Widespread transcriptional regulation from within transcribed regions in plants

Yoav Voichek[#,1], Gabriela Hristova[1], Almudena Mollá-Morales[1], Detlef Weigel[2], Magnus Nordborg[#,1]

[1] Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter (VBC), Vienna, Austria
[2] Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany
Corresponding author. E-mail: yoav.voichek@gmi.oeaw.ac.at & magnus.nordborg@gmi.oeaw.ac.at

Abstract:

Animals and plants have evolved separately for over 1.5 billion years and independently invented multicellularity, suggesting a possible divergence in their mode of transcription regulation. Here, we set out to elucidate fundamental features of transcription regulatory sequences in plants. Using a massively parallel reporter assay in four species, we show that sequences downstream of the transcription start site (TSS) play a major role in controlling transcription. Swapping regulatory sequences from one side of the TSS to the other yields different outcomes, unlike animal enhancers which act independently of their position. A GATC-containing DNA motif, positioned downstream of the TSS, was sufficient to enhance gene expression in a dose-dependent manner. Its effect on gene expression was tissue-dependent and conserved across vascular plants. These results identify a unique characteristic of transcriptional regulation in plants, and suggest fundamental differences in gene regulation might exist between plants and animals.

Introduction:

Differences in gene expression are the basis of diverse morphologies[1] and finely tuned responses to external stimuli. Given the independent evolution of multicellularity in plants and animals, their contrasting developmental patterns, and their distinct responses to environmental cues, it would not be surprising to find profound differences in their mechanisms for the regulation of gene expression. Indeed, many aspects of transcriptional regulation in plants seem to be unique, including special core promoter DNA-motifs[2], expanded[3] and new[4] families of specific and general transcription factors (TFs), and long-range enhancers with different features from animals[5,6]. Yet, how regulatory sequences function and are organized near genes is widely assumed to be similar to animals and fungi[7]. Here, we reveal new aspects of transcriptional regulation based on genome-wide, unbiased tests for regulatory capacity of sequences near genes in four different species of flowering plants.

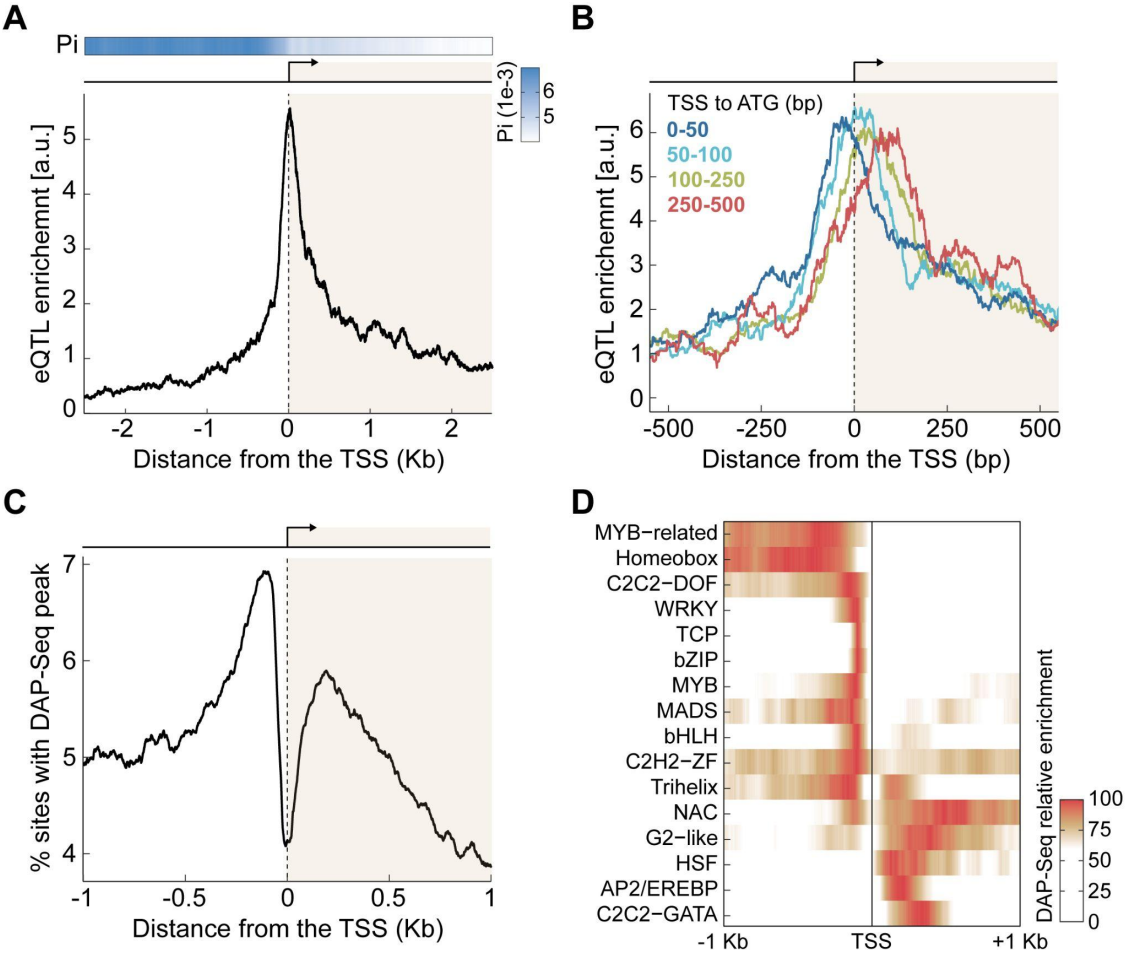
Results:

We first set out to determine the typical locations of regulatory regions near genes in Arabidopsis by large-scale mapping of expression quantitative trait loci (eQTL). Therefore, we analyzed genotypic and transcriptomic data from the Arabidopsis 1,001 Genomes Project to identify cis-eQTL within 10 Kb of each gene[8,9]. While we expected to find most eQTL in proximal promoters upstream of the transcription start site (TSS), we discovered a similar proportion of eQTL downstream of the TSS (Fig. 1A). As eQTL are more likely to occur where the density of single-nucleotide polymorphisms (SNPs) is higher, the lower sequence diversity downstream of TSSs made the downstream eQTL enrichment even more unexpected (Fig. 1A). This pattern was consistent across multiple gene expression datasets, and accounting for linkage between SNPs only intensified it (Fig. S1). Our eQTL analysis pointed to a potential central regulatory region downstream of the TSS in Arabidopsis.

