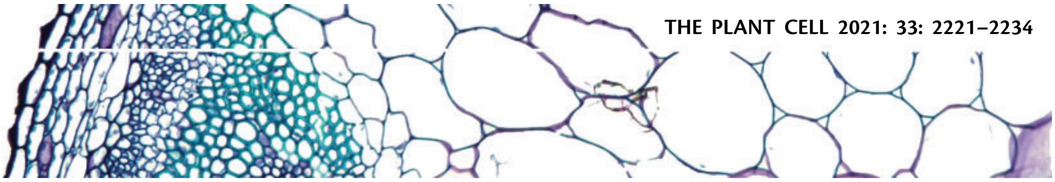
doi:10.1093/plcell/koab107

THE PLANT CELL

THE PLANT CELL 2021: 33: 2221–2234



Research Article

Length variation in short tandem repeats affects gene expression in natural populations of *Arabidopsis thaliana*

William B. Reinar ,[1,2,*] **Vilde O. Lalun**,[1,†] **Trond Reitan** ,[2,†] **Kjetill S. Jakobsen** [2,‡] and **Melinka A. Butenko** [1,*,‡]

1 Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, 0316 Oslo, Norway
2 Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, 0316 Oslo, Norway

*Author for correspondence: w.b.reinar@ibv.uio.no (W.B.R.), m.a.butenko@ibv.uio.no (M.A.B.)
[†]These authors contributed equally to this work (V.O.L. and T.R).
[‡]Senior authors.

K.S.J. and M.A.B. conceived the project. W.B.R., with input from T.R., designed and carried out all computational modeling and statistical work. T.R. performed all work related to validation of models. V.O.L. conducted all experimental work. W.B.R. wrote the manuscript with input from all authors.

The authors responsible for distribution of materials integral to the findings presented in this article is accordance with the policy described in the instructions for Authors (https://academic.oup.com/plcell) are: William B. Reinar (w.b.reinar@ibv.uio.no) and Melinka A. Butenko (m.a.butenko@ibv.uio.no).

Abstract

The genetic basis for the fine-tuned regulation of gene expression is complex and ultimately influences the phenotype and thus the local adaptation of natural populations. Short tandem repeats (STRs) consisting of repetitive DNA motifs have been shown to regulate gene expression. STRs are variable in length within a population and serve as a heritable, but semi-reversible, reservoir of standing genetic variation. For sessile organisms, such as plants, STRs could be of major importance in fine-tuning gene expression as a response to a shifting local environment. Here, we used a transcriptome dataset from natural accessions of *Arabidopsis thaliana* to investigate population-wide gene expression patterns in light of genome-wide STR variation. We empirically modeled gene expression as a response to the STR length within and around the gene and demonstrated that an association between gene expression and STR length variation is unequivocally present in the sampled population. To support our model, we explored the promoter activity in a transcriptional regulator involved in root hair formation and provided experimentally determined causality between coding sequence length variation and promoter activity. Our results support a general link between gene expression variation and STR length variation in *A. thaliana*.

Introduction

The control of gene expression in plant cells depends on the dynamic interplay between regulatory molecules. At the DNA level, gene expression is commonly regulated through modulation of the access for transcription factors (TFs) to chromatin, and on the RNA level by the control of RNA processing and/or degradation. Differential gene expression driving a vast number of biological processes, from embryonic morphogenesis to responses to environmental stimuli, is a result of variations and modifications of TFs, the DNA template, and the RNA produced by transcription. Several studies have addressed how gene regulatory networks composed of genes, noncoding RNAs, proteins, metabolites, and signaling components act in a combinatorial manner to specify developmental programs during plant development (Long et al., 2008). The causal relationship between genome structure and gene regulation is also well

Received February 19, 2021. Accepted April 07, 2021. Advance access publication April 12, 2021

IN A NUTSHELL

Background: Plants such as natural populations of *Arabidopsis thaliana* have adapted to different and changing seasonality and climates. A key adaptive genetic mechanism is to regulate genes—gene expression needs to be tuned in response to the external environment. We know that this is a highly controlled process involving the chromatin structure and promoter binding proteins. At the same time, a puzzling but common feature of genes is the presence of short stretches of repeated DNA (short tandem repeats) near or within the gene sequence. Recent research has shown that mutations in these sites, leading to variation of the repeat length, might affect gene transcription and thus provide a regulatory mechanism for efficient adaptation. The importance of repeat length variation in the regulation of gene expression in natural plant populations is not clear.

