Core promoterome of barley embryo

Simon Pavlu[1,2], Sarvesh Nikumbh[3], Martin Kovacik[1,2], Tadaichi An[4], Boris Lenhard[5,6], Hana Simkova[1]*, Pavla Navratilova[1]*

ORCID:

Simon Pavlu	0009-0009-2917-6337
Sarvesh Nikumbh	0000-0003-3163-4447
Martin Kovacik	0000-0002-7470-1585
Tadaichi An	0009-0007-2293-6239
Boris Lenhard	0000-0002-1114-1509
Hana Simkova	0000-0003-4159-7619
Pavla Navratilova	0000-0003-3719-2897

Affiliations:

[1]Institute of Experimental Botany of the Czech Academy of Sciences, Slechtitelu 31, Olomouc 77900, Czech Republic

[2]Department of Cell Biology and Genetics, Faculty of Science, Palacky University, Slechtitelu 27, 78371 Olomouc, Czech Republic

[3]Merck Sharp & Dohme (UK) Limited, 120 Moorgate London, EC2M 6UR, UK

[4]DNAFORM Precision Gene Technologies, 230-0046 Yokohama, Kanagawa, Japan

[5]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London, UK

[6]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, London, UK

* Corresponding authors:

Pavla Navratilova navratilova@ueb.cas.cz

Hana Simkova simkovah@ueb.cas.cz

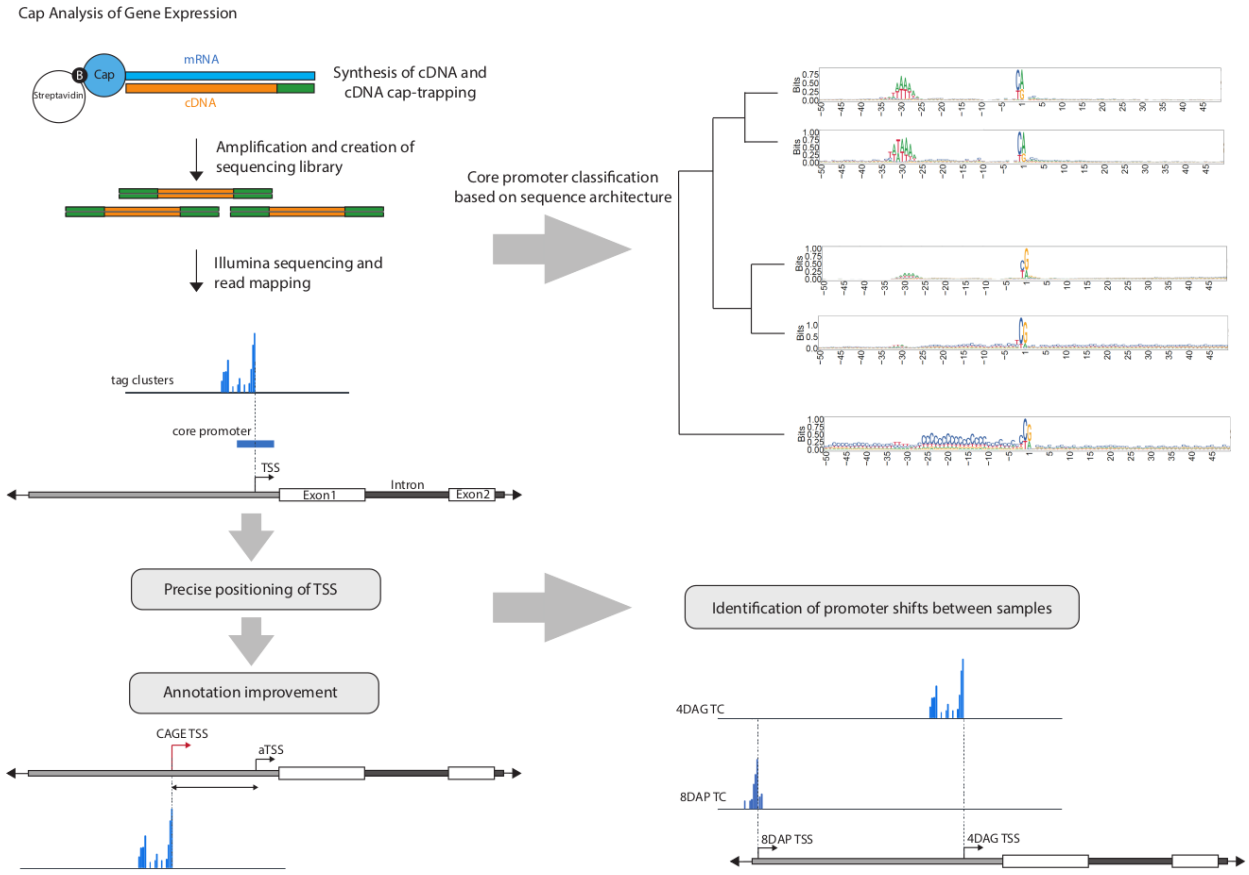
Highlights

- CAGE data provided reliable support and means of improvement to the most recent barley reference gene annotation
- Previously undescribed alternative promoters and promoter shifts between the analyzed embryo stages were revealed.
- Barley core promoter architecture types linked to the gene categories and related to epigenetic profiles were postulated using novel compuational approach.
- Exhaustive barley embryo promoterome datasets are provided.
- New, more complete barley GO annotation for all genes is provided.

Abstract

Precise localization and dissection of gene promoters are key to understanding transcriptional gene regulation and successful bioengineering. Core RNA polymerase II initiation machinery is highly conserved in eukaryotes leading to a general expectation of equivalent underlying mechanisms. Still, less is known about promoters in the plant kingdom. We used cap analysis of gene expression in three embryonic developmental stages of barley to precisely map, annotate and quantify transcription initiation events. Unsupervised discovery of *de novo* sequence clusters grouped promoters by the characteristic initiator and position-specific core-promoter motifs. Complementing by transcription-factor binding-motif annotation, integration with genome-wide epigenomic datasets and gene ontology enrichment analysis further defined chromatin environments and functions of genes driven by individual promoter categories. The TATA-box presence governs all features explored, supporting the general model of two separate genomic regulatory environments. We describe the scale and consequences of alternative transcription initiation events, including developmental stage-specific ones, that affect the protein sequence or the presence of translational regulatory regions. The generated promoterome dataset poses a valuable genomic resource to improve functional annotation of the barley genome, understand the transcription regulation of individual genes, and manipulate the promoter architecture towards enhancing traits of agronomic importance.

Graphical abstract



Keywords

Core promoter, Cap Analysis of Gene Expression, Morex, initiator, TOR-signaling, embryo

1. Introduction

Biotechnological research aims to increase plant utility or achieve higher plant resilience through full control of gene expression on either of the two levels. First, transcription is controlled by dictating the amount of mRNA that is produced from a particular gene. The second level of control is through post-transcriptional events that regulate the translation of mRNA into proteins. Our full understanding of these processes goes through elucidating the role of each nucleotide of the specific genomic sequences engaged in the expression

control. The ultimate platform for integrating transcriptional regulation signals is a core promoter, the full set of which in a given species is called promoterome. The core promoter of eukaryotic RNA polymerase II is defined as the minimal sequence at the 5' end of a gene that is recognised and bound by general transcription factors through TFIID - the largest multiprotein complex comprising TATA-binding protein (TBP) and ~12–13 TBP-associated factors (TAFs) [1] - a prerequisite for the RNA polymerase pre-initiation complex (PIC) formation and subsequent events leading to firing off the transcription. Therefore, the core promoter regions contain sequences of the 5' termini of all mRNAs transcribed beginning from the very first nucleotide of the full-length transcript, often imperfectly defined in genome annotations. Crucially, promoter diversity is a function of different cofactor binding and activity - the key to targeting gene responsiveness to transcriptional enhancers, whose activity is still not fully resolved in plants [2–4].

Promoter mechanics and sequence composition have been extensively studied in metazoa with the aim of categorizing them and providing a roadmap to their understanding (reviewed in [5]). A number of well-defined core promoter sequence elements have been characterized, such as the initiator element (Inr), TATA box, TFIIB recognition element (BRE), core promoter motif ten elements (MTE), downstream promoter element (DPE) and others that co-occur in combinations (reviewed in [6]). Human and Drosophila transcription initiation have been known to start from a well-characterized longer Initiator (Inr) sequence (YYANWYY and TCAKTY, respectively), which might even be sufficient for transcription [7]. In spite of their importance and conservation, novel and lineage-specific promoter elements, both upstream and downstream of the transcription start site (TSS), are still being discovered in both plants (e.g., the Y patch by [8]) and animals [9–12]. Their combination with initiator variants determines the PIC composition, TSS selection, level of polymerase engagement, transcriptional burst size, cis-responsiveness and pausing [13,14] with direct consequences on gene expression. The common initiator sequence (i.e. position -1,+1 relative to the TSS) is a pyrimidine-purine (PyPu), with special cases like in genes regulated by the Target of Rapamycin (TOR) signalling, known to have a specific ‘TCT’ initiator [15], as part of a nutrient-sensing mechanism in many animals. The sequence downstream the initiator (5'UTR) determines, which TAFs can bind and may comprise the 5’ Terminal OligoPyrimidine (TOP) motif - a TOR-binding platform. TOR is a conserved eukaryotic

protein kinase that regulates metabolism by promoting anabolic processes when nutrients are available and is essential in Arabidopsis' early embryogenesis and under stress [16].

Despite a diversity of scenarios, a limited number of synergistic motif/TF combinations exist per species with the two highlighted configurations: TATA-Inr-MTE and BRE-Inr-MTE, as supported by a PIC structure study [17]. This classification aligns with two major classes of promoters in mammals with respect to the number and distribution of detected TSSs: focused (one initiation site, ~5 bp, 'sharp' initiation) and dispersed (multiple initiation sites, up to 100 bp, 'broad' initiation) [18], which tightly coheres with chromatin configuration as nucleosome positioning, histone variants [19] and their posttranslational modifications [20], determining motif-independent ‘PIC catchment area’ [21]. Intriguingly, the core promoter configurations are also associated with distinct gene categories, which seem to hold true for many organisms, including plants [22,23].

The path to characterizing promoters can be taken from gene annotation or promoter prediction methods, however, the direct proof of promoter position and distribution of TSSs is provided by sequencing of long capped RNA species called Cap analysis of gene expression (CAGE) [24] or similar methods, e.g. GRO-seq, Smar2C2, TSS-seq, [22,25,26]. The CAGE has been selected as an integral method of large genome-consortia projects, facilitating correct gene annotation and providing gene expression levels [27,28]. Compared to metazoans, only a handful of studies concentrated on systematic plant promoter characterization, limited to Arabidopsis [29], soybean, rice, sorghum, wheat and maize [22,23], which hinted that the sequence features may vary between plant species. Thorough functional analysis of promoters comparing monocot and dicot reporter systems using promoters from three species (Arabidopsis, maize and sorghum) confirmed species-specific sequence-dependent differences in promoter reactivity and strength [30] which implies differences in transcriptional machinery between and within the two groups. The authors demonstrated experimentally the promoter mutation effect, underscoring the importance of a well-informed promoter design for transgenesis.