We next examined possible explanations for the observed eQTL distribution. If eQTL within transcripts are in sequences controlling mRNA stability, eQTL should be more frequent in exons than introns, which are removed by splicing. While this is what is seen for human eQTL[10], we observed no such preference for exons in Arabidopsis (Fig. S2A-B). eQTL were also not enriched toward the end of transcripts (Fig. S3), even though 3’ untranslated regions (3’ UTRs) have known roles in controlling mRNA stability[11,12]. Finally, we asked whether eQTL were less likely to occur within coding

regions, which have strong sequence constraints. Indeed, eQTL tended to be most frequent just downstream of the TSS for genes with longer TSS-to-ATG distances (defined as 5' UTRs including introns in the 5' UTRs, Fig. S2C) and just upstream of the TSS for genes with shorter TSS-to-ATG distances (Fig. 1B). These findings suggest that the proposed downstream regulatory regions are enriched between the TSS and the start codon and affect transcription rather than mRNA stability.

Figure 1: Indications for transcription-regulatory regions downstream of the TSS



(A) eQTL enrichment (below) and nucleotide diversity (Pi, above) near gene TSS **(B)** eQTL enrichment for genes with different TSS-to-ATG distances. Group counts: 1,088 (0-50), 968 (50-100), 1,122 (100-250), and 577 (250-500). **(C)** Proportion of sites with a DAP-Seq peak center, from 529 TFs analyzed. **(D)** DAP-Seq peak enrichment, as in **C**, consolidated for each TF family with at least 10 members in the dataset; maximum signal for each TF family was scaled to 100. In **A-C** data smoothed using a 100 bp rolling window. a.u., arbitrary units.

If the location of eQTL in the proximity of genes results from variation in transcription across Arabidopsis strains, then chromatin and TFs are likely to be involved. The first observation in agreement with this was that histones H3.1 and H3.3 enrichment downstream of the TSS was moved away from the TSS as TSS-to-ATG distances increased (Fig. S4). Second, binding sites of 529 TFs, as measured by DNA affinity purification sequencing (DAP-Seq)[13], showed two peaks, upstream and downstream of the TSS (Fig. 1C). Individual TFs had a preference for binding on only one side of the TSS, with similar preferences for members of the same TF family (Fig. 1D & S5). In-vivo data of three TFs binding confirmed the preference of TFs to bind on either side of the TSS (Fig. S6). We do not think that inaccurate TSS annotations greatly affect our results, given that the clear dip in TF binding sites is centered on annotated TSSs. These analyses support a transcription-regulatory region downstream of the TSS.

To systematically investigate the role of sequences downstream of the TSS in controlling gene expression, we designed a massively parallel reporter assay[14] (MPRA, Fig. 2A). We synthesized 12,000 160-long bp fragments, derived from regions 40-200 bp upstream or 40-360 bp downstream of the TSSs of highly expressed Arabidopsis genes, excluding 80 bp around the TSS (-40 bp to 40 bp) which contain the core promoter. Downstream-derived fragments included exons and introns, with only donor and acceptor splicing sites excluded. We inserted these fragments on either side of the TSS of a GFP reporter gene. Insertion-free constructs served as controls. The downstream insertion site was located in an intron of the reporter gene, to rule out effects due to altered sequence of the mature mRNA and thus minimizing effects on mRNA stability. For robust quantification, multiple variants were generated for each insertion, with a 15 bp random barcode within the transcript. Barcodes and tested regulatory fragments were linked by DNA sequencing, and transcriptional activity was read out by RNA sequencing.

We used two different GFP reporter constructs to provide different contexts (Fig. S7): the 46 bp CaMV 35S minimal promoter, commonly used to test plant enhancers, in combination with a short synthetic 5’ UTR, and a 700 bp Arabidopsis *TRP1* promoter fragment and 5’ UTR, previously used to study the effect of introns on gene expression[15]. In order to derive conclusions with broad applicability to flowering plants,