Question: We investigated if naturally occurring length variations in short tandem repeats correlate with gene expression differences in wild accessions of *Arabidopsis thaliana*. To answer this question, we analyzed whole genomes and gene expression data from a wide selection of wild samples.

Findings: Our analysis of repeats in the *Arabidopsis thaliana* genome revealed that repetitive sites cluster near genes involved in development, stress responses, and plant hormone pathways. Our statistical analysis demonstrated that the closer a repeat is to a gene, the more likely it is to influence gene expression. A surprising finding was that repetitive sites in protein-coding sequences, which encode repeated amino acids, also affect gene expression. For one protein, we experimentally verified that variation in the repeat altered the binding with its own promoter in *Nicotiana benthamiana* leaf cells and changed its gene expression—and the effect mirrored the signal in the population data.

Next steps: A natural next step is to test if repeat-associated differences in gene transcription lead to increased adaptation under specific conditions. Genetic modification of samples that only differ in a specific repeat and grown under different environmental conditions could reveal if phenotypic adaptability is affected in a controlled setting.

established (Zheng and Xie, 2019). However, the influence of DNA sequence variation, such as short tandem repeats (STRs) consisting of short repetitive stretches of DNA, at the population level has been less explored.

Analysis of RNA sequencing (RNAseq) data from rosette leaves of 998 natural *Arabidopsis thaliana* accessions revealed that natural variation in gene expression levels is linked to geography and climate (Dubin et al., 2015; Kawakatsu et al., 2016). These findings suggest that *A. thaliana* can adapt to its local environment in part by regulating gene expression. Plants are exposed to changes in the environment that could impose a strong selection pressure from one generation to the next. Genetic variants capable of regulating gene expression in response to various fluctuating abiotic and biotic stress factors are likely crucial for individual fitness. In attempts to understand the causal drivers of gene expression variation in *A. thaliana*, both epigenetic (such as methylation) and genetic mechanisms (such as transposable elements, point mutations, and short insertions and deletions) have been explored (Dubin et al., 2015; Kawakatsu et al., 2016). However, little is known about the contribution of STRs to gene expression variation.

STRs are present in genomes throughout the Tree of Life (Srivastava et al., 2019; Tørresen et al., 2019). STRs are defined as repeated units of DNA motifs ranging in size from 1 to 6 bp. Such repeats are highly mutable hotspots, with mutation rates estimated to be in the range of 10^{-3} to 10^{-7} per cell division in human genomes, which are 10-fold to a 10,000-fold higher than the estimated rates of point mutations (Legendre et al., 2007; Gemayel et al., 2010). STR mutations occur primarily in multiples of the unit size due to DNA replication slippage, which either increases or decreases the number of successive units in the STR. Interestingly, in recent large-scale human transcriptome analyses, an evidence for STRs as regulators of gene expression has emerged (Gymrek et al., 2016; Quilez et al., 2016; Fotsing et al., 2019). Furthermore, the enrichment of certain noncoding STRs in promoters and coding STRs in transcriptional regulators suggests a common functional relevance across species (Li et al., 2004; Gemayel et al., 2010; Sawaya et al., 2013; Gymrek, 2017). Experiments on specific STRs in plant genomes have provided evidence of functional length variation in a few STRs, including altered splicing, subcellular localization, protein–protein interactions, and thermoregulation (Press and Queitsch, 2017; Bryan et al., 2018; Press et al., 2018; Jung et al., 2020). A large-scale analysis of gene expression variation and STRs would shed light on the roles of these structures in regulating gene expression.