Considering the crucial role of Triticeae crops - wheat, barley and rye - in human and animal feeding, barley (*Hordeum vulgare* L.), as a diploid species with a reference genome assembled in near-T2T completion [31]; [32] and available transgenesis protocols [33], appears a suitable real-world candidate to study the promoter architecture with a chance of

information applicability in the related Triticeae crops. Here, we generated CAGE datasets from three developmental stages of barley embryos so as to detect promoters involved in a range of differentiation processes and unravel the dynamics in TSS usage during the development. The data revealed misalignments between the annotated and the detected TSSs as well as the presence of alternative TSSs, presumably affecting both regulatory and protein sequences. Application of a novel core-promoter sequence clustering approach, combined with the investigation of gene functions and epigenetic features, suggested the functional divergence of genes with different promoter categories, reflecting disparate co-factor and transcriptional regulation modes. Overall, we demonstrated the power of CAGE promoterome profiling of multiple plant developmental stages in providing nucleotide-level positions as well as the initiation measures of all possible alternative TSSs of every active gene. The resulting promoterome dataset provides a basis for understanding the transcriptional-regulation logic and poses a valuable genomic resource for research and agricultural biotechnology in barley.

2. Methods

2.1. CAGE

Barley cv. Morex was grown in growth chambers at 16/8 hrs light cycle, 16°C day/12°C night temperature. For the 4DAG seedlings, seeds were germinated on wet tissue paper at 20°C for 4 days before harvesting and removing remnants of seed coat and endosperm. The 8- and 24DAP embryos were staged according to their time of fertilization, size and phenotype and dissected as described previously [63]. Total RNA was extracted by Monarch® Total RNA Miniprep Kit (NEB cat#T2010S) and its quality was checked by Bioanalyzer (Agilent) to ensure that the RIN (RNA integrity number) values were over 7.0. CAGE libraries were sequenced using single-end reads of 150 nt on a NovaSeq instrument (Illumina). CAGE library preparation, sequencing, and read mapping on MorexV3 annotation were performed by DNAFORM (Yokohama, Kanagawa, Japan).

2.2. CAGE data and motif analysis

Obtained reads (CAGE tags) were mapped to the MorexV3 genome using BWA (version 0.7.17). Unmapped reads were then mapped by HISAT2 (version 2.0.5) while rRNA reads

were filtered. Mapping rates varied between 41 and 87% of total reads. Regions that had a 90% overlap between replicates were extracted by BEDtools (version 2.12.0). Tag count data for each of the samples were clustered using the CAGEr program, which merged neighbouring TSSs with mutual distance < 20 bp into a single tag cluster (TC). Clusters with tags per million (TPM) < 0.1 were discarded just as singletons that had a TPM signal < 5. MorexV3 annotation (both high and low confidence genes) was used to assign a genomic category to each promoter candidate by applying the 'annotatePeak' function from the ChipSeeker package [35]. The genomic features were assigned according to the following hierarchy: promoter > 5'UTR/3'UTR > exon > intron > proximal, where 'promoter' has been defined as ranging from -500 to +100 bp and 'proximal' from -1000 to -500 bp relative to the aTSS of the nearest gene (Figure 1B). For each gene ID, the hierarchically highest feature was assigned. In case two candidates had the same hierarchical significance, the candidate located closer to the aTSS was taken as the primary promoter. The rest were set aside as the 'secondary promoter dataset'. For further analysis, the 'consensus' clusters were determined using the 'aggregateTagClusters' CAGEr function, which aggregates TSSs from the three samples, merging those with mutual distance < 100 bp . The set of consensus clusters created by this method was filtered and split between the primary and secondary TC datasets as described above.

The consensus candidates from the CAGEr were further clustered according to their sequence similarity using a seqArchR program (v1.2.0, Nikumbh 2023). The seqArchR configurations were kept default with the exceptions of 'bound' being set to 10^-8 and 'chunk size' to 5000 for primary promoters (500 for the secondary set) to better suit our dataset. The resulting 49 clusters for the primary dataset were further collated into nine final seqArchR clusters based on sequence logo similarity using the seqArchRplus utilities. The 27 secondary clusters were collated into the final seven. Scripts with more details can be found in the github repository.

The +/- 50 bp around all TC`s dominant TSSs were subjected to peak-motif position-analysis by RSAT with the settings -max_seq_len 1000 -disco positions -nmotifs 5 -minol 6 -maxol 7 -no_merge_lengths -1str -origin center -motif_db footprintDB.plants transfac footprintDB.plants.motif.tf -scan_markov 1 -task purge,seqlen,composition,disco,merge_motifs,split_motifs,motifs_vs_motifs,timelog,archive,

synthesis,small_summary,motifs_vs_db,scan -noov. This was followed by the hierarchical matrix-clustering method to generate a summary radial tree with the following settings: matrix-clustering -v 1 -max_matrices 300 -hclust_method average -calc sum -metric_build_tree 'Ncor' -lth w 5 -lth cor 0.6 -lth Ncor 0.4 -radial_tree_only.

The TOP score was calculated according to [45] from the 4DAG CAGE BAM file. The calculation script is available at: https://github.com/carsonthoreen/tss_tools/blob/master/tss_analyzer.py

2.3. Tissue specificity and GO-term annotation of promoter clusters

The tissue specificity values were calculated by applying the Tau algorithm written as R script (deposited at https://rdrr.io/github/roonysgalbi/tispec/man/) on tpm (transcripts per million) matrix from 18 distinct samples from the EoRNA (https://ics.hutton.ac.uk/eorna/index.html) datasets.

To increase the accuracy of barley GO-term annotation we have utilized the Gene Ontology Meta Annotator for Plants (GOMAP)-singularity pipeline [48], which combines three different annotation techniques using the genome protein fasta file as an input. With the new barley GO-term annotation in hand, the GO-terms were assigned to each gene of the seqArchR clusters. Using the 'Enricher' function from the 'clusterProfiler' package with default settings we determined the enrichment of GO-terms within the clusters, calculating the p-value based on the hypergeometric model. The enriched GO-terms together with their p-values were then loaded into the REVIGO web interface [64] to reduce the redundancy of the enriched set and differentiate between the BP (biological process), CC (cellular compartment) and MF (molecular function) GO-term categories. The raw TSV REVIGO data for each of these categories were then used to produce dot plots, showcasing only the top 5 most enriched GO-terms for each of the seqArchR clusters.

2.4. Histone modification ChIP-seq and MNase-seq data analysis

Barley embryos grown and collected for the CAGE were used for nuclei isolation and MNase digestion followed by native ChIP-seq as described previously [65], with modifications detailed in [32]. These ChIP-grade antibodies against modified histones were used: H3K4me3 (Abcam ab8580) and H3K27me3 (Diagenode C15410195). Resulting

sequencing libraries, including those generated from MNase-digested input, were sequenced on the NovaSeq6000 platform. Reads from the ChIP-seq pipeline went through qualitative trimming by Trim Galore (v. 0.6.2) and mapping to the MorexV3 reference genome was performed using the Bowtie 2 package (version 2.4.2). Duplicated reads were then removed using Picard tools (version 2.9.0) and MACS2 software was used to call peaks. For creating heatmaps, the deeptools functionalities computeMatrix and plotHeatmap were utilized, with the kmeans clustering set to 2, to identify any divergence within the promoter clusters. ChromHMM analysis was performed according to the ChromHMM protocol [66]. MNase-digested input WIG files were used to calculate nucleosome positions around dominant CTSSs using DANPOS software https://github.com/sklasfeld/DANPOS3. The resulting WIG files were plotted using deeptools` computeMatrix and plotProfile functions.

2.5. RNAseq and CAGE correlation data analysis

The 4DAG [46] and 8/24DAP (Kovacik 2023, bioRXiv) RNA-seq datasets were analyzed using the RSEM software with STAR mapping pipeline [67,68]. The CAGEr datasets were constructed using CAGEr, merging replicas and normalizing tag counts to TPM. ChrUn records were filtered out of both datasets, CAGE and RNAseq and only records that had a TPM value higher than 0.1 were considered.

3. Results

3.1. Cap Analysis of Gene Expression profiles transcription initiation events in barley embryos

To identify transcription initiation events genome-wide in barley cv. Morex, we applied CAGE to total RNA isolated from three embryonic developmental stages: eight days after pollination (8DAP), 24 days after pollination (24DAP) and four days after germination (4DAG) in two highly correlated replicates each (Pearson correlation 0.99; Figure S1a). By sequencing the CAGE samples, we obtained 52,313,604/57,927,548/60,825,231 CAGE tags from the 8DAP/24DAP/4DAG samples, respectively. Data analysis was performed following the workflow depicted in Figure 1A. We utilized CAGEr software [34] to merge

replicates, identify CAGE TSSs (CTSSs) and aggregate them into 49,848/51,903/54,276 CTSS tag clusters (TCs) with custom parameter specifications (Figure S1b). Each TC is featured by its interquantile width (IQW), which characterizes promoter breadth, tag counts (tags per million, TPM), corresponding to the expression level, the position of a dominant CTSS and the associated gene ID. All TCs were annotated using ChipSeeker [35] by genomic features and gene IDs, considering both high- and low-confidence genes annotated on MorexV3 pseudomolecules [31] (Figure S1c). Distal intergenic TCs, located >1000 bp from the nearest annotated TSS (total of 7662/7296/7390 TCs from 8DAP/24DAG/4DAG, respectively), were deemed unassociated with the closest gene and were analyzed as a 'secondary TC set'. We searched for the hierarchically highest ranking type to keep a single TC (the most likely promoter) per gene; see the Methods section and Figure 1B for details. This process left us with a 'primary TC set' of 19,289/19,567/20,878 promoter candidates for 8DAP/24DAP/4DAG, respectively. The rest were assigned to the secondary TC set, which involved alternative promoters and intragenic and intergenic TSSs, including promoters of putative unannotated genes.

