. lycopersicum* and *S. pennellii* were correlated (Spearman's ρ = 0.46; p-value = 5.062e−06). This indicated that the MSR model successfully captures a differential expression of functional gene sets in the two tomato species. Almost all of the significant categories were observed

and predicted to be up-regulated in *S. pennellii* with respect to *S. lycopersicum*. This concerns in particular phytohormone action, external stimuli response, polyamine, lipid and flavonoid metabolism, as well as protein and RNA biosynthesis, and transcription regulation with their multiple sub-categories. Expression of several MapMan categories was overestimated by the MSR model, including *e.g.*, chromatin structure, cell wall proteins, and some TF families that indeed are very unlikely to be found in the upper quartile of expressed genes. Conversely, no MapMan categories were falsely predicted to be down-regulated. The only category significantly down-regulated in *S. pennellii* was the general term “not annotated / not assigned”.

Finally, we zoomed in on enzymes of polyamine metabolism, as metabolic pathway that has been observed and predicted to be significantly up-regulated in *S. pennellii*, and supported by previous observations of strongly differential polyamine accumulation in *S. pennellii* and *S. lycopersicum* leaves [35] (Fig. 7d). Out of 40 genes annotated in S. lycopersicum 10 were observed as highly expressed, covering each of the 10 reactions with at least one highly expressed isozyme transcript except the *ornithine decarboxylase ODC.* Remarkably, six of these were predicted correctly to be highly expressed and no false positives were recorded, resulting in a prediction precision of 1 for the high expression. At the same time out of six labelled as low expressed all six were predicted to be low expressed with four false negatives (precision = 0.6). Summarising, all pathway enzymes in *S. lycopersicum* were predicted to be highly expressed as at least one isozyme, except for the ODC which has not been observed as highly expressed in the corresponding RNA-seq data either. Analogous results were obtained for *S. pennellii*, with the precision of high expression prediction of 1 (each positive expression prediction associated with respectively high or medium expression level) and precision for low expression of 0.66. Remarkably, despite the high sequence homology to *S. lycopersicum*, the *S. pennellii* orthologs showed distinct patterns of gene expression and respective predictions. This is especially visible for ADC, SAMDC, SPDS, peroxisomal PAO and the SSAT enzymes. Additionally, ODC has been both observed and predicted to be highly expressed in *S. pennellii*, while AIH showed lower expression. In summary, the MSR model correctly indicated the polyamine biosynthetic pathway as highly expressed in both species, identified with high precision the isozymes that are most likely highly expressed for each reaction, and finally highlighted differences between *S. lycopersicum* and *S. pennellii* in terms of expression of enzymes and their homologues.

Discussion

In this study we presented a new simple approach to the challenge of plant cistrome annotation. In contrast to other approaches *e.g.*, screening for a very broad range of protein-DNA interactions [12], identifying specific TF-DNA binding events [36,37], or extracting motifs enriched in genes exhibiting a stimulus response [38], here we focused on a very specific aspect of cistrome function, namely the determination of the mRNA copy number, and we have built a mathematical model for robust and accurate prediction of it from the DNA sequence information.

There are several important features differentiating our approach from those used in most other studies on cistrome annotation. First, our study represents a genome-wide identification of predictive features related to the mRNA turnover and not those associated with the regulation of gene expression in response to any stimulus. Additionally, while we highlighted some key DNA-TF binding events, a wider range of functional sequence features at the level of DNA and mRNA (e.g., polyadenylation signals) was captured. In fact, most of the identified features were located in close proximity to the TSS and TTS sites including gene UTRs. These regions are well known to be associated with mRNA turnover [39,40] and were recently targeted by synthetic biology approaches proving their central role in gene regulation, as e.g, plant biotechnology chimeric 3' flanking regions strongly enhance gene expression in plants [41,42]. Finally, due to the chosen training strategy, we avoided capturing sequence features conserved in all gene models (e.g., translation start and termination codons or classic signals for transcription start; [43]) but recovered only those predictive for high or low gene expression. While this provides a wealth of valuable and novel information, it is difficult to define a ground truth reference for our results; *e.g.* many structural and regulatory elements of the gene model are well-characterised and relatively easy to annotate [44,45], this is not yet the case for the turnover-related regulatory features [4].

Indeed, the majority of the identified sequence features (12 of 19 EPM clusters) were not matching with previously described CREs despite their clear sequence and positional specificity. This has likely three main reasons. First, the incompleteness of the reference databases; the cognate TF-binding motifs might be uncharacterised or not included. Second, due to *e.g.*, model regularisation, highlighted predictive sequences often represent only partial CREs. For example, important but redundant sequence features, *e.g.*, palindromic and near-palindromic repeats, might be only partially captured. Finally, some EPMs were not TF-binding motifs but likely *e.g.*, intron-, transposon- and UTR-associated sequence features, that are left out on prominent CRE databases.

The achieved model performance is rather striking taking into account the complexity of plant gene regulation [46] and the relatively small training set, especially for the SSR models. In comparison to more complex models like the Enformer, which capture long range interactions by scanning sequences longer than 100kb, our model captures only the information present in the proximal *cis*-regulatory sequences. Both the SSR and MSR models achieved an accuracy of around 80% and generalised well across the species. This indicates that the *cis*-regulatory landscape influencing the gene expression level in plants is relatively simple and very conserved. This is in agreement with the conservation of TF-binding sites in *Arabidopsis* ecotypes [47]. While the stimulus-induced changes in gene expression rarely translate directly to changes in the accumulation of respective proteins, the mRNA turnover and the related amount of mRNA in the cytosol remains a key factor determining the general amount of respective protein product [48–50]. Thus, the determinants of the mRNA level are important elements of the system homeostasis and are unlikely to exhibit high genetic variation [51,52]. The extent of that conservation is reflected by a comparison of the MSR model with the cross-species SSR performance. Increasing the training set for multiple species enabled the MSR models to capture that conserved regulation component and generalise very well to new genomes. The relative simplicity of the *cis*-regulatory landscape determining

mRNA copy number was further supported by results obtained using the RF approach. The RF classifiers achieved comparable performance to the CNN models for SSR binary classification and highlighted the known role of general UTR properties such as UTR length or GC content [42]. While being a simple and accurate alternative to the deep learning approach, the RF model is sensitive to low-quality gene annotations, especially those of the UTR regions, which significantly limits the applicability of RF models to the newly sequenced genomes and draft annotations.

One advantage of the parallel sequence feature and function annotation is the identification of the positional specificity of the EPM occurrence and function. In general, CREs may overlap with genic features like the UTRs, exons or introns, or are located distantly to the up- and downstream of the TSS or TTS, respectively [53]. In *Arabidopsis thaliana*, for example, several transcription factors are known to bind specifically at around 100 bp upstream of the TSS around the promoter [54]. Other studies show that certain CREs may preferentially occur in the first intron of *A. thaliana* genes [55]. While the ratio of CREs to genes correlates, the distances in between likely correlate with the overall genome sizes due to mobility transposons or accumulation of repeats [56]. Previous experiments on reporter genes and comparative genomics in *A. thaliana* and *Z. mays* demonstrated changes in gene regulation for transcription factors binding sites situated within introns, where the distance between cognate CRE and the TSS was changed [57].