In this work, we scored genome-wide STR length variation in a subset of sequenced accessions provided by the 1001 Arabidopsis genome project (1001 Genomes Consortium, 2016). Whole genome sequences from 472 natural accessions of *A. thaliana* sampled primarily from Europe and Asia were analyzed by applying a bioinformatic tool previously demonstrated to be applicable for profiling STRs from short read sequencing data (Gymrek et al., 2017; Tang et al., 2017; Willems et al., 2017). We reanalyzed available genome-wide rosette leaf RNAseq data from Kawakatsu et al. (2016) and used statistical modeling to investigate to what extent natural allelic STR variation can explain the gene expression patterns. The natural accessions in the gene expression dataset were grown in growth chambers under identical conditions, ensuring that variation in gene expression was primarily driven by genetic composition. First, we describe the distribution of the STRs included in our models in relation to annotated genes in the *A. thaliana* genome. Next, we describe the statistical model utilizing natural allelic variation in these STRs and empirical gene expression data. Furthermore, we present the results of modeling along with control groups and present candidate genes whose expression is affected by STR length variation. Finally, we experimentally verify the importance of length variation in an STR within the protein coding region of one candidate gene, *ALFIN-LIKE 6* (*AL6*), for the regulation of gene expression. Most importantly, and as stated by our empirical modeling results, the STR length variant in *AL6* was experimentally found to regulate the activity of the *AL6* promoter in vivo, possibly by modulating the chromatin state of the promoter.

Results

Distribution of STRs in the *A. thaliana* genome

To investigate the relationship between STRs and gene expression, we quantified genome-wide natural allelic variation in the global *A. thaliana* population. For this purpose, we employed HipSTR, a polymerase chain reaction (PCR)-stutter aware STR profiler, to examine sequenced genomes from 1,135 *A. thaliana* accessions (1001 Genomes Consortium, 2016; Willems et al., 2017). After quality control and validation of the variant calling results (see "Methods"), the sample set was reduced to 770 high-quality accessions. Of these 770 accessions, 472 overlapped with accessions where rosette leaf RNAseq data from Kawakatsu et al. (2016) were available. Our initial scan for STRs in the *A. thaliana* reference (TAIR10) genome located 37,462 STR sites. However, this set was reduced to 14,195 sites after (1) omitting sites without reliable variant calling data, (2) omitting sites not within 100 kb of genes with available expression level data, and (3) omitting sites without at least two common variants in the natural accessions (see "Methods"). To explore if the distributions of these 14,195 STRs were indicative of a regulatory role in gene expression, we examined how these STRs were distributed in relation to annotated genes in the *A. thaliana* reference genome. In general, the STR sites show distinct clustering upstream of the transcription start site (TSS) (Figure 1A; Supplemental Figure S1). There are, however, some notable differences: Dimer STR motifs (unit size: 2) cluster more closely to the TSS than homopolymer, tetramer, and pentamer STR motifs (unit sizes: 1, 4, and 5). Trimer STR motifs (unit size: 3) are found primarily within genes, as expected due to conservation of the reading frame. Of the trimer STR motifs, we found TTC_n and GAA_n to be the most common (Figure 1B). The most recurrent STR motifs are the homopolymer STRs A_n and T_n, followed by the dimer STRs motifs TA_n, GA_n, and TC_n. Interestingly, STR sites did not display a uniform localization within gene features, except for sites in protein coding sequences (Figure 1C). In promoters, sites were skewed toward the TSS, and in 5′ untranslated regions (UTRs) toward the translation start site. In introns, site localization was skewed toward splice junctions, and in 3′-UTRs toward the translation termination site.

Functional classification of genes with STRs near TSSs

A total of 7,692 genes were found to have common STR sites 500-bp upstream of or in the transcribed region, with at least two common STR variants in the natural population. We tested whether specific Gene Ontology (GO) terms were enriched for these genes compared with what is expected by chance. For this purpose, we performed functional enrichment analysis using plant-specific GO terms from the PANTHER database (Thomas et al., 2003). All enriched and depleted biological process GO terms are available in Supplemental Data Set S1. We found three broad categories that were enriched: Developmental processes, hormone pathways, and responses to stimuli. A high number of genes were linked to more than one of the categories (Figure 1D). More specifically, the developmental processes included the development of flowers, seeds, gametophytes, leaves, the shoot system, the meristem, the xylem tissue, roots, and root hairs. Of genes responsive to stimuli, we enriched for terms related to abscisic acid and auxin hormone pathways and terms, such as the response to water deprivation, osmotic stress, radiation, and biotic interactions (including defense responses). Statistically, our enrichment tests support the notion that the 7,692 genes are not a random subset of *A. thaliana* genes but instead seem to be particularly responsive to environmental stimulus and heavily involved in directing aspects of plant organ development.