As the goal was to provide a generalized classification of promoter types, the primary TCs from all three embryonic stages were pooled and the overlapping ones merged into a set of 34,897 consensus clusters. These were subjected to the same filtering and annotation, leaving us with 21,610 'primary consensus' and 13,287 'secondary consensus' clusters (Figures 1b, c, Table S1). These consensus clusters became the core data used for most analyses. A complete list of the TCs, including consensus-TC coordinates in the MorexV3 genome, the assigned-gene ID, position of the dominant TSS, feature annotation, classification, IQW and TPM, are provided as Data S1 and are shared in the EPD database (https://epd.expasy.org/epd/). Since the promoter width has been recognized as an important characteristic distinguishing different functional classes of promoters, we compared IQW distributions of the primary and the secondary cluster sets and observed a trend towards narrower TCs for the latter one (Figure 1D; Figure 2). A group of extremely sharp (width 1-2 bp) TCs, termed singletons and found within the secondary TC set, comprised the initiation at the first intron-exon boundary (CAG sequence) with notably high TPM values.

3.2. CAGE data improve gene annotation and provide transcription initiation levels

Promoters are characterized by isolated clusters of CAGE peaks. Defining the most frequently used dominant CTSSs enables positioning the TSSs with 1-bp precision, which should ideally correspond to the gene annotation. To compare the CTSSs positions with the TSSs in MorexV3 annotation, we applied dominant CTSSs from the primary set, whose distance from the annotated TSSs (aTSSs) was not greater than 500 bp. The comparison revealed a certain level of misalignment. When allowing a 20-bp deviation in each direction, 39% of expressed genes (detected by both CAGE and RNA-seq) agreed on the TSS position. Another 29% of aTSSs are located within 100 bp and the remaining 31% are within 500 bp of the dominant CTSS (Figure 1E, F). From this primary set, 2583 promoters belonged to the low-confidence gene category. Considering this, the defined CAGE signal within close proximity of the aTSS might be grounds to reconsider the categorization of these specific genes. Based on the CAGE signal supported by RNA-seq and epigenomic data, we detected multiple putative unannotated genes, with an example shown in Figure S1d.

A summary of tag counts within a CTSS serves as a measure of 5' transcription initiation level giving a measure of gene expression (TPM). We compared the CAGE and RNA-seq datasets both qualitatively (i.e., which genes were detected by one or both methods) and quantitatively by calculating correlations of tag/transcript counts for active genes detected by both methods (Figure S2, Data S2). The comparison revealed that CAGE has a detection limit on the low-expression side (Figure S2c). On the contrary, genes detected by CAGE, but not found in RNA-seq data were not predominantly low-expressed (Figure S2d) and there might be other biological or technical reasons for the discrepancy.

3.3. Promoter architecture clustering and the initiator

Based on previous extensive work on other species` promoters (reviewed in [36] and elsewhere), we considered 50 bp up- and downstream from the dominant TSS to contain a PIC-binding sequence. To cluster core promoter sequences according to their characteristic sequence architectures, we used seqArchR [37], a recently developed method using non-negative matrix factorization. The clusters are characterized by *de novo*-identified

sequence elements, such as position-specific motifs or nucleotide composition of the input sequences. Initially, we analyzed the primary TC sets from all three developmental stages, resulting in 15/16 sequence architecture clusters for each (Figure S3). To provide a generalized classification of promoter types across stages, we generated a set of primary consensus promoters and subjected it to the SeqArchR analysis. Resulting clusters, defined by the sequence architectures (composed of the initiator sequence, positioned TATA-box and other sequence motifs) and supplemented with the IQW and expression levels, were manually collated into nine final clusters (Figure 2A). The initiator, corresponding to the dominant CTSS and visible as the sequence logo around positions -1/+1, represents a highly significant motif in all clusters. Barley lacks a longer Inr sequence motif or the 'TCT'. The TATA-box promoters (clusters 1-3 of the primary set) are transcribed predominantly from 'CA' initiator, driving mostly genes with high expression levels. This points to the key role of the TATA box in defining the initiator sequence as well as the promoter activity (transcriptional burst frequency). This characteristic contrasts with the more variable PyPu constellation ('CG' or 'TG') predominant in the promoters without the actual TATA box, which tend to have a lower expression (clusters 4-9, Figure 2A, 3A).

The same clustering method was applied also on the secondary TCs (Figure 2B). Seven distinct clusters resolve best all motif configurations, which can be split into three main groups: clusters 1-3 represent TCs originating from the first intron-exon boundary, which are extremely sharp and have TPM values often higher than the primary TC of the same gene. They are transcribed in the sense direction and the base at the TSS corresponds to one of the three bases forming the conserved eukaryotic splice-acceptor site cAG. An example - a gene coding for 40S ribosomal protein assigned to the secondary cluster 1 - is shown in Figure S4. On the contrary, clusters 4 and 5 of the secondary TCs relate to antisense transcripts initiated at the G-rich initiation sequences at the intron-exon boundary without preference for the intron position in the gene. Lastly, clusters 6 and 7 likely represent true alternative promoters, which are generally TATA-less, and also promoters of anticipated unannotated genes (Figure S1d).

3.4. The TATA box and other previously defined motifs in barley promoters

The primary clusters 1-3, characterized by a positioned TATA-box-like motif, define the TSS in significantly narrower regions (sharp promoters), unlike the broader non-TATA promoters. This is in line with all studied organisms, but the TATA-box position and frequency of its use vary. To reproducibly quantify the TATA-box sequence, we used the position weight matrix (PWM) of the canonical plant TATA box (inset in Figure 3B) generated from 134 unrelated plant promoter sequences from the Eukaryotic Promoter Database [38]. Strand-specific TATA-box frequency search with this PWM using the FIMO tool [39] under the default p-value threshold 1e-4 resulted in 7.8/7.9/9% of all promoters active in 8DAP/24DAP/4DAG barley stages, respectively. Using a more relaxed p-value of 1e-3, which included motifs quite distant from the canonical TATAWAW, resulted in 20/20/23% of active promoters.

Position analysis using the same canonical PWM and TA pentamers as in [40] showed that the starts of the TATA-box and more diverged W-box motifs are positioned at -29 to -36 upstream from the initiator site, with the peak at -32 and higher distances observed for more conserved TATA-box motifs (Figure 3B). A certain proportion of TATA-box promoters is associated with distinct C nucleotides right upstream of the TATA box as revealed by subsequent motif discovery (Figure 4), which may be a sign of the presence of BRE-upstream sequence (BREu, SSRCGCC), although we did not detect full BREu or BREd (BRE-downstream sequence, RTDKKKK) (Figure S5). Except for the TATA and distinct initiators, other previously known core promoter motifs were not detected (Figure S5). However, dinucleotide heat maps generated by seqArchRplus provide an overview of the common PyPu dinucleotides at TSSs, including ominously present W boxes (W=A/T) and the high SS (S=G/C) content in the majority of barley promoters (Figure 3C, D).

Aiming to detect less pronounced motifs within the core promoter regions of each cluster and to find, which regulatory proteins might putatively target them, we complemented the cluster analysis with an analysis using the RSAT (Regulatory Sequence Analysis Tools) toolkit [41]; https://github.com/rsa-tools/rsat-code). Specifically, we utilized a 'peak-motif positioned' function that detects oligonucleotides showing a positional bias, i.e. having a non-homogeneous distribution in the sequence set, followed by the transcription factor binding site (TFBS) search against footprintDB.plants [42]. For a comprehensive summary, the pool of motifs from all nine consensus clusters from the primary set was subjected to hierarchical matrix clustering represented by the radial tree (Figure 4, Data S5). This

analysis confirmed that besides the notoriously known canonical TATA box (i.e., TATAWAWR), extended TA repeat, i.e., stretches of 6-15 TA dinucleotides occur in about 5.5% cases (p-value 1e-4), not matching the TBP matrix in the footprintDB. Instead, it appears as a target of zinc finger protein ZHD10 involved in establishing polarity during leaf development through the gibberellic acid (GA) signalling pathway. Other putative W-box motifs match the matrix of AGL3/PHE1 - both MADS-box homeotic transcription factors.

Also, other motifs in barley core promoters have a low-complexity nature, such as the previously described pyrimidine-rich Y patch [8], shown experimentally to stimulate gene expression in the maize reporter system [30]. This motif, defined as a CTTCTTCCTC (or its reverse-complement GAGGAAGAAG) sequence, occurs in over 14% and up to 70% of barley core promoters using strict (p-value 1e-5) and relaxed (p-value 1e-4) search settings, respectively. TFBS analysis by RSAT assigns to this motif BPC1 (BASIC PENTACYSTEINE1), an octodinucleotide GA-repeat-binding protein, having a homolog in barley (BBR). This protein likely participates in the Polycomb-mediated transcriptional regulation of developmental genes via binding Polycomb Repressive Elements [43]. Other significant hits were 3xHMG-box proteins, linked to cell proliferation with a role in the organization of plant mitotic chromosomes [44], and hormone-responsive factors TF3A, FRS or SHN3. Another, low complexity motif group dominating the barley core promoters is a GCC box, known to be bound by a GCC box-binding factor (GBF) and/or ethylene-responsive factors (ERFs), all of which are hormone-responsive proteins involved in stress responses and developmental processes. The same analysis of extended sequence (+/-100bp) failed to add any new motifs (Figure S6). Other TF binding in core promoters includes E2F, MYB and NAC factors.

3.5. CT-rich motifs downstream TSSs might serve as a nutrient-sensing TOP motif

Although the search for TFBS matching the promoter sequence resulted in many hits, the region right downstream of the TSS might have an other than transcriptional regulation function. It has been shown to serve as a translational signal, specifically through the TOP motif, which begins with a cytosine at the 5′ cap and is followed by several uracils and/or cytosines, and no or very few adenines or guanines. To assess the possibility that some

barley promoter sequences serve as a platform for TOR, we calculated TOPscore values for our CAGE dataset in 4DAG embryos using the algorithm by [45]. This revealed a set of mRNAs that begin with a likely 5′TOP motif as potential subjects to the TOR-5′TOP nutritional signalling. Overall, each cluster's distribution and means of TOPscores were significantly skewed towards lower values in the TATA-box clusters 1-3 and cluster 6, while the remaining clusters showed values significantly higher (Figure 5A). 558 barley genes had TOPscore values >3, considered to carry the bonafide TOP-motif signature in their 5'UTRs. Gene ontology (GO)-term analysis of these candidate genes revealed that they are significantly enriched in ribosomes, Golgi apparatus and plastids and are functionally involved primarily in sugar metabolism, cell division and plant growth (Figure 5B). This indicates that they are likely to be direct targets of the nutrient-sensing pathway.

3.6. Tissue specificity and GO enrichment analysis

To assess how features of genes driven by particular promoters relate to the promoter categories, we first looked at how broadly or specifically they are expressed during plant development and across different tissues. We calculated a tau value for each of the genes from a TPM matrix of a broad range of tissues represented in the developmental transcriptomic dataset [46], available in the EoRNA database [47]. The tau value distributions of individual promoter clusters again clearly separate clusters with the TATA boxes, having significantly higher tissue specificity than the TATA-less, which tend to be expressed more ubiquitously (Figure 2A, Figure S7). Notably, the most ubiquitous are the genes that exhibit a sharp capped-mRNA signal at the intron-exon boundary, i.e., the TCs included in secondary clusters 1-3 (Figure 2B).