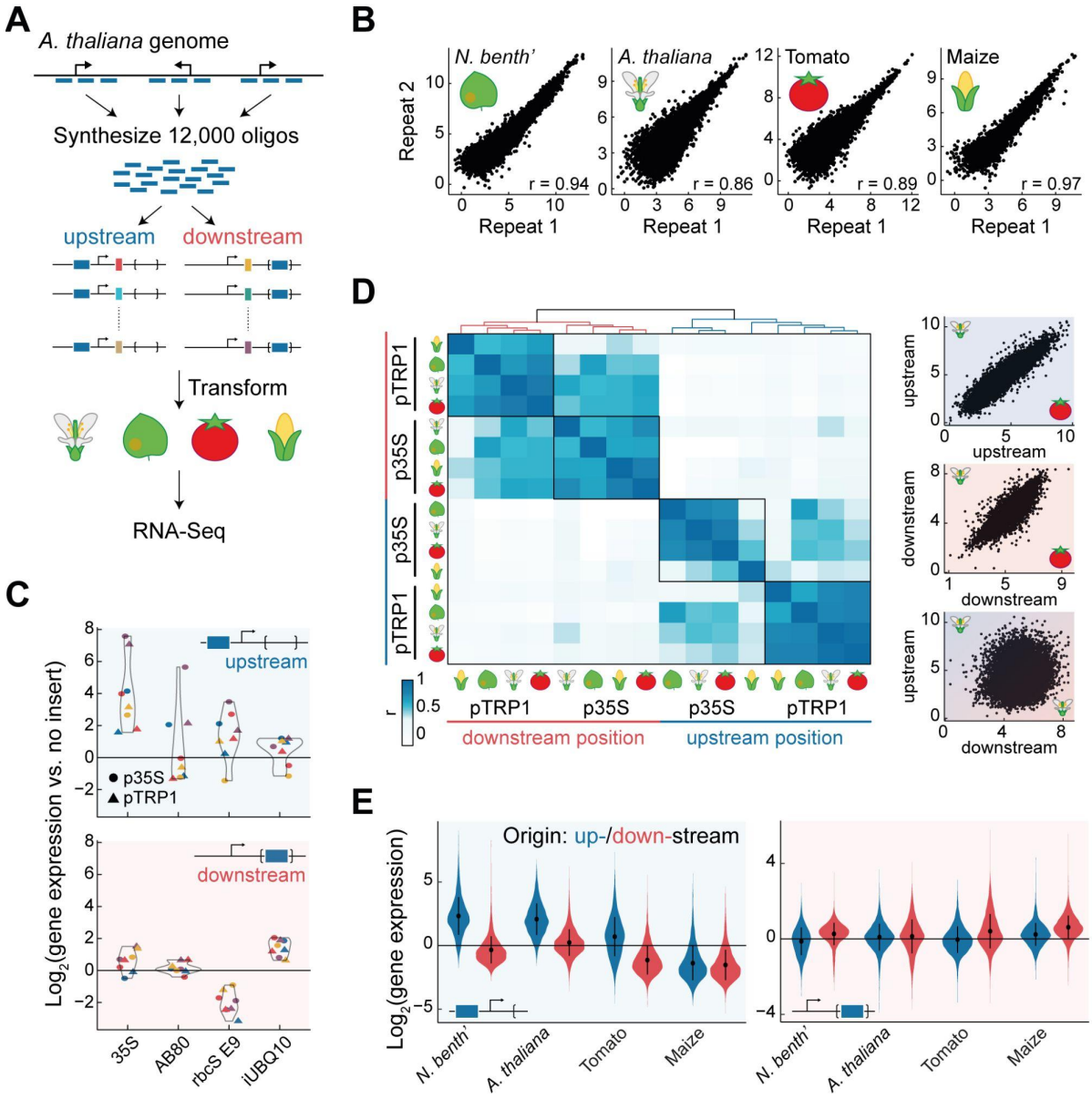
we quantified activity of the libraries in four different species - in Arabidopsis, tomato, and maize using transfection of leaf protoplasts, and in *N. benthamiana* using leaf infiltration of *Agrobacterium tumefaciens* into mesophyll cells. We reasoned that the use of two different transformation methods would increase the robustness of our conclusions. Reproducibility was ensured through three to four replicated experiments (Fig. 2B and S8).

As a control, a small fraction of the synthesized fragments were parts of known enhancers, previously examined in MPRA[14]. In most cases, these fragments increased expression when placed upstream of the TSS (Fig. 2C), but not when placed inside introns, in agreement with previous results[14]. Segments of the UBQ10 intron, known to enhance expression[16], were also included in the library. These intron-derived fragments drove higher expression when inserted into the intron of the reporter gene rather than when inserted upstream of the TSS (Fig. 2C).

The position-dependent effects observed for the known enhancers seem to be common. We found that fragments had similar activity independent of which species they were assayed, promoter, or how they were introduced into the host cell (Fig. 2D). In contrast, the relative activity greatly changed when the same fragment was inserted upstream or downstream of the TSS, even when using the same backbone and species. These results suggest that, unlike the position-independence seen for animal enhancers, the activity of flowering plant enhancers is strongly dependent on their position relative to the TSS.

Furthermore, the original genomic location of the fragment played a substantial role. Generally, fragments increased expression when positioned in their original position relative to the TSS, but the extent varied between backbone and species (Fig. 2E and S9). Enhancers were more effective in the CaMV 35S promoter construct, than in the *TRP1* promoter construct, in which fragments insertions disrupted the *TRP1* genomic sequence (Fig. S9). In maize fragments often reduced reporter expression when inserted upstream, regardless of genomic origin. This finding, juxtaposed with the strong correlation among fragments relative activity from all libraries, suggests that while absolute levels are strongly influenced by backbone and species, the relative effects of different fragments in the same position are similar across species.

Figure 2: Position-dependent enhancers reside inside transcribe regions



(A) Massively Parallel Reporter Assay (MPRA) overview: 12,000 fragments (160 bp), originating from upstream or downstream of the TSS of Arabidopsis genes, were synthesized, pooled, and inserted upstream of the TSS or within the intron of a reporter gene, and tagged by barcodes. Following transient transformation into one of four species, barcoded RNA sequencing quantified expression. **(B)** High reproducibility in MPRA experiment replicates, demonstrated here for CaMV 35S-minimal promoter-based libraries. Pearson's correlation coefficients r are indicated. **(C)** Expression comparison of constructs with control enhancer fragments and no-insert constructs, for upstream (top) and downstream (bottom) insertions. Depicted for *N. benthamiana* (purple), *A. thaliana* (blue), tomato (red), and maize (yellow); construct backgrounds are shape-coded. **(D)** Left, Pearson's correlation coefficients between all libraries,

hierarchically clustered. Construct background and insertion position are indicated. Right, activity of pTRP1-based constructs compared between Arabidopsis and tomato for upstream-upstream (top, r=0.92), downstream-downstream libraries (middle, r=0.85), and Arabidopsis-Arabidopsis for upstream-downstream (bottom, r=0.13). **(E)** Comparison of activity of upstream- (blue) and downstream-derived (red) fragments relative to no-insertion constructs, when fragments are positioned upstream (left) or downstream (right) of the TSS. Constructs are p35S-based. Error bars depict the mean and ±1 standard deviation.