We demonstrated two EPMs that exhibit consistent positional preference and retained regulatory function, while also displaying variable positional preference that correlates with their regulatory activity across diverse flowering plant model organisms. In all four analysed species, *A. thaliana, S. bicolor, S. lycopersicum* and *Z. mays*, 2CWY + type EPMs exhibit strict positional preference of the transcription start site (TSS and TSS-downstream) and the transcription termination site (TTS and upstream-TTS). The 2CWY + cluster is strictly associated with the prediction of low gene expression, too. EPMs of the 2GCB + cluster, however, identified in *A. thaliana, S. bicolor* and *Z. mays* do occur rather up- or downstream of the TSS and TTS region, respectively or, in contrast, within the transcribed region, with conversely regulatory effects **(Fig. 5d)**. These findings indicate that the linear distance of CREs to their cognate gene plays a defined role in gene regulation. Interestingly, it has also been demonstrated that most similar to EPMs 2CT+, CT(GA)-binding motifs of paralogous transcription factor BPC6 are localised at the TSS in barley (*Hordeum)* genes too [58]. These TFs are responsive to cytokinin stimulus [58], congruent with findings about BPC6 in response to brassinosteroids in *A. thaliana* [31]. Somehow contradictory to our findings, studies of paralogous transcription factor BPC1 in *A. thaliana* showed that by binding its cognate motif (GA-rich region) situated in a 5'-terminal intron is up-regulating gene expression [59]. Likely, the EPMs identified here do not cover all regulatory elements and additional EPMs associated with more specific types of expression together with already characterized ones in other organisms will likely be identified in the future.

Our analyses on the occurrence of EPMs and associated prediction of gene expression throughout different *Solanum* genotypes further shows how the EPMs could be used for extended annotation and

characterization of genes **(Fig. 6b-g)**. If indel or point mutations are present within the identified EPM or if the EPM is located outside of its typical positional preference, the CNN model disregards them for prediction. Future characterization of EPMs should consider both their sequence and positional preference and their association with genomic features such as UTRs, introns, transposons, or repeats. The insertion of SVs within the *cis*-regulatory sequences has been associated with perturbations in gene expression as shown for *Solanum* cultivars [32]. Similarly, insertions of transposable elements (TEs) in Brassicaceae species could recently be associated with perturbations in gene expression associated with the transition from C3 to C4 photosynthesis [60]. Here, mining and characterisation of EPMs could greatly improve the identification of the regulatory determinants in such evolutionary processes.

We have also demonstrated that our approach is able to highlight some genotype-determined phenotypic and metabolic features of newly sequenced plants. The MSR model highlighted known differences between *S. lycopersicum* and *S. pennellii* in terms of the increased biotic and abiotic stress resistance [61–63], growth [64,65], cell wall [66] and specialised metabolism [67,68]. This is striking taking into account that the compared species are homologous to each other, in comparison to the *A. thaliana*, *Z. mays* and *S. bicolor* genomes that were used to train the MSR model. The result suggests that the observed differential expression of gene functional groups results from changes in a relatively limited pool of conserved regulatory sequences. Application of the MSR model to the metabolic pathway of polyamine biosynthesis, i.e., compounds with highly differential accumulation in *S. lycopersicum* and *S. pennellii* [35], showed that the approach identifies active pathways (e.g., those with multiple reactions represented by at least one highly expressed enzyme), selects the most likely expressed isozymes for a given reaction, and highlights differences between the expression of orthologous genes in *S. lycopersicum* and *S. pennellii.* While these differences do not have to be directly associated with the observed differential accumulation of polyamines, the high precision and accuracy of expression prediction indicate that the MSR model is an efficient tool for increasing the confidence in gene candidate selection for molecular validation [69]. Namely, enzymes that are observed and predicted to be highly expressed based on a wealth of data across several species, might be a safer choice for follow-up characterization than those that were observed to be expressed in a single experiment. Namely, enzymes that are observed and predicted to be highly expressed based on a wealth of data across several species, might be a safer choice for follow-up characterization than those that were observed to be expressed in a single experiment.

In summary, we showed that training a relatively simple deep learning model on publicly available large-scale genomic data provides a wealth of new biological information if combined with a proper model interpretation approach and set in an evolutionary context. This highlights the great potential of deep learning for high-throughput cistrome annotation in newly sequenced genomes, as exemplified by our MSR models. Another important application is the prediction of the effects of genetic variation [70]. While current genomics approaches rarely enable functional annotation of the effect of polymorphisms [71,72] and rely mostly on statistical association, an increasing availability of long-read data enables making mechanistic links between observed variation and the down-stream molecular network [32]. The presented

example of gene expression across 14 tomato accessions represents an important step in this direction and opens perspectives for routine prediction pipelines for new genomes, tissues and environmental scenarios. The ability to predict which genes are likely to be expressed and which are not might be also crucial for the selection of gene candidates in GWAS studies [68,73,74], or in the selection of functional enzyme homologues in reconstructed metabolic pathways [69]. While RNA-seq and metabolomic data are often used for that purpose [68, 75–77], a purely sequence-based approach will significantly accelerate the process.

Material and Methods

Reference genomes, proteomes, CDNA, gene models and data processing

The reference genomes, proteomes, CDNAs and gene models were downloaded from the Ensembl Plants database (v52). RNA-seq data was downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database using the fasterq-dump (version 2.11.3). These were trimmed with sickle (version 1.33), mapped unto reference CDNAs using Kallisto (version 0.46.2) and Kallisto outputs were processed using the tximport package of the R statistical language to obtain the normalised counts in Transcripts Per Million (TPM) **(Supplementary Table 1)**.

Upstream flanking sequences were extracted by anchoring at the gene start and end coordinates. Taking DNA strandedness into consideration, sequences from the negative strand were reverse-complemented. Since neural network models perform mathematical computations, sequences were one hot-encoded and combined into a single continuous sequence separated by a 20 nt zero-padding.

Producing nonhomologous training and validation sets

To produce nonhomologous training and validation sets, we constructed local protein databases using proteomes for every species. The proteomes were blasted against their respective databases (Protein-Protein BLAST 2.9.0+). Two sequences were considered homologous if the blast hit had an e-value < 0.001 and a bit score > 50. Genes in the validation chromosome that had homologous pairs in the training chromosomes were dropped.

Architecture of Convolutional neural networks and training strategy

Convolutional neural networks were built using the sequential API of tensorflow-keras. All models had three convolutional blocks, each convolutional block had two convolutional layers followed by a max pooling and a dropout layer.

The final convolutional block was followed by a fully connected block with two fully connected layers interspaced by dropout layers. All dropout layers used a dropout rate of 0.25, all convolutional and fully connected layers used Rectified Linear unit (ReLu) activation functions. The output layer contained a

single unit and the sigmoid activation function. Models were trained with a maximum number of 100 epochs. To further mitigate overfitting, we used the *EarlyStopping callback (patience = 10)*, the *ModelCheckpoint callback* was used to restore the best seen model with lowest validation loss. Finally, the *ReduceLROnPlateau callback (patience = 5)* was used to decrease the learning rate to improve model training.

Stochastic gradient descent as implemented by the *Adam optimizer (learning rate = 0.001)* was used to take gentle steps towards the lowest achievable binary cross entropy loss.

Model Interpretation

We computed hypothetical and actual contribution scores for all validation sets across all chromosomes. These scores are used in two ways: firstly, contribution scores were averaged across all nucleotides per sequence to obtain single scores per nucleotide and then averaged across all sequences to obtain a single representation of the most salient regions. Secondly, the hypothetical and actual contribution scores are used as inputs to TFModisco, which extracts the predictive regions from the contribution scores as seqlets. These seqlets are then clustered into groups and each group is aggregated into a representative motif. Due to memory and runtime constraints, we used the default value for the maximum number of seqlets per meta-cluster.