To conduct a GO enrichment analysis, we generated a detailed plant-centred GO annotation using the GOMAP toolkit [48]. Compared to the published MorexV3 GO-term annotation [31], generated by the Automated Assignment of Human Readable Descriptions (AHRD) pipeline [49], GOMAP generated additional 3491 terms, assigning one to each barley gene, including those with a missing GO term in the published annotation (37% of the whole gene set). Annotation of several well-defined gene categories (e.g., histones, auxin-responsive genes, MADS-box) confirmed a better definition of gene functions compared to the

published version (Figure S8). The new barley GO-term annotation is available in GAF format as Data S3.

The resulting enrichment of GO terms in the individual promoter clusters revealed that the genes driven by the most active and narrow TATA-box promoters were annotated as responsive to environmental triggers, stress, and hormonal, developmental and organ growth signalling (Figure 6A), topped by the response to karrikin - a strigolactone-like plant growth regulator implied in seed germination, nitrate response, peroxidase activity and protein folding. In a biased approach, a manually extracted group of 116 auxin-responsive genes clearly associated with the first three clusters (Figure 6C), just as did genes encoding histones. The non-TATA gene clusters 4-9 were largely overlapping in their functions, which could be assigned to housekeeping and some metabolic functions as well as translational regulation (RNA modification, Golgi apparatus), represented, among others, by ribosomal protein genes (Figure 6C). Interestingly, the terms related to transcriptional regulation (transcription cis-regulatory binding or chromatin binding), typically associated with developmental genes, are represented in both promoter categories in the primary set. Surprisingly, the first five secondary clusters have apparently a lot in common regarding their functions (Figure 6B), which could be defined as strictly metabolic, ribosome-related and involved in glucose metabolism, photorespiration and RNA methylation and binding.

3.7. Promoter developmental shifts across barley embryo stages

Altering promoter choice is one of the mechanisms for cell type differentiation. Usage of an alternative promoter, which manifests as a separate TC associated with the same gene, can be accompanied by a change in promoter architecture, typically switching between TATA-box and TATA-less promoter types. The putative alternative promoters identified in our study largely belong to the secondary cluster 6. A proportion of the alternative promoters are developmentally regulated, which we dubbed 'moving promoters'. The average length of the promoter shift was around 500 bp, typically including 5'UTR, the first exon or intron and therefore affecting the length of the UTR or the coding sequence. We searched for these cases in our CAGEr consensus dataset, considering only those TSS pairs that involved at least one of promoter, 5’UTR and promoter-proximal sequence (-500 to -1000 bp relative to the TSS). The potential TSS shifts between embryonic stages were then evaluated based

on the TPM values of each TC in the pair. The comparison of pairs of stages with the highest stage-specific TPM values revealed 182/154/160 genes with developmentally changed TSSs between pairs of stages 8DAPx4DAG/8DAPx24DAP/24DAPx4DAG, respectively (Data S4). Of these, 60/34/51 involved the coding region, potentially changing the amino-acid sequence, and 110/115/103 affected the promoter/5'UTR region, with possible impact on transcriptional, translational and transport signals. As an example, an alternative promoter in the first intron of an Argonaute protein gene is active in 4DAG embryos and produces a 5' truncated transcript compared to the 8DAP stage. The transcription-initiation alternatives were also reflected in RNA-seq and ATAC-seq data (Figure 7A). A 5'UTR/promoter-shift example is shown for a kinase-like protein gene (Figure 7B).

3.8. Epigenetic characteristics of barley promoters

The two main regulation modes - TATA-dependent, generating sharp TCs, vs. those associated with broad TCs - are known to be related to different promoter-proximal nucleosome positioning in metazoans [50]. To explore how nucleosome positioning relates to barley promoter architectures, we used information from MNase-digested DNA sequencing of 24DAP embryos. Plots showing nucleosome distributions reveal a dominant nucleosome right downstream of the TSS – a feature common for all promoters – but two distinct profiles of other surrounding nucleosomes (Figure 7C, detailed in Figure S9). In the region upstream of the TSS, the narrower, TATA-box, promoters have a well-positioned nucleosome, in contrast to non-TATA promoters with a nucleosome-free region. On the contrary, several nucleosomes downstream of the TSS seem to be very precisely positioned in the latter promoter category but less so in the TATA-box group. The key promoter nucleosomes are typically decorated by histone-3 post-translational modifications, namely H3K4me3 and H3Kac, in active genes. We performed native ChIP using histone modification-specific antibodies and profiled both these activating modifications as well as Polycomb- and facultative heterochromatin-related H3K27me3. K-means clustering of each cluster profile revealed that TATA clusters contained two distinct composite profiles - the H3K4me3 plus H3K9ac and H3K4me3 plus H3K27me3, resembling bivalent marking considered to poise expression of developmental genes (Figure S10a, b). In contrast, the

silencing H3K27me3 mark was almost absent from the non-TATA promoters, which is consistent with their lower tissue specificity. Chromatin-state analysis, done for particular stage-specific clusters by ChromHMM [51] (Figure 7D), confirmed that all TATA-box promoter clusters contained both the activating marks (H3K4me3 and H3K9ac) and – to a lesser extent - the silencing H3K27me3. The proportion of promoters carrying the silencing mark was increasing in hand with the cell differentiation.

An important promoter feature, tightly related to the epigenetic landscape, is a transposable element content. In the human genome, around 18% of human start sites had been defined by CAGE-seq overlap with TEs [52]. Intriguingly, when inspecting the degree of overlap of 100 bp around the TSSs with TEs as annotated by [46,53], only 5-6% of TATA box-containing TCs overlapped whereas 10 (cluster 6) to 15% (cluster 5) of the non-TATA clusters did (Figure S10c). Predominant TE families were DNA transposons CACTA and retrotransposons *Copia*, *Gypsy* and others unknown (LTR-RLX class). The overlap of these elements with promoters re-opens questions regarding their role in plant gene regulation, development and evolution.

3. Discussion

Here, we analyzed barley core promoter sequences using precisely delineated TSSs by CAGE and a novel unbiased data analysis approach that takes into account positional restriction of the motifs comprised in the promoters and groups them by their overall sequence architecture. This approach is independent of previous motif knowledge and avoids non-conserved-noise detection, which arises when inaccurate promoter sequences and non-positioned motifs are examined. Our approach also keeps away from the previously emphasized distinction between sharp and broad promoters with the need to set the boundary between the two categories. However, we came to a similar conclusion to that in studies on metazoan promoteromes [5] that the presence of the TATA box is key to governing TSS selection in a more restricted region with CA initiator, whereas other W-box sequences or their absence leads to more relaxed or flexible TSS selection including different PyPu initiator configurations.

Although applied methods slightly differ, the proportion of promoters with the relaxed form of the TATA-box-like motif (around 20%) is similar in barley and human promoterome [54]. A

significantly larger proportion (38%) was observed for maize genes active in root and shoot tissues [23], which may reflect the fact that the TATA-box promoters are associated with tissue-specific expression primarily in adult tissues [55], contrasting to the early developmental stages analyzed in our study.

TATA-box position is not strictly conserved across species and the distance to initiator does not seem to be directly proportional to the genome size. The TATA-box-position range in barley (-29 to -36) is wider and slightly shifted compared to what is usually seen in metazoan genomes, where the most common position for a TATA box is tightly restricted at the 31st or the 30th nucleotide upstream from the TSS [54] and could be related to wider TCs associated with TATA box in plants compared to animals. Similar distances were reported in arabidopsis (around -30, [30]) and maize, sorghum, rice and wheat (-34 [22]). We also observed a positioned W-box variant not fitting the consensus TBP-motif PWM. It remains a question, whether other factors than TBP, both evolutionarily related and unrelated, are able to replace this key protein in the PIC as predicted by motif analysis tools, or the other way around: whether other motifs than TATA box get bound by TBP protein as reported from yeast [56]. The BREu, a C-rich element upstream of the TATA box, when inserted into an artificial promoter in the transgenic assay, was demonstrated to increase promoter strength in maize, unlike in tobacco, contrasting to the finding that the motif is absent in maize [30]. We found only a partial match to this element, leaving its significance in barley to be tested functionally.

The two main TSS selection modes relate to nucleosome positions, tissue specificity and epigenetic profile, all in line with the model of two distinct regulatory environments. However, due to the complicated interplay of plant hormones with both developmental and housekeeping functions, and the greater potency of plant cells, the distinction between housekeeping and developmental gene promoters is partially blurred as reflected by our GO-term analysis. Related to that, the bivalent chromatin states observed in the TATA promoters are known to allow timely activation while maintaining repression in the absence of differentiation signals [57]. Considering the relatively complex nature of our samples, the H3K27me3 presence could be also attributed to the repression of tissue-specific promoters in a subset of cell types and so the bivalent histone modification status would have to be confirmed or excluded by a sequential ChIP experiment.

Our comparison of the dominant TSSs from CAGE datasets to the most recent (MorexV3) barley annotation resulted in relatively frequent misalignment: 61% CTSSs were located more than 20 bp and 32% more than 100 bp apart from the aTSS. The discrepancy may be partly due to the promoter shifts between tissues and developmental stages since the embryonal samples examined in our study were only marginally represented in the RNA-seq dataset [46] that was used for the MorexV3 annotation. This points to the significance of generating a promoterome atlas for multiple tissues and cell types of a given organism, which allows focus on the relevant regulatory region in gene cloning and editing projects. In this context, CAGE appears as an affordable complementary technology that can be used to greatly improve even high-quality genome annotations that were based on RNA-seq data.

To understand the differences between CAGE and RNA-seq, we must consider that the two methods detect different parts of RNA molecules: capped 5′ ends and random RNA fragments, respectively [58]. The lower detection limit of CAGE can be explained by our TPM-value threshold used to filter out frequent low intergenic signals (disregarding whether background or real) and by the shallowness of the sequencing. Required sequencing depth for barley was calculated from multiple experiments on human cells, which could lead to an underestimation due to larger promoter widths and different intergenic transcription and background levels in barley.

In our study, we found 15.4, 14.1 and 13.6% intergenic CTSSs located >1000 bp apart from the nearest gene for 8DAP, 24DAP and 4DAG embryo, respectively. This proportion was smaller in barley than in rice, maize and sorghum, which had only 69-74% transcription start regions (TSRs) in <1000-bp distance from annotated genes, and wheat with a mere 49-54% TSRs in gene-proximal regions [22]. The difference may be due to distinct detection techniques (CAGE vs. Smar2C2) and data processing pipelines or may reflect a lower proportion of genes missing in the MorexV3 annotation.