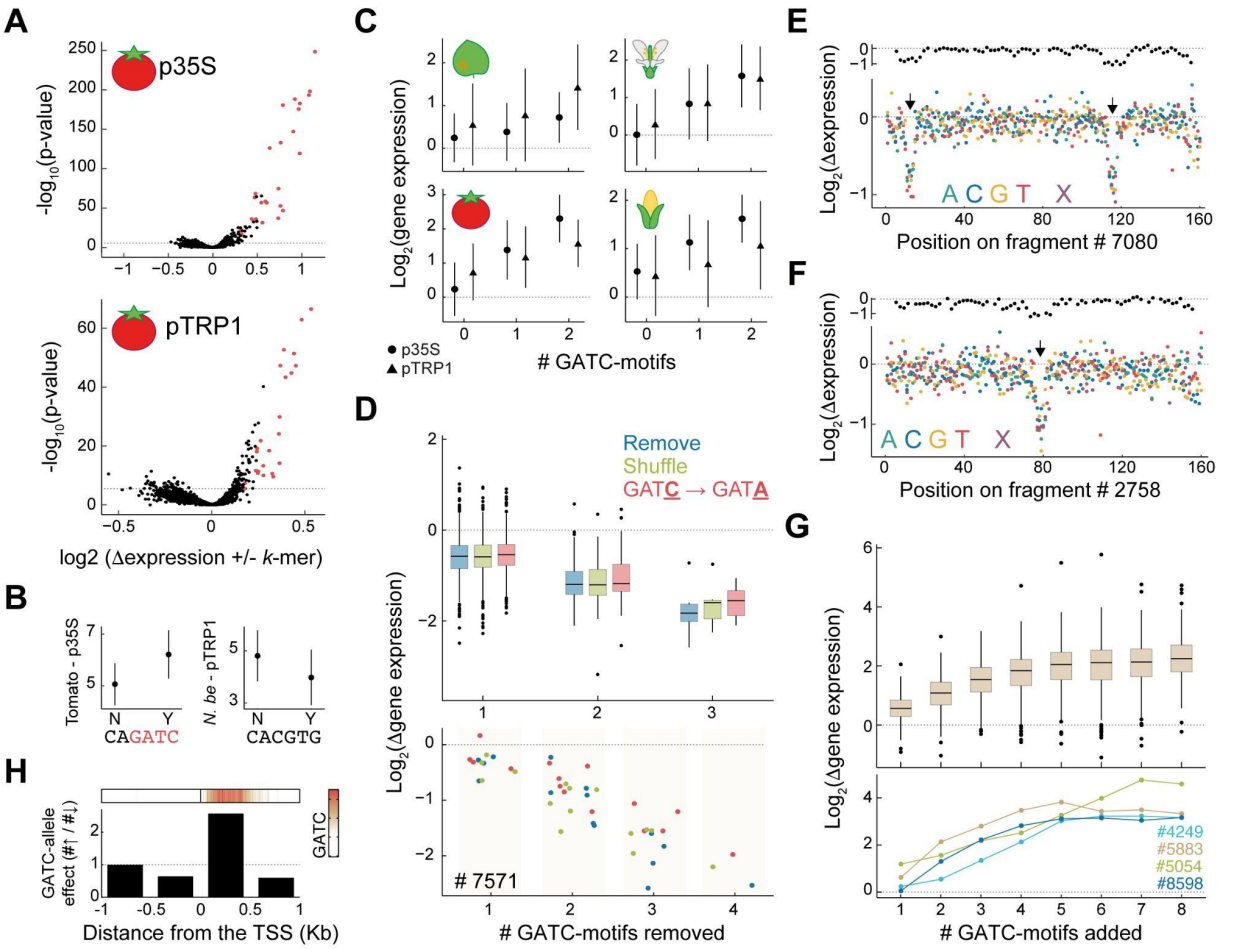
An immediate question that arises from our observation is how sequences downstream of the TSS control transcription activity. We hypothesized that transcription factors that promote expression through binding DNA motifs may determine the transcriptional activity of the region. Thus, we searched for 6-bp sequences (6-mers) whose presence downstream was associated with increased or decreased expression (Fig. 3A-B & S10A-B). We found more 6-mers that promoted expression than ones that repressed it, in agreement with downstream fragments' propensity to increase expression when positioned downstream. These 6-mers are thus potentially part of sequence motifs bound by TFs downstream to the TSS.

Across species and backbones, 6-mers including a GATC sequence had the strongest effect (Fig. 3A, S10A-B). To quantify the GATC effect we combined the six 6-mers with the strongest effect into a 8-bp YVGATCBR motif (Y=CT, V=ACG, B=CGT, R=AG, Fig. S11). Expression levels increased with the number of GATC motifs in the fragment, each copy increasing expression level on average by almost 50% (Fig. 3C).

To investigate the effects of the GATC motif further, we synthesized 18,000 additional oligonucleotides, each a variant of a fragment from our initial pool as described below. These fragments were inserted downstream of the TSS in both backbones, and their effect on gene expression were measured in Arabidopsis protoplasts across three replicates (Fig. S12). First, we tested the requirement of the motif by focusing on 841 downstream-derived fragments containing a GATC motif. By deleting, shuffling, or modifying the core GATC to GATA, we effectively removed these motifs. We found that such removal led to an average 50% decrease in gene expression, regardless of mutation type (Fig. 3D, S13A).

To supplement the GATC-focused mutation analysis, we conducted a deep mutational scan of 13 downstream-derived fragments. For each, we (i) deleted every set of 10 consecutive base pairs, and (ii) either mutated each nucleotide to its three alternatives or deleted it. This resulted in 736 derivatives from each original fragment. Any change to the core GATC motif decreased activity, underscoring the motif's strict constraints (Fig. 3E-F, S14-15). As expected, these analyses also revealed additional sequences that do not include GATC motifs as important for enhancing activity of the tested fragments.