Clustering of EPMs and comparison JASPAR database

We retrieved expression predictive motifs (EPMs) using TFModisco that returns both Cluster weighted models and position weight matrices (PWMs). PWMs were characterised with R statistical language and the package JASPAR2022 calculates various properties of PWMs, such as information content and consensus sequence [21]. In addition, we searched for potential hits between EPMs and already characterised TF binding motifs on the JASPAR plants database **(Supplementary Table 3)**. With the algorithm implemented in the R package motifStack EPMscould be converted and clustered using the fisher-waterman algorithm [78] **(Supplementary Table 3)**.

Investigation of flanking gene sequences of Solanum lycopersicum **sub-variants**

Candidate orthologous gene selection followed the identification of predicted perturbations in gene expression of fourteen wild and cultivated Solanum *spp.* [32]. We selected 14 diverse genotypes for prediction: *Solanum pimpinellifolium* (PAS014479, BGV006775), *S. lycopersicum var. cerasiforme* (BGV006865, BGV007931, BGV007989), *S. lycopersium processing* (M82), *S. lycopersicum fresh* (EA00371, Fla.8924, LYC1410) and *S. lycopersicum vintage* (PI69588, Brandywine, EA00990, PI303721). Perturbations in gene expression of wild and cultivated *Solanum spp.* were predicted using the single-species-model CNN trained on Heinz 1607 genome [79] and ITAG 4.0 annotation. Top two orthologous genes with predicted differential gene expression were extracted from all analysed species and aligned with MAFFT [80] **(Supplementary Table 5)**. EPMs extracted from the SSR model and MSR model with TFMoDisco were mapped to the *Solanum spp.* gene alignments with BLAMM [33] **(Supplementary**

Table 6). Up- and downstream gene regions were investigated manually for the presence of structural variants, previously identified in the work of Alonge and colleagues, and EPMs identified in this work using alignment visualisation tool MEGA7 [81].

Functional annotation and pathway mapping

Functional annotation of *S. lycopersicum* and *S. pennellii* genes to MapMan functional categories was performed using Mercator4 [34] on ITAG 4.1 and the latest *S. pennellii* genome Schmidt and colleagues (2017)[82], as described in Mercator4[34]. Significance of the shift of median between *S. lycopersicum* and *S. pennellii* shown in Fig. 7c was performed by a two-sided Wilcoxon rank sum test with Benjamini-Hochberg p-value correction. Metabolic pathway mapping for polyamine biosynthetic pathway was done using MapMan “bin 8” and the super-pathway of polyamine biosynthesis POLYAMINSYN3-PWY from the current TomatoCyc annotation v6.0.0. All statistics and visualisation were done using R 4.2.2 and Bioconductor [83].

Declarations

Competing interests statement

The authors declare no competing interests.

Availability of data and materials

Custom code used in this study is available from the Github repository (https://github.com/NAMlab/DeepCRE). The raw fastq files used in this project can be retrieved from NCBI Sequence Read Archive under PRJNA319904, PRJDB8552, PRJNA503076 and PRJNA171684.

Funding

Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany
Simon Maria Zumkeller

Author contribution

F.F.P. performed bioinformatic analyses, designed and trained the CNN models, evaluated their performance and performed feature selection. S.Z. contributed to the characterisation of EPMs and motif search in *Solanum* cultivars, A.S. and M.G. consulted the results and corrected the manuscript. J.S. designed and supervised the study and performed the functional enrichment and pathway analysis.

F.F.P., S.Z. and J.S. designed figures and wrote the manuscript. All authors edited and approved the final manuscript version.

Corresponding author

All manuscript correspondence should be directed to Jedrzej Jakub Szymanski szymanski@ipk-gatersleben.de

Acknowledgement

The authors wish to thank CEPLAS for the financing of S.Z., IPK Gatersleben IT team for providing and maintenance of the computing cluster, and all members of the Network Analysis and Modelling lab at the IPK Gatersleben for discussions and suggestions.

References

1. Cooper, G. M. Regulation of transcription in eukaryotes. The cell: A molecular approach (2000).

2. Tafvizi, A., Mirny, L. A. & van Oijen, A. M. Dancing on DNA: kinetic aspects of search processes on DNA. Chemphyschem 12, 1481–1489 (2011).

3. Fesenko, I. et al. Alternative splicing shapes transcriptome but not proteome diversity in Physcomitrella patens. Sci. Rep. 7, 2698 (2017).

4. Sorenson, R. S., Deshotel, M. J., Johnson, K., Adler, F. R. & Sieburth, L. E. Arabidopsis mRNA decay landscape arises from specialized RNA decay substrates, decapping-mediated feedback, and redundancy. Proc. Natl. Acad. Sci. U. S. A. 115, E1485–E1494 (2018).

5. Bernardes, W. S. & Menossi, M. Plant 3' Regulatory Regions From mRNA-Encoding Genes and Their Uses to Modulate Expression. Front. Plant Sci. 11, 1252 (2020).

6. Whitington, T., Frith, M. C., Johnson, J. & Bailey, T. L. Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Res. 39, e98 (2011).

7. Gordân, R., Hartemink, A. J. & Bulyk, M. L. Distinguishing direct versus indirect transcription factor–DNA interactions. Genome Res. 19, 2090–2100 (2009).

8. Li, X.-Y. et al. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol. 12, R34 (2011).

9. Yan, J. et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell 154, 801–813 (2013).

10. Stormo, G. D. & Zhao, Y. Determining the specificity of protein–DNA interactions. Nat. Rev. Genet. 11, 751–760 (2010).

11. Akagi, T. et al. Genome-wide cis-decoding for expression design in tomato using cistrome data and explainable deep learning. Plant Cell 34, 2174–2187 (2022).

12. Savadel, S. D. et al. The native cistrome and sequence motif families of the maize ear. PLoS Genet. 17, e1009689 (2021).

13. Quang, D. & Xie, X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods 166, 40–47 (2019).

14. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI 31, (2017).

15. Vaswani, A. et al. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, (2017).

16. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

17. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods 18, 1196–1203 (2021).

18. Washburn, J. D. et al. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proceedings of the National Academy of Sciences vol. 116 5542–5549 (2019).

19. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. in Proceedings of the 34th International Conference on Machine Learning (eds. Precup, D. & Teh, Y. W.) vol. 70 3145–3153 (PMLR, 06--11 Aug 2017).

20. Shrikumar, A. et al. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv [cs.LG] (2018).

21. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research vol. 50 D165–D173 (2022).

22. Li, X.-Q. & Du, D. Motif types, motif locations and base composition patterns around the RNA polyadenylation site in microorganisms, plants and animals. BMC Evol. Biol. 14, 162 (2014).

23. Arimbasseri, A. G., Rijal, K. & Maraia, R. J. Transcription termination by the eukaryotic RNA polymerase III. Biochim. Biophys. Acta 1829, 318–330 (2013).

24. Pařenicová, L. et al. Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis : New Openings to the MADS World[W]. Plant Cell 15, 1538–

1551 (2003).

25. Dorca-Fornell, C. et al. The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. Plant J. 67, 1006–1017 (2011).

26. Chen, M.-K. et al. The MADS box gene, FOREVER YOUNG FLOWER, acts as a repressor controlling floral organ senescence and abscission in Arabidopsis. Plant J. 68, 168–185 (2011).

27. Chen, W.-H. et al. Regulatory network for FOREVER YOUNG FLOWER-like genes in regulating Arabidopsis flower senescence and abscission. Commun Biol 5, 662 (2022).