Besides transcripts overlapping with the annotated promoters, we also found over 2,000 mainly single base-wide CTSSs originating from the first intron-exon boundary at the splice acceptor. These secondary-set transcripts, transcribed in a gene-sense direction, were previously detected in mammalian CAGE datasets and described as products of post-transcriptional cleavage and recapping. They were proposed to result in truncated mRNA isoforms, potentially translated to C- or N-terminal-truncated proteins [59]. The

authors called them ‘intraexonic CAGE tags’ and noted that they tended to be tissue-specific forms, suggestive of a tightly regulated process obviously increasing the mRNA abundance, same as observed here. Alternatively, these can occur as a by-product of co-transcriptional splicing or RNA polymerase slow-down and re-capping. Last, they could be a product of an intron-dependent looping and an associated mechanism of transcription enhancement, as reviewed in [60]. It will be interesting to subject these transcripts to further confirmation by RACE and detection of translation products at the protein level.

Our results prove that in barley, only a PyPu initiator, W box and Y patch - the previously known core promoter motifs - can be identified, complemented by stretches of low-complexity sequences with some degree of dynamics and flexibility during development. The promoter sequence architecture likely evolves by the exaptation of new regions including TEs or by changes in the ancient promoters by degeneration and binding-preference changes. We hypothesize that ancient cellular functions are more likely directed from the TATA box and simple TF-binding motif-containing sequences while more specialized functions are associated with less conserved promoter sequences enabling polymerase scanning [61], contacts with co-activators and initiation in response to distant TFBS.

Artificial promoter design or regulatory sequence manipulation are common engineering methods to drive or influence transcription levels. The knowledge of PIC interactions have been exploited in human to create a highly active core promoter, termed the ‘super core promoter’, that is capable of engaging in high-affinity interactions with TFIID through the presence of optimal versions of the TATA, Inr, MTE, and DPE motifs [17]. Similarly, a plant ‘super-promoter’ has been designed by [30] who also indicated that for optimal results, species-specific promoters might be desirable in transgenic design. Although a functional validation of each individual promoter type is beyond the scope of this article, in future, it can be easily addressed by an *in vitro* assay using a reporter gene, including tests of hormone, stress and nutrient effects on the promoter activity.

Our study, conducted on developing and germinating barley embryos, provided comprehensive information about transcription initiation at 21,610 genes active in the targeted stages. To increase that number and get more complete knowledge of transcription regulation in barley, we intend to generate and analyze additional CAGE datasets, namely

from floral tissues, expected to involve generative-tissue-specific regulatory mechanisms. Besides, we anticipate that species-specific promoter models built from a limited number of tissue-specific CAGE datasets could be used for genome-wide promoter prediction without the need to generate an exhaustive number of new datasets [62].

5. Code and data availability

CAGE data were deposited into the Gene Expression Omnibus (GEO) database under accession number GSE227219 (reviewer access token: **orcbgqwspzodrct**).
The data generated in the study will be shared in the EPD database (https://epd.expasy.org/epd/). The scripts, high-resolution figures and other large data files are available at: https://github.com/MorexV3CAGE.

Author Contributions

P.N. and H.S.: Project conceptualizaton; P.N.: Investigation; S.N., B.L. and S.P: Software design; S.P., P.N., and T.A. Formal analysis; M.K.: resources (RNA-seq data); P.N.: Writing - Original Draft; H.S. and S.P.: Writing - Review & Editing. P.N. and H.S.:
Funding acquisition.

Funding

This project was supported by the Czech Science Foundation [grant number 21-18794S] and by the European Regional Development Fund project “Plants as a tool for sustainable global development” [No. CZ.02.1.01/0.0/0.0/16_019/0000827].
Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Katerina Holusova and Helena Tvardikova for sequencing assistance and Zdenka Bursova for plant maintenance.

Appendix A. Supplemental Data

The following materials are available in the online version of this article.

Figure S1 CAGE data initial analysis: replica correlation, CAGEr analysis settings and TC annotation across stages

Figure S2 Comparison of 4DAG CAGE and RNAseq datasets

Figure S3 Clustering of 8DAP, 24DAP and 4DAG CAGE promoter sequence architectures

Figure S4 An example of a secondary CTSS

Figure S5 Heat maps for known core promoter sequence motifs in the primary consensus promoter clusters

Figure S6 Clustering of +/- 100bp CAGE promoter sequence architectures

Figure S7 Tissue specificity of individual stage-specific promoter clusters

Figure S8 Comparison of the published and the newly generated Morex GO-term annotation

Figure S9 Nucleosome positioning profiles around dominant CTSS in 24DAP embryo

Figure S10 Chromatin profiles and TE promoter occupancy

Data S1 A complete list of the TCs, including consensus-TC coordinates in the MorexV3 genome, the assigned-gene ID, position of the dominant TSS, feature annotation, classification, IQW and TPM

Data S2 Interactive correlation plots of CAGE and RNA-seq

Data S3 The new barley GO-term annotation in GAF format

Data S4 Genes with developmental TSS shifts between embryonic stages

Data S5 Interactive result from hierarchical matrix clustering of motifs found in all barley promoters

Table S1 The annotation of consensus promoters corresponding to Figure 1C and containing count values for each annotation category.

Table S2 Dynamics of the stage-specific promoter clusters across three stages of embryo development.

Table S3 TATA-box-like motif sets according to the level of correlation with the TATA-box PWM.

References

1. Burley SK, Roeder RG. Biochemistry and structural biology of transcription factor IID (TFIID). Annu Rev Biochem. 1996;65: 769–799.
2. Bergman DT, Jones TR, Liu V, Ray J, Jagoda E, Siraj L, et al. Compatibility rules of human enhancer and promoter sequences. Nature. 2022;607: 176–184.
3. Martinez-Ara M, Comoglio F, van Arensbergen J, van Steensel B. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. Mol Cell. 2022;82: 2519–2531.e6.
4. Neumayr C, Haberle V, Serebreni L, Karner K, Hendy O, Boija A, et al. Differential cofactor dependencies define distinct types of human enhancers. Nature. 2022;606: 406–413.
5. Haberle V, Lenhard B. Promoter architectures and developmental gene regulation. Semin Cell Dev Biol. 2016;57: 11–23.
6. Vo Ngoc L, Wang Y-L, Kassavetis GA, Kadonaga JT. The punctilious RNA polymerase II core promoter. Genes Dev. 2017;31: 1289–1301.
7. Smale ST, Baltimore D. The “initiator” as a transcription control element. Cell. 1989. pp. 103–113. doi:10.1016/0092-8674(89)90176-1
8. Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, et al. Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics. 2007;8: 67.
9. Cordon-Obras C, Gomez-Liñan C, Torres-Rusillo S, Vidal-Cobo I, Lopez-Farfan D, Barroso-Del Jesus A, et al. Identification of sequence-specific promoters driving polycistronic transcription initiation by RNA polymerase II in trypanosomes. Cell Rep. 2022;38: 110221.
10. Marbach-Bar N, Bahat A, Ashkenazi S, Golan-Mashiach M, Haimov O, Wu S-Y, et al. DTIE, a novel core promoter element that directs start site selection in TATA-less genes. Nucleic Acids Res. 2016;44: 1080–1094.
11. Bernard V, Brunaud V, Lecharny A. TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. BMC Genomics. 2010;11: 166.
12. Danks GB, Navratilova P, Lenhard B, Thompson EM. Distinct core promoter codes drive transcription initiation at key developmental transitions in a marine chordate. BMC Genomics. 2018;19: 164.
13. Shao W, Alcantara SG-M, Zeitlinger J. Reporter-ChIP-nexus reveals strong contribution of the Drosophila initiator sequence to RNA polymerase pausing. eLife. 2019. doi:10.7554/elife.41461
14. Vo Ngoc L, Cassidy CJ, Huang CY, Duttke SHC, Kadonaga JT. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. Genes Dev. 2017;31: 6–11.

15. Parry TJ, Theisen JWM, Hsu J-Y, Wang Y-L, Corcoran DL, Eustice M, et al. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. Genes Dev. 2010;24: 2013–2018.
16. Menand B, Desnos T, Nussaume L, Berger F, Bouchez D, Meyer C, et al. Expression and disruption of the Arabidopsis TOR (target of rapamycin) gene. Proc Natl Acad Sci U S A. 2002;99: 6422–6427.
17. Cianfrocco MA, Kassavetis GA, Grob P, Fang J, Juven-Gershon T, Kadonaga JT, et al. Human TFIID Binds to Core Promoter DNA in a Reorganized Structural State. Cell. 2013;152: 120–131.
18. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38: 626–635.
19. Deal RB, Henikoff S. Histone variants and modifications in plant gene regulation. Current Opinion in Plant Biology. 2011. pp. 116–122. doi:10.1016/j.pbi.2010.11.005
20. Vermeulen M, Mulder KW, Denissov S, Pijnappel WWMP, van Schaik FMA, Varier RA, et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell. 2007;131: 58–69.
21. Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, et al. Two independent transcription initiation codes overlap on vertebrate core promoters. Nature. 2014;507: 381–385.
22. Murray A, Mendieta JP, Vollmers C, Schmitz RJ. Simple and accurate transcriptional start site identification using Smar2C2 and examination of conserved promoter features. Plant J. 2022;112: 583–596.
23. Mejía-Guerra MK, Li W, Galeano NF, Vidal M, Gray J, Doseff AI, et al. Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites. Plant Cell. 2015;27: 3309–3320.
24. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003;100: 15776–15781.
25. Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. Plant J. 2009;60: 350–362.
26. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322: 1845–1848.
27. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309: 1559–1563.
28. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. Nature. 2014;507: 462–470.
29. Thieffry A, Vigh ML, Bornholdt J, Ivanov M, Brodersen P, Sandelin A. Characterization of Promoter Bidirectionality and Antisense RNAs by Inactivation of Nuclear RNA Decay Pathways. Plant Cell. 2020;32: 1845–1867.
30. Jores T, Tonnies J, Wrightsman T, Buckler ES, Cuperus JT, Fields S, et al. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. Nat Plants. 2021;7: 842–855.
31. Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, et al. Long-read sequence assembly: a technical evaluation in barley. Plant Cell. 2021;33: 1888–1906.