Figure 3: A GATC motif is sufficient to increase expression when positioned downstream of the TSS



(A) Association of 6-mers with downstream-MPRA activity. Each 6-mer's presence (or absence) in downstream-derived fragments was compared based on the difference in average $\log_2$ expression (x-axis) and Mann-Whitney U test p-value (y-axis). Displayed for p35- (top) and pTRP1-based (bottom) constructs in tomato. Red points denote 6-mers with GATC; dashed line marks the 5% Bonferroni multiple testing

threshold. **(B)** Activity distribution for downstream-derived fragments with (Y) or without (N) 6-mers when inserted downstream. Plotted for CAGATC in tomato with p35S-based construct (left), and for CACGTG (G-box[17]) in *N. benthamiana* with pTRP1-based construct (right). **(C)** Relative activity of downstream fragments when inserted downstream, as a function of the number of YVGATCBR consensus motifs in the tested fragments. Group sizes: 6,855 (no motif), 956 (1 motif), and 119 (2 motifs). Backbone and species are indicated. **(D-G)** Activity difference in the p35S-downstream library between original and mutated fragments. **(D)** Effect of removing GATC motifs on activity of 823 GATC-containing fragments (top), and one specific example with 4 motifs (bottom). Motifs were removed by deletion, 8 bp shuffle, or GATC-to-GATA mutation. **(E)** Deep-mutagenesis of fragment #7080: effects of 10 bp deletions (top), 1 bp deletions (X), or 1 bp mutations (bottom). Arrows highlight GATC motifs. **(F)** As in E, for fragment #2578. **(G)** Effect of adding GATC-motifs for 221 fragments with incremental motif additions (top), and four specific examples (bottom). **(H)** GATC-motif-gain/loss alleles in the 1,001 Genomes population of accessions linked to nearby gene expression. The bar-graph showcases the ratio of number of significant associations with higher vs. lower expression in the GATC-motif allele, grouped by distance to the TSS (bottom). Top, GATC motif's enrichment in proximity to the TSS. Error bars in B-C represent mean ± 1 standard deviation. Boxplots in D and G display median, IQR, and 1.5x IQR, with outliers as points.

We next explored the GATC motifs' sufficiency for enhancing expression. We took a random set of 221 fragments from our initial set (166 downstream- and 55 upstream-derived) and incrementally added 1 to 8 GATC motifs to the fragments. Expression consistently increased with each added copy, even for upstream-derived fragments (Fig. 3G, S13B-C). Remarkably, 97% of these fragments enhanced expression once at least 4 GATC motifs were added. The enhancement was a function of base activity of each fragment, with the increased activity of highly active fragments becoming saturated after a single GATC addition, and the activity of the initially least active fragments remained unsaturated even after adding eight GATC motifs (Fig. S13D). This finding suggests that GATC and non-GATC activity-enhancing sequences might work through the same mechanism to increase expression of the reporter constructs.

Finally, to confirm our synthetic MPRA mutational analysis of the GATC motif, we explored the effects of natural variation in the GATC motif by returning to the Arabidopsis 1001 Genomes Project data[8,9]. We identified gains and losses of GATC motifs near TSSs, and asked how these correlated with expression of the affected genes. We categorized significant associations based on whether the allele with the GATC motif had higher or lower expression. Consistent with our MPRA findings, an

enrichment of higher expression was observed exclusively in the GATC-motif allele situated downstream of the TSS, particularly within the initial 500 bp (Fig. 3H, S16). Intriguingly, this is also where the GATC motif is predominantly found, reinforcing its role in enhancing gene expression when located downstream of the TSS.

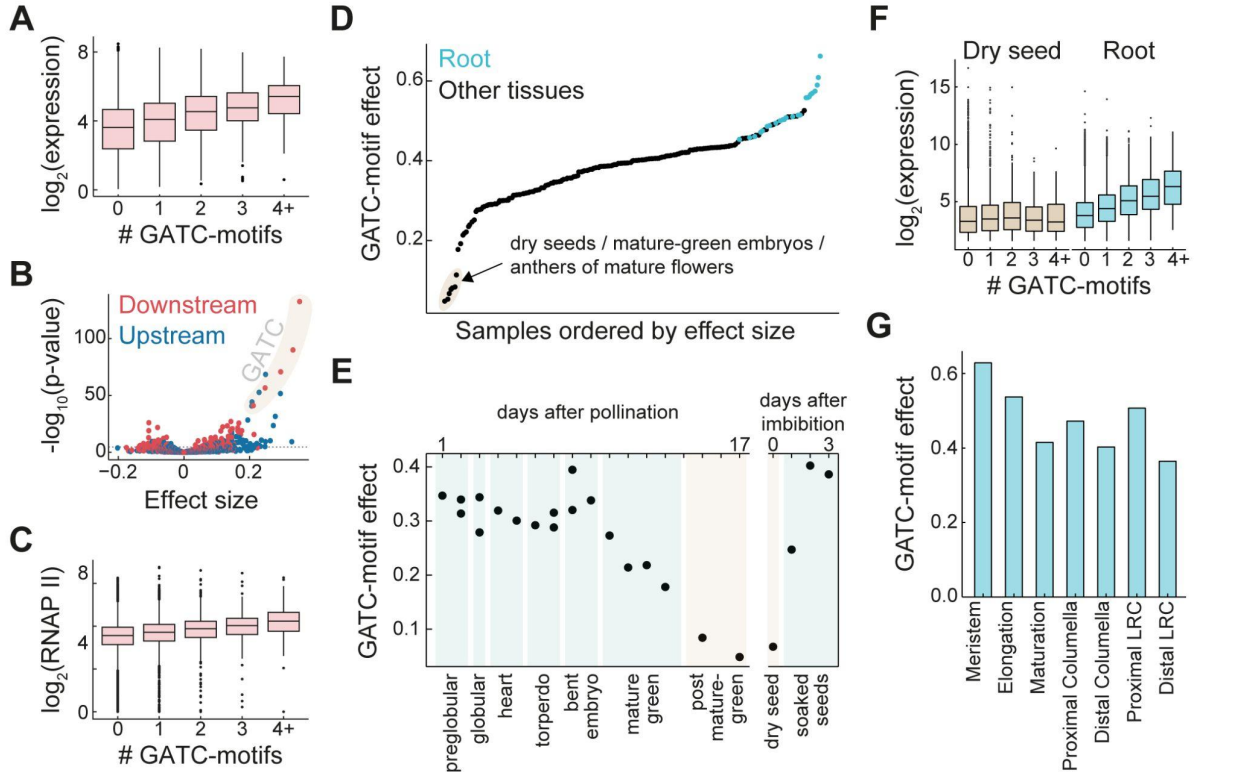
In plants, the GATC motif is recognized by GATA TFs[13] (see supplementary note S1), which are linked to diverse biological functions[18]. The Arabidopsis genome encodes 30 of these TFs. DAP-Seq data reveal GATA factor binding enrichment within 500 bp downstream of the TSS (Fig 1D, 3H). In this region, 7,397 genes have at least one GATC motif (Fig. S17A). Gene ontology analysis shows these genes to be enriched in processes related to the Golgi apparatus, endoplasmic reticulum, endosomes, and vesicle-mediated transport (table S6). Given its prevalence and association with the secretion system, the GATC motif likely acts as a widespread regulatory signal in diverse biological functions.