28. Chen, W.-H., Li, P.-F., Chen, M.-K., Lee, Y.-I. & Yang, C.-H. FOREVER YOUNG FLOWER Negatively Regulates Ethylene Response DNA-Binding Factors by Activating an Ethylene-Responsive Factor to Control Arabidopsis Floral Organ Senescence and Abscission. Plant Physiol. 168, 1666–1683 (2015).

29. Kimotho, R. N., Baillo, E. H. & Zhang, Z. Transcription factors involved in abiotic stress responses in Maize (Zea mays L.) and their roles in enhanced productivity in the post genomics era. PeerJ 7, e7211 (2019).

30. Zhang, L. et al. Function Analysis of the ERF and DREB Subfamilies in Tomato Fruit Development and Ripening. Front. Plant Sci. 13, 849048 (2022).

31. Theune, M. L., Bloss, U., Brand, L. H., Ladwig, F. & Wanke, D. Phylogenetic Analyses and GAGA-Motif Binding Studies of BBR/BPC Proteins Lend to Clues in GAGA-Motif Recognition and a Regulatory Role in Brassinosteroid Signaling. Front. Plant Sci. 10, 466 (2019).

32. Alonge, M. et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell vol. 182 145–161.e23 (2020).

33. Fostier, J. BLAMM: BLAS-based algorithm for finding position weight matrix occurrences in DNA sequences on CPUs and GPUs. BMC Bioinformatics vol. 21 (2020).

34. Bolger, M., Schwacke, R. & Usadel, B. MapMan Visualization of RNA-Seq Data Using Mercator4 Functional Annotations. Methods Mol. Biol. 2354, 195–212 (2021).

35. Santa-Cruz, A., Acosta, M., Perez-Alfocea, F. & Bolarin, M. C. Changes in free polyamine levels induced by salt stress in leaves of cultivated and wild tomato species. Physiol. Plant. 101, 341–346 (1997).

36. Chow, C.-N. et al. PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. Nucleic Acids Res. 47, D1155–D1163 (2019).

37. Bartlett, A. et al. Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat. Protoc. 12, 1659–1672 (2017).

38. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. & Van de Peer, Y. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. (2009).

39. Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z. & Zhu, J.-K. UTR-Dependent Control of Gene Expression in Plants. Trends Plant Sci. 23, 248–259 (2018).

40. Prall, W., Sharma, B. & Gregory, B. D. Transcription Is Just the Beginning of Gene Expression Regulation: The Functional Significance of RNA-Binding Proteins to Post-transcriptional Processes in Plants. Plant Cell Physiol. 60, 1939–1952 (2019).

41. Peyret, H., Brown, J. K. M. & Lomonossoff, G. P. Improving plant transient expression through the rational design of synthetic 5’ and 3' untranslated regions. Plant Methods 15, 108 (2019).

42. Diamos, A. G. & Mason, H. S. Chimeric 3’ flanking regions strongly enhance gene expression in plants. Plant Biotechnol. J. 16, 1971–1982 (2018).

43. Yandell, M. & Ence, D. A beginner’s guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342 (2012).

44. Li, Y., Shi, W. & Wasserman, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics 19, 202 (2018).

45. Holst, F. et al. Helixer—de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. bioRxiv (2023).

46. Ko, D. K. & Brandizzi, F. Network-based approaches for understanding gene regulation and function in plants. Plant J. 104, 302–317 (2020).

47. Korkuc, P., Schippers, J. H. M. & Walther, D. Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. Plant Physiol. 164, 181–200 (2014).

48. Edfors, F. et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Syst. Biol. 12, 883 (2016).

49. Jovanovic, M. et al. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science 347, 1259038 (2015).

50. Szymanski, J. et al. Label-free deep shotgun proteomics reveals protein dynamics during tomato fruit tissues development. Plant J. 90, 396–417 (2017).

51. Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Human gene essentiality. Nat. Rev. Genet. 19, 51–62 (2018).

52. Blanchette, M. & Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res. 12, 739–748 (2002).

53. Schmitz, R. J., Grotewold, E. & Stam, M. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell 34, 718–741 (2022).

54. Yu, C.-P., Lin, J.-J. & Li, W.-H. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci. Rep. 6, 25164 (2016).

55. Bradnam, K. R. & Korf, I. Longer first introns are a general property of eukaryotic gene structure. PLoS One 3, e3093 (2008).

56. Maher, K. A. et al. Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. Plant Cell 30, 15–36 (2018).

57. Sieburth, L. E. & Meyerowitz, E. M. Molecular dissection of the AGAMOUS control region shows that cis elements for spatial regulation are located intragenically. Plant Cell 9, 355–365 (1997).

58. Shanks, C. M. et al. Role of BASIC PENTACYSTEINE transcription factors in a subset of cytokinin signaling responses. Plant J. 95, 458–473 (2018).

59. Kooiker, M. et al. BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic Arabidopsis gene SEEDSTICK. Plant Cell 17, 722–729 (2005).

60. Triesch, S. et al. Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species. bioRxiv 2022.12.06.519256 (2022) doi:10.1101/2022.12.06.519256.

61. Smart, C. D., Tanksley, S. D., Mayton, H. & Fry, W. E. Resistance to Phytophthora infestans in Lycopersicon pennellii. Plant Dis. 91, 1045–1049 (2007).

62. Sharlach, M. et al. Fine genetic mapping of RXopJ4, a bacterial spot disease resistance locus from Solanum pennellii LA716. Theor. Appl. Genet. 126, 601–609 (2013).

63. Bolger, A. et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat. Genet. 46, 1034–1038 (2014).

64. Gur, A. & Zamir, D. Unused natural variation can lift yield barriers in plant breeding. PLoS Biol. 2, e245 (2004).

65. Eshed, Y. & Zamir, D. A genomic library of Lycopersicon pennellii in L. esculentum: A tool for fine mapping of genes. Euphytica 79, 175–179 (1994).

66. Caruso, G. et al. Exploring tomato Solanum pennellii introgression lines for residual biomass and enzymatic digestibility traits. BMC Genet. 17, 56 (2016).

67. Nunes-Nesi, A. et al. Identification and characterization of metabolite quantitative trait loci in tomato leaves and comparison with those reported for fruits and seeds. Metabolomics 15, 46 (2019).

68. Szymański, J. et al. Analysis of wild tomato introgression lines elucidates the genetic basis of transcriptome and metabolome variation underlying fruit traits and pathogen response. Nat. Genet. 52, 1111–1121 (2020).

69. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. 46, D633–D639 (2018).

70. Sarkar, D. & Maranas, C. D. SNPeffect: identifying functional roles of SNPs using metabolic networks. Plant J. 103, 512–531 (2020).

71. Liao, P.-Y. & Lee, K. H. From SNPs to functional polymorphism: The insight into biotechnology applications. Biochem. Eng. J. 49, 149–158 (2010).

72. Haleem, A., Klees, S., Schmitt, A. O. & Gültas, M. Deciphering Pleiotropic Signatures of Regulatory SNPs in Zea mays L. Using Multi-Omics Data and Machine Learning Algorithms. Int. J. Mol. Sci. 23, (2022).

73. Schaefer, R. J. et al. Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell 30, 2922–2942 (2018).

74. Zhao, H. et al. Multi-omics analysis reveals the mechanism of seed coat color formation in Brassica rapa L. Theor. Appl. Genet. 135, 2083–2099 (2022).

75. Sonawane, P. D. et al. Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism. Nat Plants 3, 16205 (2016).

76. Cárdenas, P. D. et al. Pathways to defense metabolites and evading fruit bitterness in genus Solanum evolved through 2-oxoglutarate-dependent dioxygenases. Nat. Commun. 10, 5169 (2019).