32. Navrátilová P, Toegelová H, Tulpová Z, Kuo Y-T, Stein N, Doležel J, et al. Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the MorexV3 reference genome. Plant Biotechnol J. 2022;20: 1373–1386.
33. Schreiber M, Mascher M, Wright J, Padmarasu S, Himmelbach A, Heavens D, et al. A Genome Assembly of the Barley “Transformation Reference” Cultivar Golden Promise. G3 . 2020;10: 1823–1827.
34. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 2015;43: e51.
35. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31: 2382–2383.
36. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424: 147–151.
37. Nikumbh S, Lenhard B. Identifying promoter sequence architectures via a chunking-based algorithm using non-negative matrix factorisation. bioRxiv. 2023. p. 2023.03.02.530868. doi:10.1101/2023.03.02.530868
38. Cavin Périer R, Junier T, Bucher P. The Eukaryotic Promoter Database EPD. Nucleic Acids Res. 1998;26: 353–357.
39. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–1018.
40. Wragg JW, Roos L, Vucenovic D, Cvetesic N, Lenhard B, Müller F. Embryonic tissue differentiation is characterized by transitions in cell cycle dynamic-associated core promoter regulation. Nucleic Acids Res. 2020;48: 8374–8392.
41. Santana-Garcia W, Castro-Mondragon JA, Padilla-Gálvez M, Nguyen NTT, Elizondo-Salas A, Ksouri N, et al. RSAT 2022: regulatory sequence analysis tools. Nucleic Acids Res. 2022. doi:10.1093/nar/gkac312
42. Sebastian A, Contreras-Moreira B. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. Bioinformatics. 2014;30: 258–265.
43. Xiao J, Jin R, Yu X, Shen M, Wagner JD, Pai A, et al. Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. Nat Genet. 2017;49: 1546–1552.
44. Pedersen DS, Coppens F, Ma L, Antosch M, Marktl B, Merkle T, et al. The plant-specific family of DNA-binding proteins containing three HMG-box domains interacts with mitotic and meiotic chromosomes. New Phytol. 2011;192: 577–589.
45. Philippe L, van den Elzen AMG, Watson MJ, Thoreen CC. Global analysis of LARP1 translation targets reveals tunable and dynamic features of 5′ TOP motifs. Proceedings of the National Academy of Sciences. 2020. pp. 5319–5328. doi:10.1073/pnas.1912864117
46. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544: 427–433.
47. Milne L, Bayer M, Rapazote-Flores P, Mayer C-D, Waugh R, Simpson CG. EORNA, a barley gene and transcript abundance database. Sci Data. 2021;8: 90.
48. Wimalanathan K, Lawrence-Dill CJ. Gene Ontology Meta Annotator for Plants (GOMAP). doi:10.1101/809988
49. Boecker F. AHRD: Automatically annotate proteins with human readable descriptions and Gene Ontology terms. Universitäts- und Landesbibliothek Bonn. 2021. Available:

https://bonndoc.ulb.uni-bonn.de/xmlui/handle/20.500.11811/9344
50. Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, et al. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet. 2011;7: e1001274.
51. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9: 215–216.
52. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489: 101–108.
53. Wicker T, Schulman AH, Tanskanen J, Spannagl M, Twardziok S, Mascher M, et al. The repetitive landscape of the 5100 Mbp barley genome. Mob DNA. 2017;8: 22.
54. Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biol. 2006;7: 1–18.
55. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19: 621–637.
56. Seizl M, Hartmann H, Hoeg F, Kurth F, Martin DE, Söding J, et al. A Conserved GA Element in TATA-Less RNA Polymerase II Promoters. PLoS One. 2011;6. doi:10.1371/journal.pone.0027595
57. Voigt P, Tee W-W, Reinberg D. A double take on bivalent promoters. Genes Dev. 2013;27: 1318–1338.
58. Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, et al. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. Genome Res. 2014;24: 708–717.
59. Mercer TR, Dinger ME, Bracken CP, Kolle G, Szubert JM, Korbie DJ, et al. Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. Genome Res. 2010;20: 1639–1650.
60. Dwyer K, Agarwal N, Gega A, Ansari A. Proximity to the Promoter and Terminator Regions Regulates the Transcription Enhancement Potential of an Intron. Front Mol Biosci. 2021;8: 712639.
61. Qiu C, Jin H, Vvedenskaya I, Llenas JA, Zhao T, Malik I, et al. Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. Genome Biol. 2020;21: 1–31.
62. Wang Y, Peng Q, Mou X, Wang X, Li H, Han T, et al. A successful hybrid deep learning model aiming at promoter identification. BMC Bioinformatics. 2022;23: 1–20.
63. Kovacik M, Nowicka A, Pecinka A. Isolation of High Purity Tissues from Developing Barley Seeds. J Vis Exp. 2020. doi:10.3791/61681
64. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6: e21800.
65. Neumann P, Navrátilová A, Schroeder-Reiter E, Koblížková A, Steinbauerová V, Chocholová E, et al. Stretching the rules: monocentric chromosomes with multiple centromere domains. PLoS Genet. 2012;8: e1002777.
66. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12: 2478–2492.
67. Applied Research Applied Research Press. RSEM: Accurate Transcript Quantification from RNA-Seq Data with Or Without a Reference Genome. 2015.
68. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12: 323

Figures

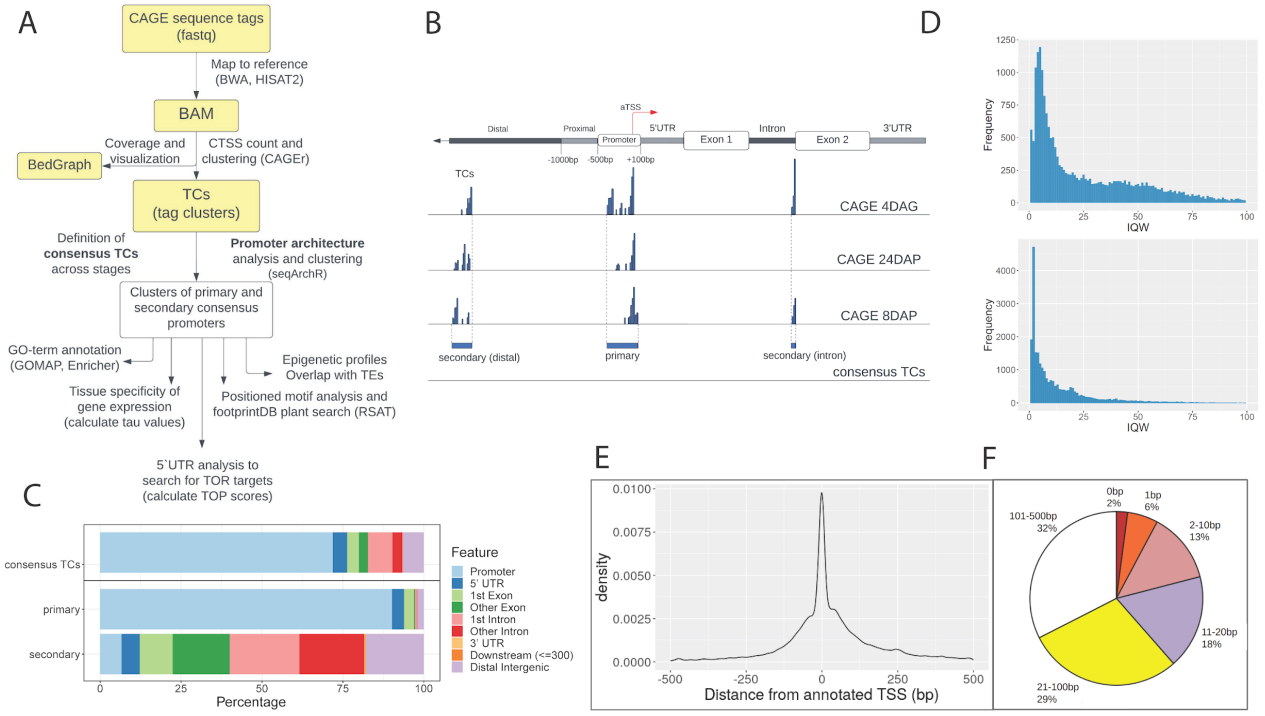


Figure 1. Initial analysis of the barley CAGE dataset. A) CAGE data analysis workflow. **B**) Definition of the consensus tag clusters (TCs) and of the primary and the secondary promoter set. The aTSS corresponds to the TSS as in MorexV3 annotation. **C**) Annotation of consensus promoter candidates from CAGEr. The consensus TCs dataset was split into the primary and secondary sets using a filtering method described in the main text. **D)** Distribution of 24DAP promoter interquantile-width (IQW) values for primary (top) and secondary (bottom) promoters. **E**) Distribution of distances between dominant CTSSs and aTSSs for barley genes expressed in 8DAP, 24DAP and 4DAG embryos. **F**) Proportions of expressed genes with a dominant CTSS present within the designated distance from the aTSS. Analysis in **E**, **F**) was done for the dominant CTSSs whose distance from the aTSS was not greater than 500 bp.