Enhancer sequences typically consist of multiple DNA motifs that are targeted by specific TFs, of which Arabidopsis has more than 1,500[19]. Given this diversity, individual regulatory motifs typically offer limited predictive power of absolute gene expression. We found nevertheless a strong positive relationship between the occurrence of the GATC motif within 500 bp downstream of the TSS and gene expression (Fig. 4A). A comprehensive analysis of all 6-mer counts, both downstream and upstream of the TSS, highlighted that the GATC motif's association with gene expression is unique (Fig. 4B), with few other 6-mers coming even close. This positions the downstream GATC motif as a singularly potent regulatory sequence.

As the GATC-motif effect is strong enough to be observed in genome-wide measurements, we can investigate its function using the many resources available for Arabidopsis. If the GATC motif indeed works primarily through transcription and not mRNA stability, we expect it to affect chromatin measurements. Indeed, occurrence of GATC-motifs is correlated with the active marks H3K4me3 and H3K36me3[20], as well as RNA polymerase II occupancy[21] (Fig. 4C, S18A-B). Moreover, we observed a correlation with genome-wide measurements of mRNA synthesis but not mRNA half-life[22] (Fig. S18C-F). These results further support an effect through transcription.

Because our MPRA inferences came from enhancer activity in leaf cells, we wanted to know whether the GATC motif was also effective in other tissues. Analyzing a compendium of gene expression in different tissues and developmental stages verified once more its regulatory activity in increasing expression, yet also revealed a roughly 3-fold fluctuation in the impact of the GATC motif[23–25] (Fig. 4D). Its influence was smallest in specific seed stages: decreasing from mature green embryo stages through seed drying, then rebounding upon germination[25] (Fig. 4E-F).

Figure 4: Cell-type specificity of the GATC-motif effect on gene expression



(A) In aerial parts of Arabidopsis seedlings[26], gene expression correlates with the number of GATC motifs within 500 bp downstream of the TSS. Expression values are depicted for various motif counts, with "4+" representing 4-9 motifs. A linear fit reveals a GATC-motif effect size of 0.4 (p-value: $5*10^{-128}$), indicating the average expression increase for each added motif. **(B)** For all 6-mers within 500 bp up- or downstream of the TSS, effect size and p-value are determined as in A. The five most significant downstream 6-mers, all containing the GATC sequence, are highlighted. A 5% Bonferroni threshold indicated by a dashed line. **(C)** Average RNA polymerase II occupancy at genes plotted as in **A**, with an effect size of 0.17 (p-value: 10^{-107}). **(D)** GATC-motif effect sizes from a compendium of 200 tissue-specific gene expression data sets[23–25], as determined in **A**. Samples with the lowest effect sizes are shaded and detailed. **(E)** Chart of

GATC-effect size during embryo/seed development and upon imbibition[24,25,27]. **(F)** Expression values in dry seeds and seedling roots[24] plotted as in **A**, with effect sizes of 0.07 (p-value: $2.6*10^{-3}$) and 0.57 (p-value: $4*10^{-400}$), respectively. **(G)** GATC-effect sizes across different root developmental stages, averaged from single-cell expression data[28]. Boxplots in A, C, and F display median, IQR, and 1.5x IQR with outliers as points.