77. Luzarowska, U. et al. Hello darkness, my old friend: 3-KETOACYL-COENZYME A SYNTHASE4 is a branch point in the regulation of triacylglycerol synthesis in Arabidopsis thaliana. Plant Cell (2023) doi:10.1093/plcell/koad059.

78. Ou, J., Wolfe, S. A., Brodsky, M. H. & Zhu, L. J. motifStack for the analysis of transcription factor binding site evolution. Nat. Methods 15, 8–9 (2018).

79. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641 (2012).

80. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research vol. 30 3059–3066 (2002).

81. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution vol. 33 1870–1874 (2016).

82. Schmidt, M. H.-W. et al. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. Plant Cell 29, 2336–2348 (2017).

83. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004).

Figures

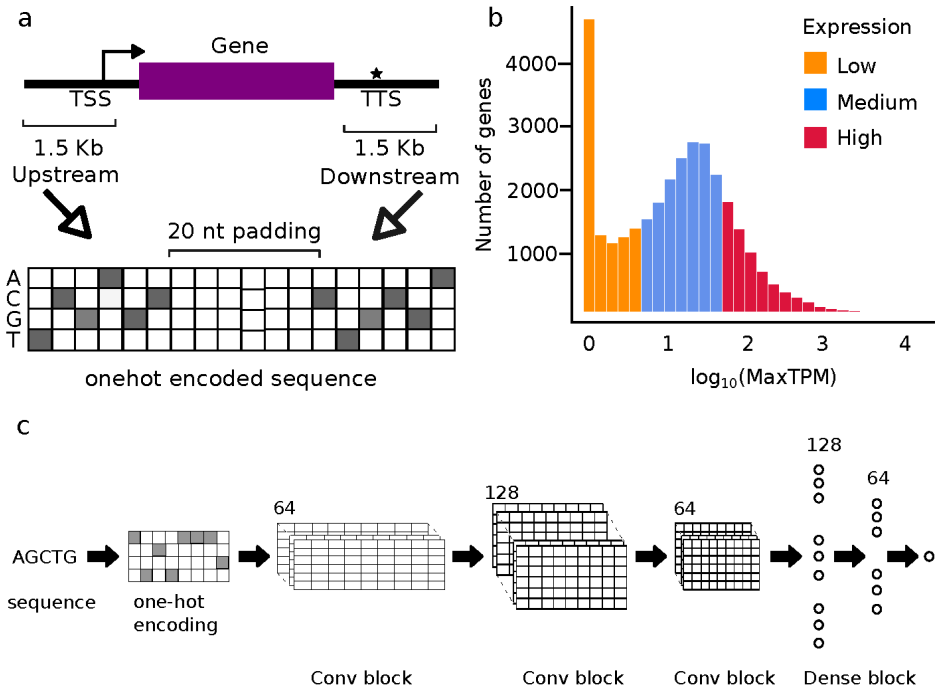


Figure 1

(a) DNA regions were extracted 1 kbp upstream and 0.5 kbp from the annotated gene start and end. Extracted sequences were converted into matrices by one-hot encoding, separated by a 20 nt padding.

(b) Genes were assigned into low (dark orange), medium (blue) and high (red) expression classes based on the logMaxTPM distributions.

(c) End-to-end depiction of model training. One-hot encoded sequences are fed into convolutional neural networks that predict the expression classes.

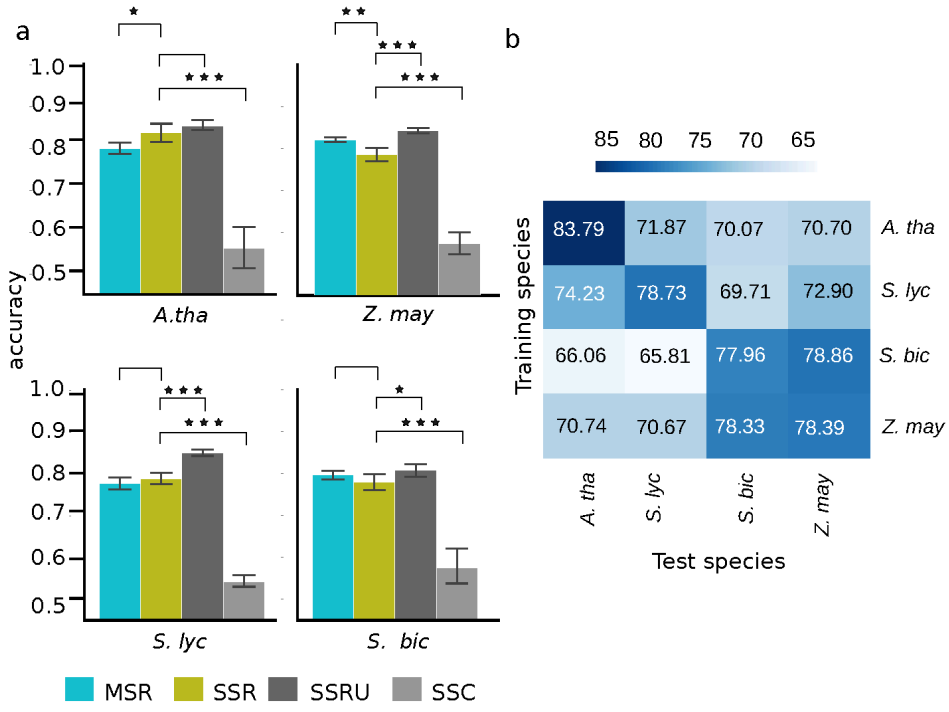


Figure 2

(a) Model performances estimated with binary accuracy. Models were trained for the four reference species A. thaliana (A. tha), Z. mays (Z. may), S. lycopersicum (S. lyc) and S. bicolor (S. bic) with the

strategies of multi-species reference (MSR), species-specific reference (SSR), species-specific reference with homologous sequences (SSRU), and randomly shuffled sequences (SSC).

(b) The performance of the SSR-models when tested on other species. The estimated accuracies range from 65.81% to 83.79 % (light blue to dark blue).