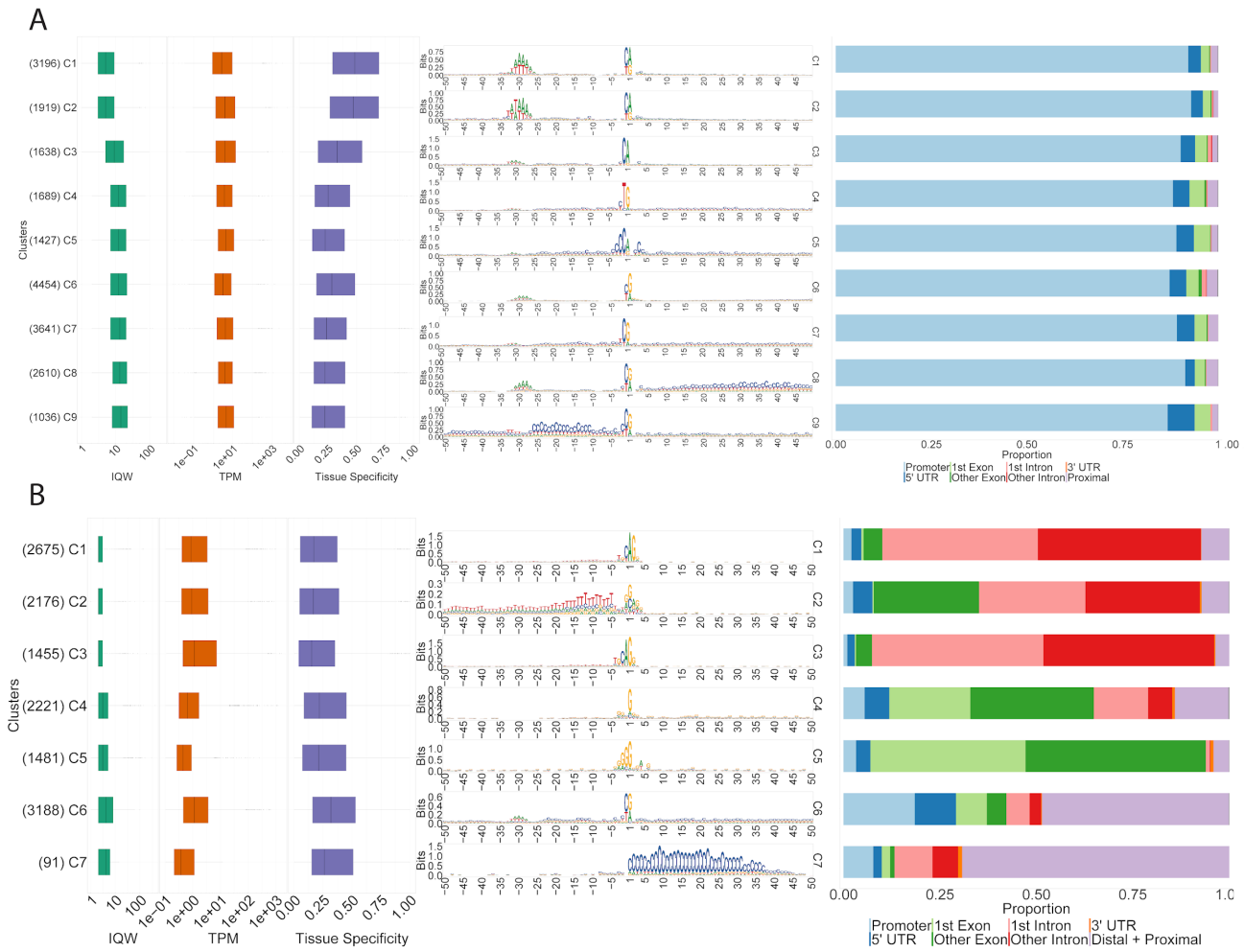


Figure 2: Clustering of CAGE promoter sequence architectures. Clusters of consensus promoters were generated by the seqArchR algorithm and ordered by median interquantile widths (IQWs). The composed plots for **A**) primary and **B**) secondary promoter clusters include boxplots for IQW, gene expression level values (tags per million (TPM), log-transformed) and tissue specificity (tau values), followed by sequence logos and genomic feature annotation bar plots. The numbers of genes per cluster are given in parentheses.

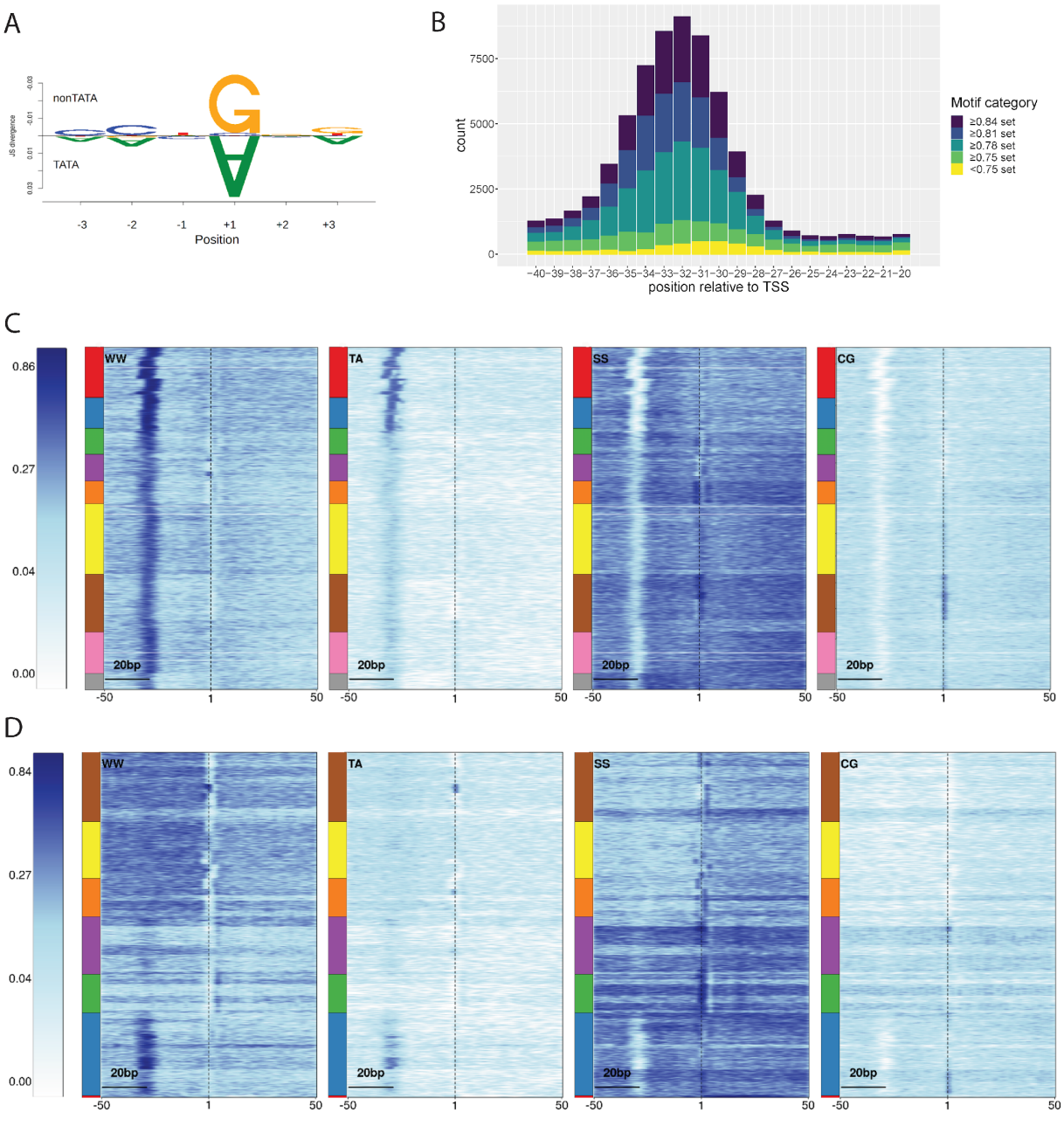


Figure 3. Sequence analysis of the initiator, TATA-box and dinucleotide content in barley promoter clusters. A) TATA vs. non-TATA box promoters differ in the +1 position (A vs. G). **B**) Variability in the sequence and position of TATA-like motifs in the consensus promoter set. The motif categories are based on the degree of Pearson correlation with the canonical TATA-box position weight matrix (PWM). Pentamers included in particular sets are listed in Table S3. **C**, **D**) Dinucleotide motif heat maps for the primary **C**) and the secondary **D**) promoter clusters. The multi-coloured bars left of the heat maps indicate boundaries between the clusters, ordered as in Figure 2A and 2B, respectively. Note the presence of the W box in all primary promoters and the differences in CG distributions.

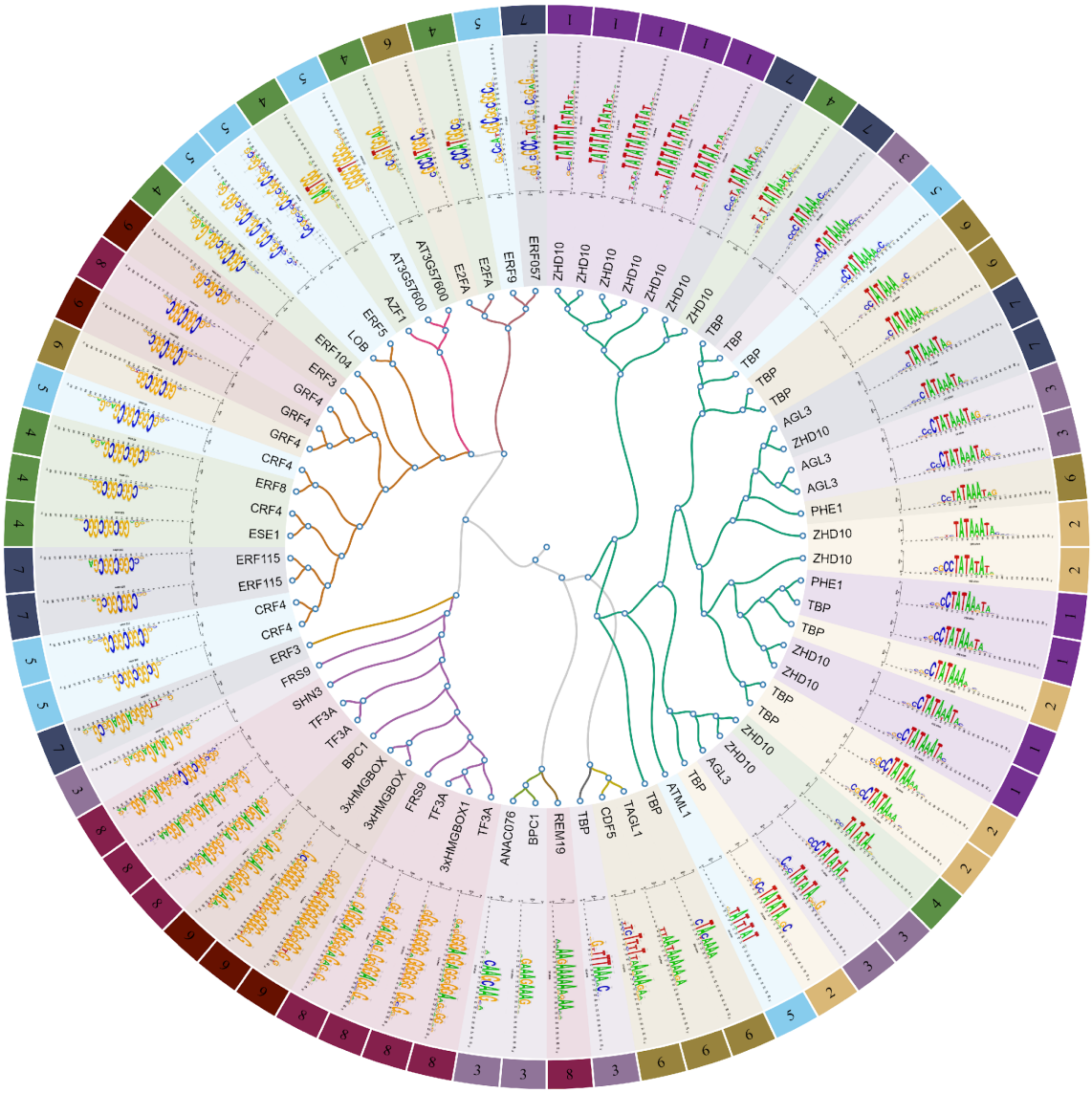


Figure 4. Positioned sequence motifs in consensus core promoters. Region +/-50 bp relative to the TSS was searched for TFBS motifs collected in footprintDB.plants database. The proteins at the tree branches correspond to the highest probability hits of the motif search. Colors and numbers in the outer circle correspond to individual primary promoter clusters.