Modeling of STR variation and gene expression

Next, we constructed models to test whether natural allelic variation in STRs could explain gene expression pattern variation in the 472 accessions. The geographic distribution of the accessions is shown in Figure 2A. STR variant calling results employed in modeling are available in Supplemental Data Set S2. More specifically, we modeled gene expression (y) as a function of the number of units present in the STR (G). In addition to a term capturing random noise (ε), we included a genetic covariance matrix (X) as a random effect in the model (Figure 2B). As such, we controlled for relatedness between samples through the genetic covariance (Supplemental Data Set S3). Further descriptions and discussion of our model choice can be found in "Methods"; Supplemental Methods. After filtering and validating the STR calls (see "Methods"; Supplemental Figure S2 and Supplemental Data Set S4), we evaluated the significance of the STR for every STR within 100 kb of a gene by comparing models with and without the STR as an explanatory variable. In total, we performed 665,364 tests, 99.6% of which had a sample size more than 100 (Supplemental Figure S3). The mean number of STR sites tested for each gene was 29 (Supplemental Figure S4).

Given the scenario that STR length variation influences the expression of genes in the *A. thaliana* genome, an intuitive expectation is that a STR in close proximity to a gene is

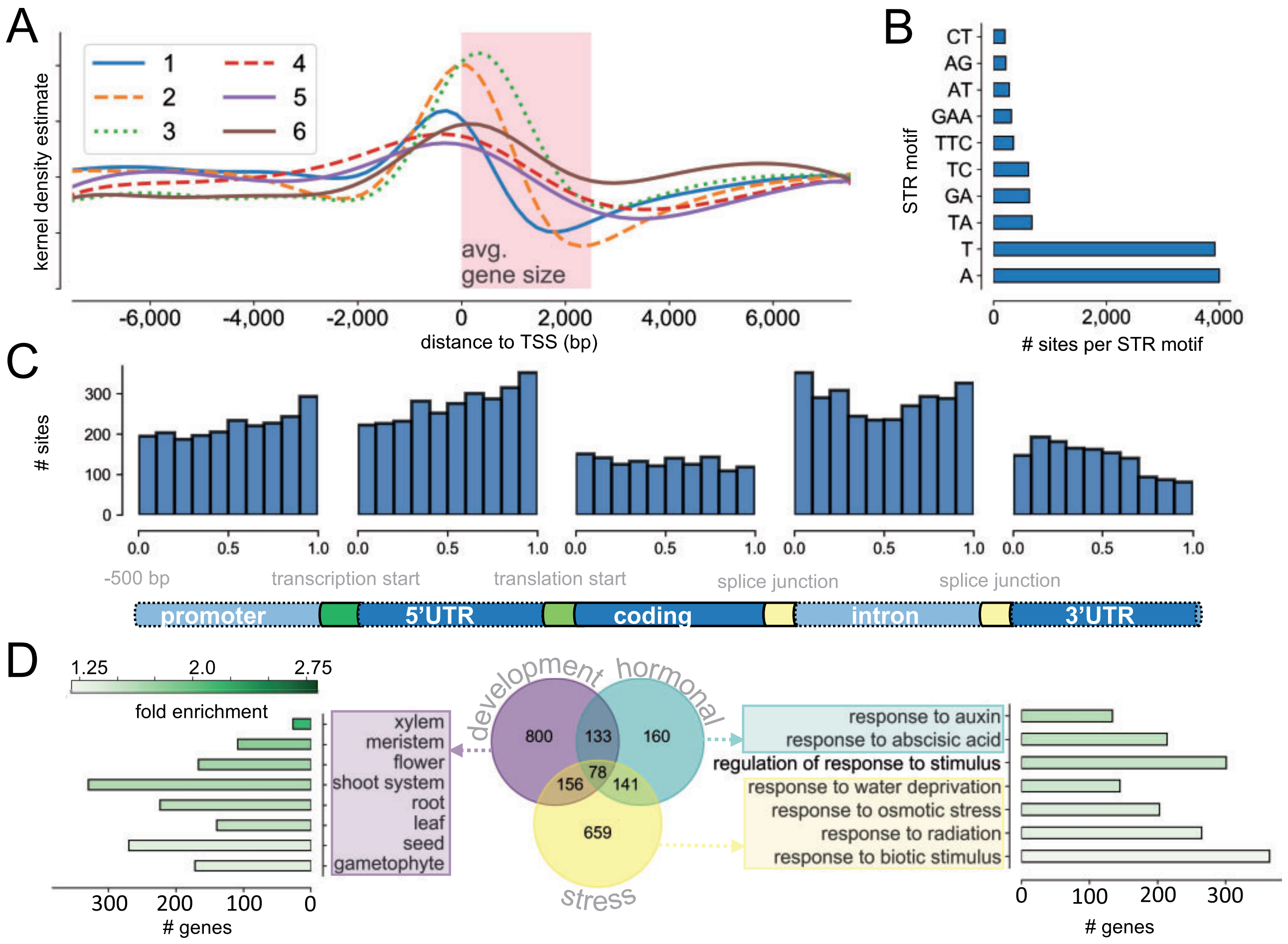


Figure 1 Description of the STRs and the genes included in gene expression modeling. A, The lines show densities of STRs in relation to gene TSSs. Peaks are present upstream and within the gene space. Densities were smoothed using kernel density estimation. Different line colors denote different STR unit sizes (see legend). The average gene size in *A. thaliana* (2,500 bp) is indicated by the pink rectangle. B, Top ten genotyped STR motifs included in gene expression modeling. C, The bar charts show the distribution of STRs in relation to the gene features (linked to the gene cartoon). Here, a value of 0.5 denotes the middle of the feature, read from 5′ to 3′. D, GO enrichment of biological processes linked to genes with STRs in the gene space or up to 500-bp upstream of the TSS. The Venn diagram shows the overlap of genes in GO terms related to development (purple), hormone pathways (blue), and stress (yellow). The bar charts show the number of genes in subcategories related to the three primary categories. The bars are colored by fold enrichment of the GO term, ranging from 1.25 to 2.75 (see colorbar).