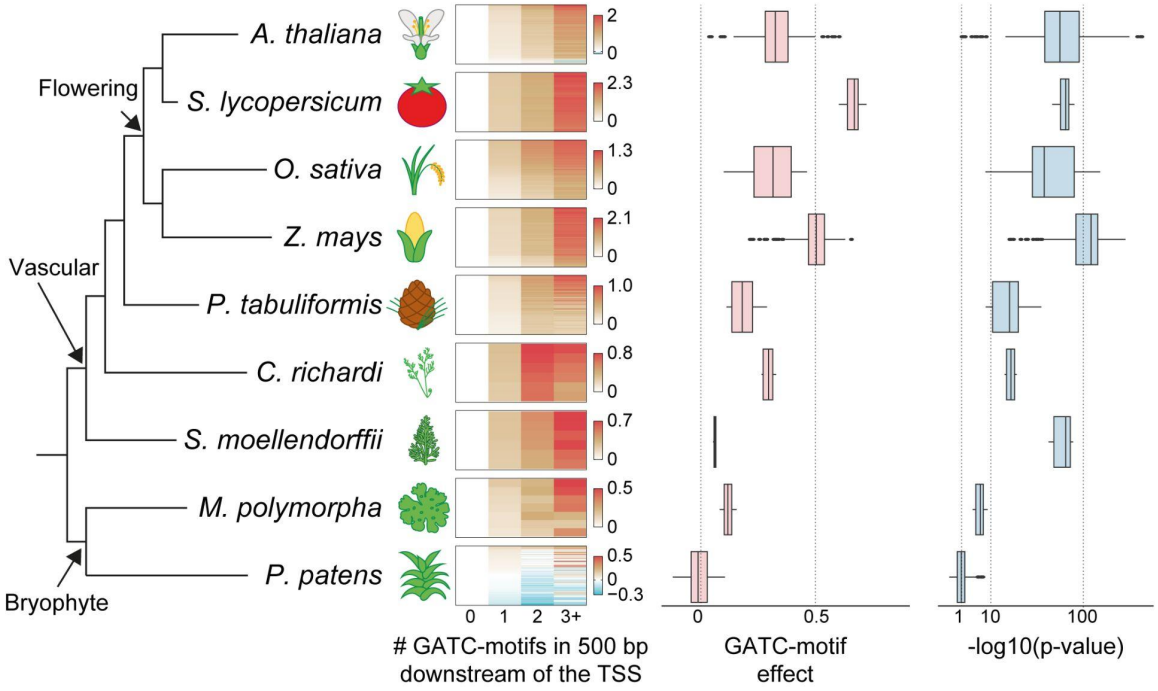
Conversely, the strongest effects were seen in roots (Fig. 4D). Single-cell expression data from Arabidopsis roots[28] pinpointed the meristem as the region most sensitive to GATC-motif stimulation, with decreasing effects through the elongation and maturation zones (Fig. 4G). This trend held true across various root cell types (Fig. S19). Similarly, in the vegetative shoot apex[29], the GATC motif's impact diverged between cell types - for example, mesophyll cells showing muted effects compared to the pronounced effects in epidermal cells (Fig. S20). Overall, the GATC motif's regulatory role spans the entire body plan of the plant, being modified by tissue and cell type. This suggests that the GATC motif functions like a general rheostat, modulating gene expression of thousands of genes across plant cell types.

To evaluate the conservation of the GATC-motif's influence on gene expression, we correlated the number of GATC motifs in the 500 bp downstream region with gene expression across various land plants[24,30–37]. Consistent with our MPRA findings in four flowering plants, GATC-motif count correlated with gene expression in all flowering plants examined (Fig. 5). This conservation extended to the gymnosperm *Pinus tabuliformis* and the fern *Ceratopteris richardii*, albeit with reduced effect-size compared to flowering plants. In the lycophyte *Selaginella moellendorffii*, the association, though significant, was markedly weaker. Among bryophytes, there was a modest effect in *Marchantia polymorpha*, with relatively very weak statistical support, and no clear effect in *Physcomitrium patens*. Overall, the impact of the GATC-motif, and by extension the downstream regulatory sequences, is conserved in vascular plants, with a weaker influence outside flowering plants.

In summary, we have identified the 500 bp region downstream of the TSS as a critical site for transcription regulation of a large fraction of plant genes. We demonstrate that the function of regulatory sequences near the TSS is dependent on their position relative to the TSS, making them distinct from animal enhancers. We further examined a

specific downstream GATC-motif that modulates transcription in a dose-dependent manner. In our analysis, the effect size of the GATC motif easily surpassed that of any other short DNA motif, even those located upstream of the TSS. The motif apparently acts as a regulatory module, operating much like a rheostat in tuning gene expression between cell types, throughout vascular plants.

Figure 5: Downstream GATC motif correlates with gene expression in vascular plants



Heatmap illustrates the average $\log_2$ expression for genes, categorized by 0, 1, 2, or ≥3 GATC-motifs within 500 bp downstream of the TSS, across land plant species. Expression is normalized to the 0-motif group, with color scales specific to each species. The effect on expression (slope) and significance of association (p-value) of the GATC motif, as in Fig. 4, are presented. Boxplots show median, IQR, and 1.5x IQR, with outliers. The right x-axis is square-root scaled.

Our findings are consistent with the frequent identification of TSS-proximal plant introns in driving gene expression[38]. Specifically, research into the role of introns has highlighted a motif similar to the GATC motif[15]. As with introns, the TSS downstream regulatory region, and especially the ability of one motif to control transcription in a

dose-dependent manner, is promising for biotechnology and synthetic biology applications.

Intragenic enhancers could impede RNA polymerase II due to factors recruited to the transcribed region. While the presence of nucleosomes at genes and intronic enhancers in animals indicate that the RNA polymerase can navigate proteins obstructing its path[39], it remains unclear how enhancers might function differently depending on their positioning relative to the TSS. We propose that the distinct 3D genome architecture in Arabidopsis, characterized by densely packed genes[40], might create different local environments on either side of the TSS, but many other scenarios can be imagined as well.

Finally, the GATC-motif regulatory program exerts a widespread influence, modulating the gene expression of a significant proportion of genes throughout the plant body. The adaptive advantages this mechanism offers, and how it has evolved across different lineages, promises to be a fertile ground for future exploration.

Acknowledgment:
We thank J. Neuhold, M. Clavel, V. Nizhynska for technical assistance, A. Levy and D. Ben-Tov for help establishing the protoplast system, Y. Eshed and J. Bindics for sharing seeds, and The Plant Sciences and Next Generation Sequencing facilities at the Vienna BioCenter Core Facilities (VBCF). We also thank F. Berger, L. Dolan, A. Stark, RK. Papareddy, BP. de Almeida, and FK. Lorbeer for fruitful discussions. This work was supported by core funding to MN from the Gregor Mendel Institute, and postdoctoral fellowships to YV from the EU Horizon 2020 via the VIP^2 program and the Marie Skłodowska-Curie individual fellowships (101028014).

Data availability:
Sequencing data have been deposited in the SRA database with accession number PRJNA1009032. Processed data are available in Tables S3 and S4.

References:

1. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).