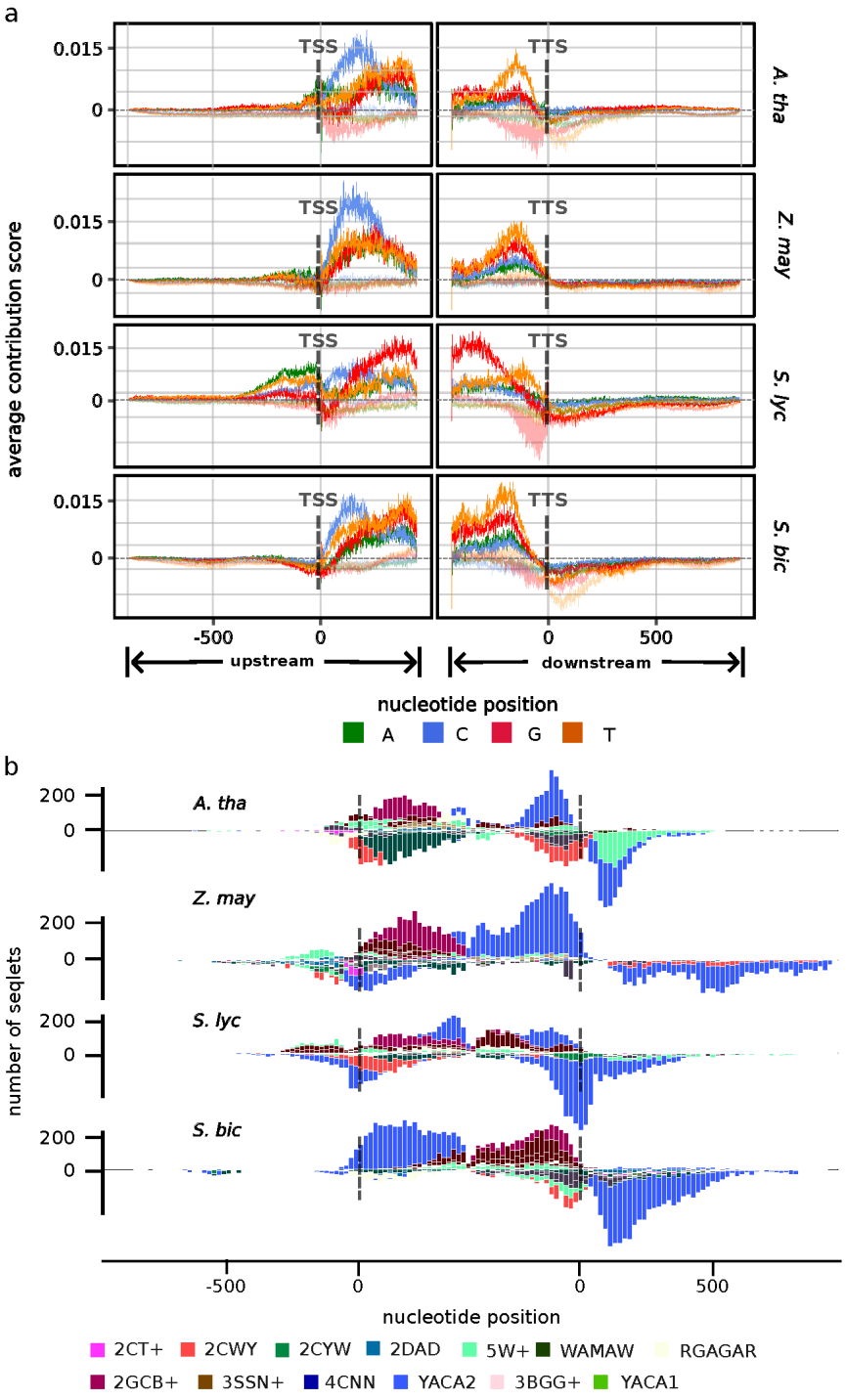


Figure 3

(a) Per-base contribution scores computed with DeepLIFT for A. thaliana, Z. mays, S. lycopersicum and S. bicolor. The positive and negative effect corresponding to high and low expression, respectively, of each nucleotide A (green), C (blue), G (red) and T (orange) per position was determined.

(b) EPMs for the SSR models have distinct distributions across the transcription start and stop sites.

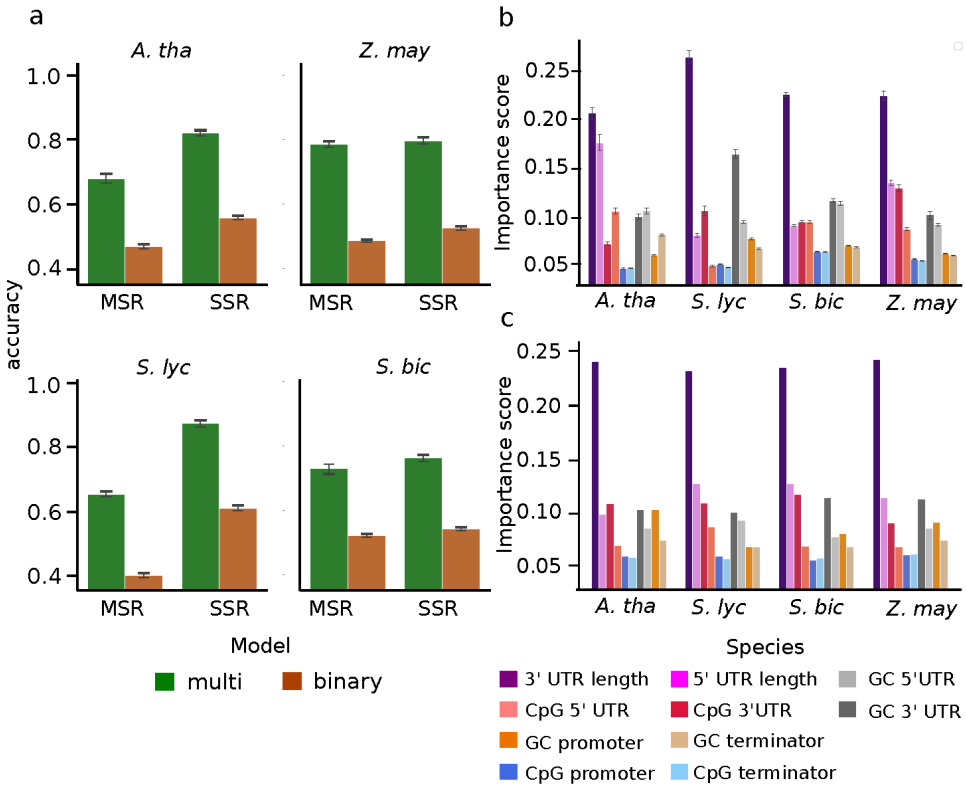


Figure 4

(a) Accuracy of random forest classifiers for binary (low-high) and multi (low-medium-high) classification models, are shown for the reference species A. thaliana, Z. mays, S. lycopersicum and S. bicolor.

(b) Impurity-based feature importance from SSR random forest classifiers.

(c) Impurity-based feature importance from MSR random forest classifiers.