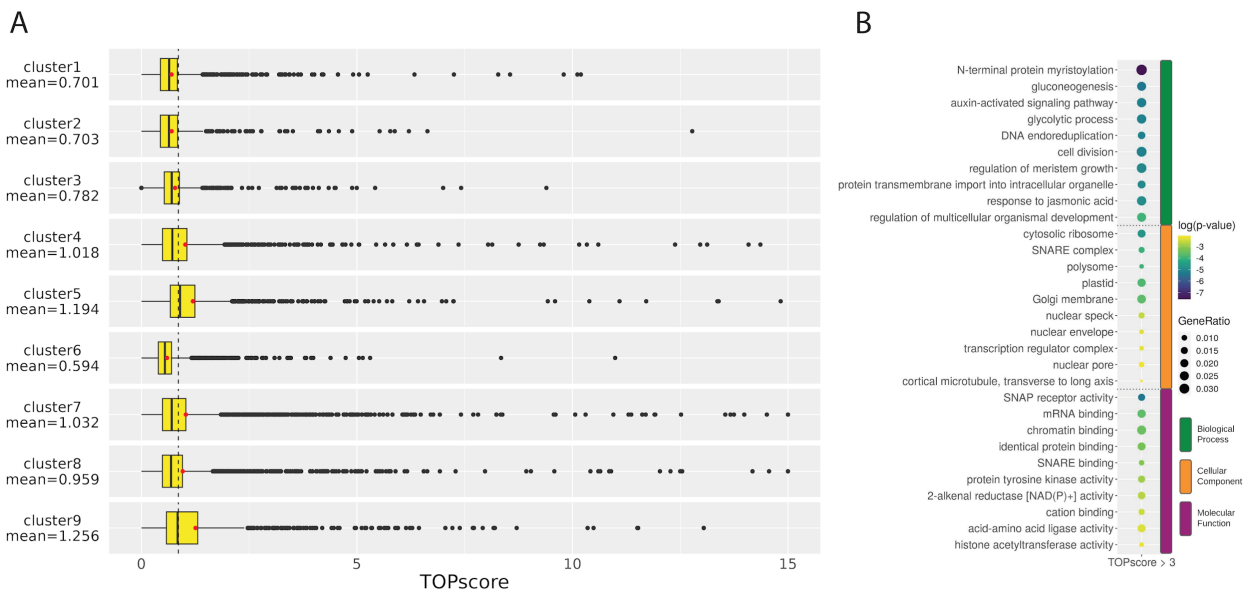


Figure 5. TOP motif analysis across promoter clusters. A) TOPscore distributions in the primary consensus clusters. Each of the clusters had either a significantly lower or a significantly higher (p-value < 0.001, Welch Two Sample t-test) TOPscore mean (red dot) compared to the mean value of the whole dataset (mean=0.86, shown as a dashed line). **B**) GO enrichment analysis of genes with promoters containing the candidate TOP motif (TOPscore >3).

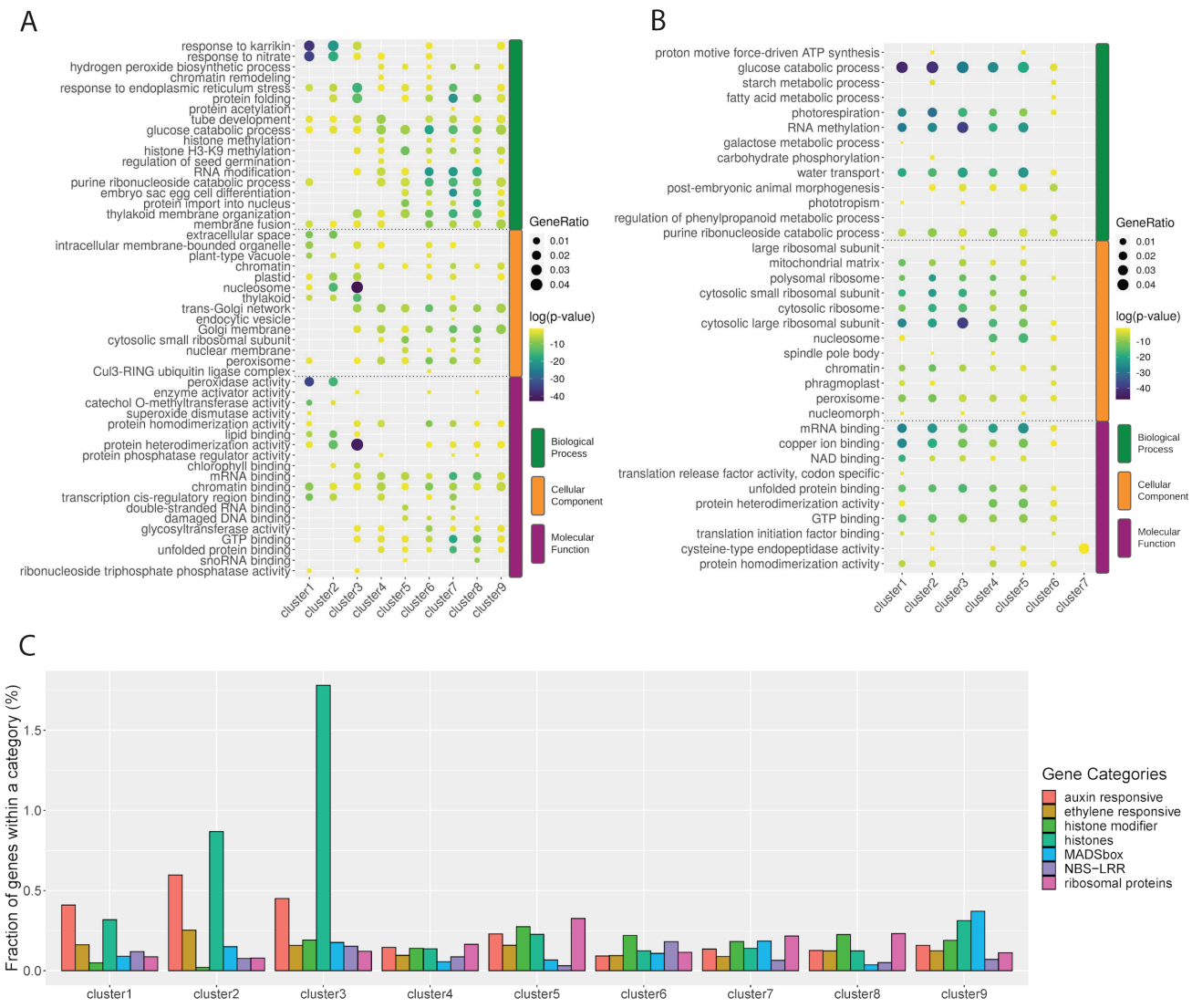


Figure 6. GO enrichment analysis. The plot shows the top five GO terms for primary **A**) and secondary **B**) consensus clusters. Clusters 1-3 in A) represent TATA-box clusters. **C**) Representation of selected gene categories in the clusters of the primary set.

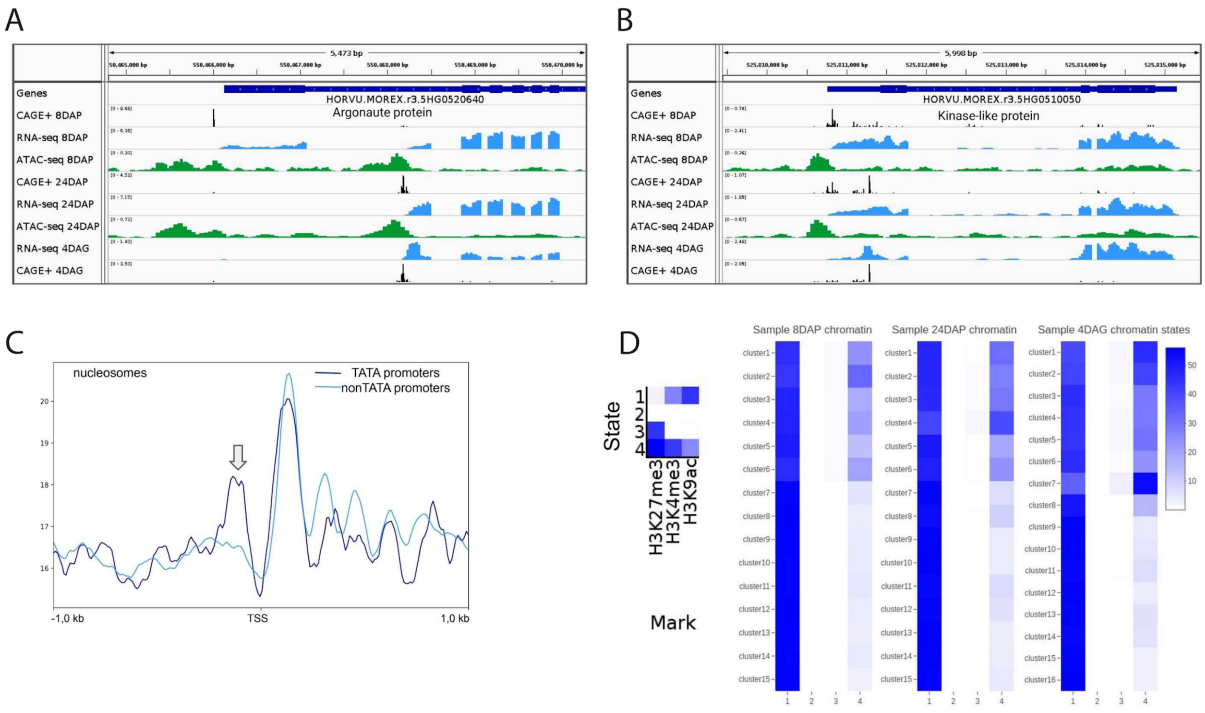


Figure 7. Promoter shifts and epigenetic context of barley promoters. A) An example of a developmentally regulated alternative first exon of an Argonaute protein gene. Relating dynamics of the epigenetic environment are illustrated by open-chromatin profiles (green peaks) for three developmental stages of the embryo. **B**) An example of an alternative TSS shifted during development and changing the 5`UTR sequence. Differences in RNA-seq and open-chromatin profiles confirm the shift. **C**) Nucleosome positioning of the TATA (merged clusters 1-3 of the primary consensus set) and non-TATA promoters (clusters 4-9). TATA-box presence is correlated with a well-positioned upstream nucleosome (arrow), in contrast to the nucleosome-free region upstream and well-positioned nucleosomes downstream the TSS in promoters without the TATA box. **D**) ChromHMM emission parameters showing four-chromatin-state composition generated from 8DAP, 24DAP and 4DAG ChIP-seq datasets of three key histone modifications followed by heat maps showing enrichment of the four chromatin states across the stage-specific promoter clusters.