2. Kumari, S. & Ware, D. Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. *PLoS One* **8**, e79011 (2013).
3. Shin-Han, S. A. D. M.-C. S. A. D. L. Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals1. **139**, 18–26 (2005).
4. Blanc-Mathieu, R., Dumas, R., Turchi, L., Lucas, J. & Parcy, F. Plant-TFClass: a structural classification for plant transcription factors. *Trends Plant Sci.* (2023) doi:10.1016/j.tplants.2023.06.023.
5. Weber, B., Zicola, J., Oka, R. & Stam, M. Plant Enhancers: A Call for Discovery. *Trends Plant Sci.* **21**, 974–987 (2016).
6. Lu, Z. *et al.* The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat Plants* **5**, 1250–1259 (2019).
7. Burgess, D. G., Xu, J. & Freeling, M. Advances in understanding cis regulation of the plant gene with an emphasis on comparative genomics. *Curr. Opin. Plant Biol.* **27**, 141–147 (2015).
8. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166**, 492–505 (2016).
9. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
10. Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
11. Newman, T. C., Ohme-Takagi, M., Taylor, C. B. & Green, P. J. DST sequences, highly conserved among plant SAUR genes, target reporter transcripts for rapid decay in tobacco. *Plant Cell* **5**, 701–714 (1993).
12. Narsai, R. *et al.* Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *Plant Cell* **19**, 3418–3436 (2007).

13. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280–1292 (2016).
14. Jores, T. *et al.* Identification of Plant Enhancers and Their Constituent Elements by STARR-seq in Tobacco Leaves. *Plant Cell* **32**, 2120–2131 (2020).
15. Gallegos, J. E. & Rose, A. B. Intron DNA Sequences Can Be More Important Than the Proximal Promoter in Determining the Site of Transcript Initiation. *Plant Cell* **29**, 843–853 (2017).
16. Norris, S. R., Meyer, S. E. & Callis, J. The intron of Arabidopsis thaliana polyubiquitin genes is conserved in location and is a quantitative determinant of chimeric gene expression. *Plant Mol. Biol.* **21**, 895–906 (1993).
17. Ezer, D. *et al.* The G-Box Transcriptional Regulatory Code in Arabidopsis. *Plant Physiol.* **175**, 628–640 (2017).
18. Schwechheimer, C., Schröder, P. M. & Blaby-Haas, C. E. Plant GATA Factors: Their Biology, Phylogeny, and Phylogenomics. *Annu. Rev. Plant Biol.* **73**, 123–148 (2022).
19. Riechmann, J. L. *et al.* Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105–2110 (2000).
20. Liu, Y. *et al.* PCSD: a plant chromatin state database. *Nucleic Acids Res.* **46**, D1157–D1167 (2018).
21. Lee, T. A. & Bailey-Serres, J. Integrative Analysis from the Epigenome to Translatome Uncovers Patterns of Dominant Nuclear Regulation during Transient Stress. *Plant Cell* **31**, 2573–2595 (2019).
22. Sidaway-Lee, K., Costa, M. J., Rand, D. A., Finkenstadt, B. & Penfield, S. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. *Genome Biol.* **15**, R45 (2014).
23. Toufighi, K., Brady, S. M., Austin, R., Ly, E. & Provart, N. J. The Botany Array Resource:

e-Northerns, Expression Angling, and promoter analyses. *Plant J.* **43**, 153–163 (2005).

24. Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D. & Penin, A. A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.* **88**, 1058–1070 (2016).
25. Hofmann, F., Schon, M. A. & Nodine, M. D. The embryonic transcriptome of Arabidopsis thaliana. *Plant Reprod.* **32**, 77–91 (2019).
26. Schmid, M. *et al.* A gene expression map of Arabidopsis thaliana development. *Nat. Genet.* **37**, 501–506 (2005).
27. Schneider, A. *et al.* Potential targets of VIVIPAROUS1/ABI3-LIKE1 (VAL1) repression in developing Arabidopsis thaliana embryos. *Plant J.* **85**, 305–319 (2016).
28. Shahan, R. *et al.* A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev. Cell* **57**, 543–560.e9 (2022).
29. Zhang, T.-Q., Chen, Y. & Wang, J.-W. A single-cell analysis of the Arabidopsis vegetative shoot apex. *Dev. Cell* **56**, 1056–1074.e8 (2021).
30. Zhang, S. *et al.* Spatiotemporal transcriptome provides insights into early fruit development of tomato (Solanum lycopersicum). *Sci. Rep.* **6**, 23173 (2016).
31. Stelpflug, S. C. *et al.* An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *Plant Genome* **9**, (2016).
32. Xia, L. *et al.* Rice Expression Database (RED): An integrated RNA-Seq-derived gene expression database for rice. *J. Genet. Genomics* **44**, 235–241 (2017).
33. Perroud, P.-F. *et al.* The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J.* **95**, 168–182 (2018).
34. Xiao, Y.-L. & Li, G.-S. Differential expression and co-localization of transcription factors during the indirect de novo shoot organogenesis in the fern Ceratopteris richardii. *Research Square* (2023) doi:10.21203/rs.3.rs-2531906/v1.

35. Niu, S. *et al.* The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217.e14 (2022).
36. Huang, L. & Schiefelbein, J. Conserved Gene Expression Programs in Developing Roots from Diverse Plants. *Plant Cell* **27**, 2119–2132 (2015).
37. Sharma, N., Bhalla, P. L. & Singh, M. B. Transcriptome-wide profiling and expression analysis of transcription factor families in a liverwort, Marchantia polymorpha. *BMC Genomics* **14**, 915 (2013).
38. Rose, A. B. Intron-mediated regulation of gene expression. *Curr. Top. Microbiol. Immunol.* **326**, 277–290 (2008).
39. Zabidi, M. A. & Stark, A. Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. *Trends Genet.* **32**, 801–814 (2016).
40. Liu, C. *et al.* Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution. *Genome Res.* **26**, 1057–1068 (2016).