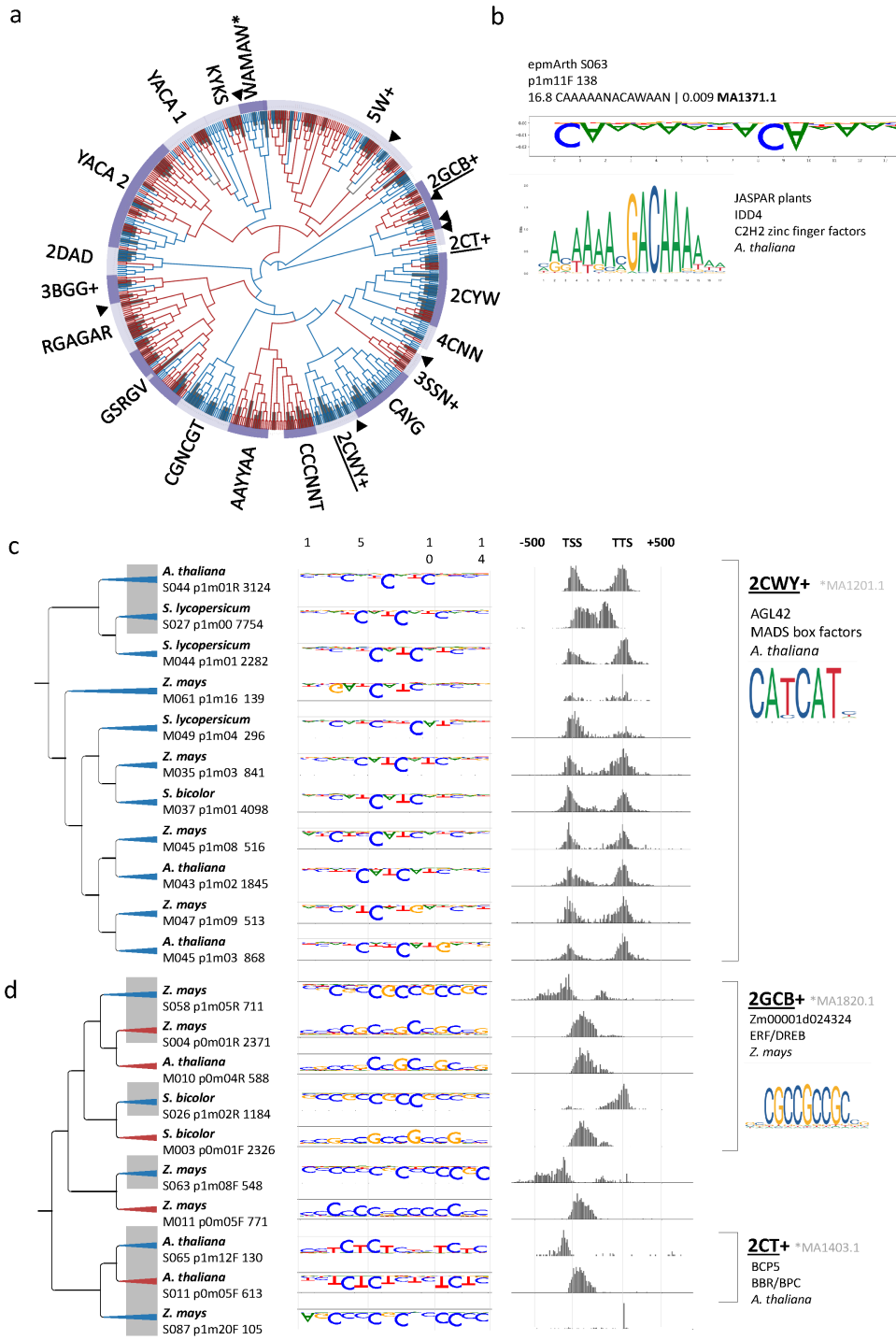


Figure 5

Expression predictive motifs (EPMs) identified by DeepLift and MoDisco determined with convolutional neural networks (CCN) trained on single species reference (SSR) and multi species reference (MSR) models of A. thaliana, S. bicolor, S. lycopersicum, and Z. mays.

(a) The EPMs are assigned into 19 clusters based on similarity using the Smith-Waterman algorithm and manual inspection of the consensus sequences following the alignment. The clusters are named after conserved DNA motifs and indicated by the IUPAC nucleotide code, along with the least number of repeats of motifs (numerals) and potential additions (+). Clusters with EPMs that match significant transcription factor binding sites from the JASPAR database with e-value >0.5 are marked with black triangles. The MSR EPMs identified by the MSR model are highlighted by grey boxes in the dendrogram, while prediction for high and low gene expression is displayed by red and blue branches, respectively. Underlined clusters, with selected representative EPMs are shown exemplarily in the panels (c) and (d). A The complete full scale version of the dendrogram EPMs of EPMs can be found in supplementary figure 1.

(b) The EPM nomenclature system assigns abbreviations to plant species based on their genus and epithet, followed by the model used to produce the EPMs (SSR or MSR). The physiological conditions of the plant are indicated by a number (0 for standard conditions). The predictability of each motif is indicated by a 'p' followed by a number of 1s and 0s for low and high rates of gene expression, respectively. The delimiter is followed by the motif number within the metacluster and its orientation (forward or reverse). Finally, the number of seqlets included, the information content, and the consensus sequences are added at the end of the EPM. For example, the top match (e-value = 0.0009) between the binding site of TF A. thaliana IDD4 and an identified EPM is shown. According to the nomenclature the EPM was identified in A. thaliana (Arth), by the SSR model (S), under standard conditions (0), predicting low gene expression rates (p1), produced from 138 seqlets with an information score, indicating nucleotide frequency, specificity, and motif heterogeneity of 16.8, along with its consensus sequence.

(c) EPMs of the 2CWY+ cluster uniformly predict low rates of gene expression (blue tips). EPMs of this cluster are identified by both the SSR and MSR (grey boxes) models. The MSR models, however, identified similar 2CWY+ motifs for all four species tested. The inverted web-logos show the EPMSaffiliation to MoDiscos metacluster 1 (p1) and histograms display the positional preference inferred from the number of seqlets relative to the transcription start and transcription termination sites (TSS, TTS) of each EPM. EPMs for the 2CWY+ type display significant similarity to the transcription factors binding site of AGL42 (JASPAR accession MA1201.1) of A. thaliana.

(d) The 2GCB+ and 3CT+ clusters contain EPMs predicting both low and high gene expression rates, identified by the SSR and MSR models, corresponding to the positional occurrence related to the TSS or TTS. Both clusters highly resemble previously determined transcription factor binding motifs for 2GCB+ and 2CT+, with MA1820.1 and MA1403.1, respectively. There are EPMs that are highly specific to species and do not have any sequence similarity to other EPMs or known TF binding sites, like Z. mays motif pair S063 and M011 or positionally highly restricted S087.

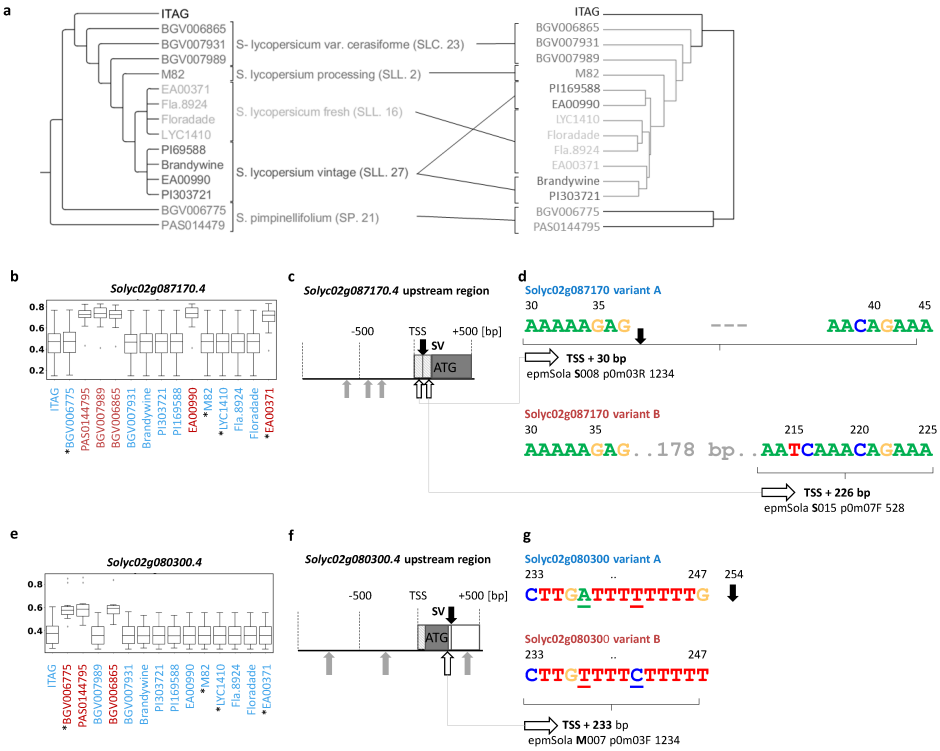


Figure 6

Comparison of the predictability of expression predictive motifs (EPMs) in 15 different Solanum genotypes with structural variations (SVs) in the transcription start site (TSS) and transcription termination site (TTS) identified and characterised by Alonge and colleagues (2020)32.

(a) The cladograms shown, the taxonomic grouping of Solanum species based on SVs inferred by Alonge and colleagues (left) in comparison to hierarchically clustered predictions of gene expression (right) display an overall similar topology, except for the groups of S. lycopersicum fresh (SLL.16) and vintage (SLL27.).

(b) Structural variations in the flanking regions altered gene expression measured in species indicated with asterisks for the exemplarily shown genes Solyc02g087170 and Solyc02g080300 (e). Congruently, high (red) and low (blue) probability scores of the multi-species reference (MSR) model indicate differential gene expression.

(c) Gene maps of upstream regions around the TSS include UTR region (striped boxes), exons ("ATG" + filled boxes) and introns (white boxes), along with the location of significant EPM matches (e-value = 0.00001) inside (white arrow) and outside (grey arrow) of their positional preference and the location of SVs (black arrows)).

(d) For closer inspection only EPMs allocated to the preferred predicted position were chosen. For sequence comparison, ITAG.3 representing genetic variant A and PAS104479 representing genetic variant B are displayed. EPMs of S. lycopersicum MSR model S008 and S015 lie within the 5'UTR of Solyc02g087170. An SV indel mutation of 178 bp located only 37 bp behind the TSS disrupts EPM S008 of S. lycopersicum in species similar to variant B, starting 30 bp behind the expected TSS in species of variant A. Due to the same SV, however, EPM S015 is now localised within its positional optimum for species of genetic variant B.

(e) Probability scores of the multi-species reference (MSR) model indicate differential gene expression for Solyc02g080300. (see caption for sub-figure b)

(f) (see caption for sub-figure c)

(g) One EPM of S. lycopersicum MSR model S007 has been identified within the first intron 233 bp downstream of the expected TSS of Solyc02g080300. An SV indel mutation of 10 bp lies 7 bp downstream of epmSoly-S007 not disrupting the EPM. In contrast to the example before, the difference between two point mutations coincides with low probabilities of high gene expression for variant A within epmSoly-S007.

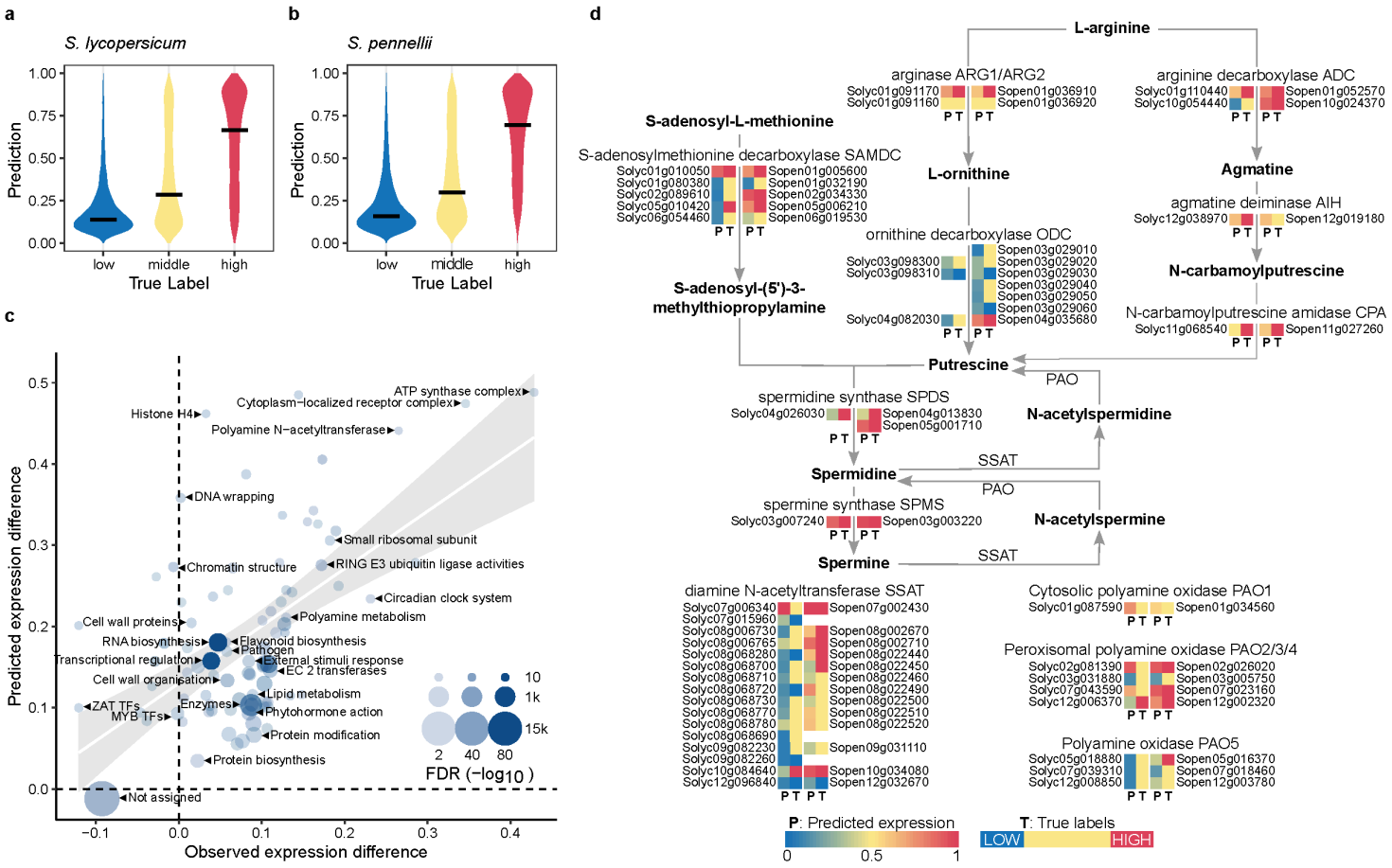


Figure 7

Prediction and comparison of gene expression between S. lycopersicum and S. pennellii by the MSR model.

(a) Distribution of MSR model-predicted expression probabilities for the low (1st TPM quartile), middle (2dn and 3rd TPM quartiles) and high (4th TPM quartile) expressed transcripts of S. lycopersicum.

(b) Distribution of MSR model-predicted expression probabilities for the low (1st TPM quartile), middle (2dn and 3rd TPM quartiles) and high (4th TPM quartile) expressed transcripts of S. pennellii.

(c) Expression of MapMan functional categories exhibiting significant differential expression between S. lycopersicum and S. pennellii (Wilcoxon rank-sum test FDR < 0.05). The difference between mean expression for the observed and predicted values provides the x and y axis coordinates, respectively. Data was scaled so that -1 indicates expression in S. lycopersicum only, 0 indicates equal expression in both species and 1 signifies unique expression in S. pennellii. Linear regression represents fit with Pearson's correlation coefficient of 0.56 (p-value 5.827e-09, R2 0.32).

(d) Mapping of predicted expression probabilities (P; predicted) and observed gene expression (T; true labels) for enzymes of the polyamine biosynthesis pathway in S. lycopersicum and S. pennellii.

Heatmaps are sorted according to gene orthology with the S. lycopersicum on the left side and the respective S. pennellii orthologs on the right.

Supplementary Files

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

- Supplfigure1DLcisregulartoryFWEMPTreeSZ20230428.pdf
- Suppltable1Pelekeetal2023readcountsspecies.ods
- Suppltable2Pelekeetal2023modelperformancecnnrf.ods
- Suppltable3Pelekeetal261EPMscharacterizedSZ20230405.xlsx
- Suppltable4PelekeetalProbabiltiesgeneexprSZv1.xlsx
- Suppltable5Pelekeetal2023DeepLearningsignificantmapmanbinsJJv1.csv