more likely to be a significant explanatory variable than a STR far away from a gene. If the STRs in general have no effect, we would not expect any systematic relationship between the significance of STRs and the distance between the gene and the STR. To evaluate the pattern, we plotted the resulting *P*-values as a function of the distance between STR-gene pairs (Figure 2C). From the patterns in Figure 2C, it is evident that our modeling resulted in an effect that increased with proximity, that is, that we find higher significance when the STR is in closer proximity to a gene or within the gene. Effect size shows the same pattern, with higher mean effect sizes when STRs are closer to genes. We repeated the modeling with “mock” STR genotypes, that is, STR genotypes that were shuffled among natural samples. None of these 665,330 tests reached the Bonferroni significance threshold, and there was no sign of increasing effect by proximity to the TSS (Figure 2C; Supplemental Data Set S5). To quantify the percentage of significant STRs as a function of distance, we binned our modeling results in 2,000-bp windows (100,000 bp: 98,000 bp, . . ., 2,000 bp: 0 bp) and calculated the percentage of total tests in each bin that was below the global Bonferroni threshold ($7.5e^{-8}$). The highest percentage of significant tests were found when STRs were located from 0 to +2,000 bp from the gene TSS. Complete results of all STR-gene expression modeling can be found in Supplemental Data Set S6.

To investigate if a similar pattern would emerge when modeling single nucleotide polymorphisms (SNPs), we repeated our modeling with gene expression as a response to SNPs (Supplemental Data Set S7). For the same genes tested in the STR analysis, we tested on average approximately 39 close and common SNPs (a total of 893,372 tests). None of the tests with SNP variation as an explanatory variable produced *P*-values below the Bonferroni multiple testing

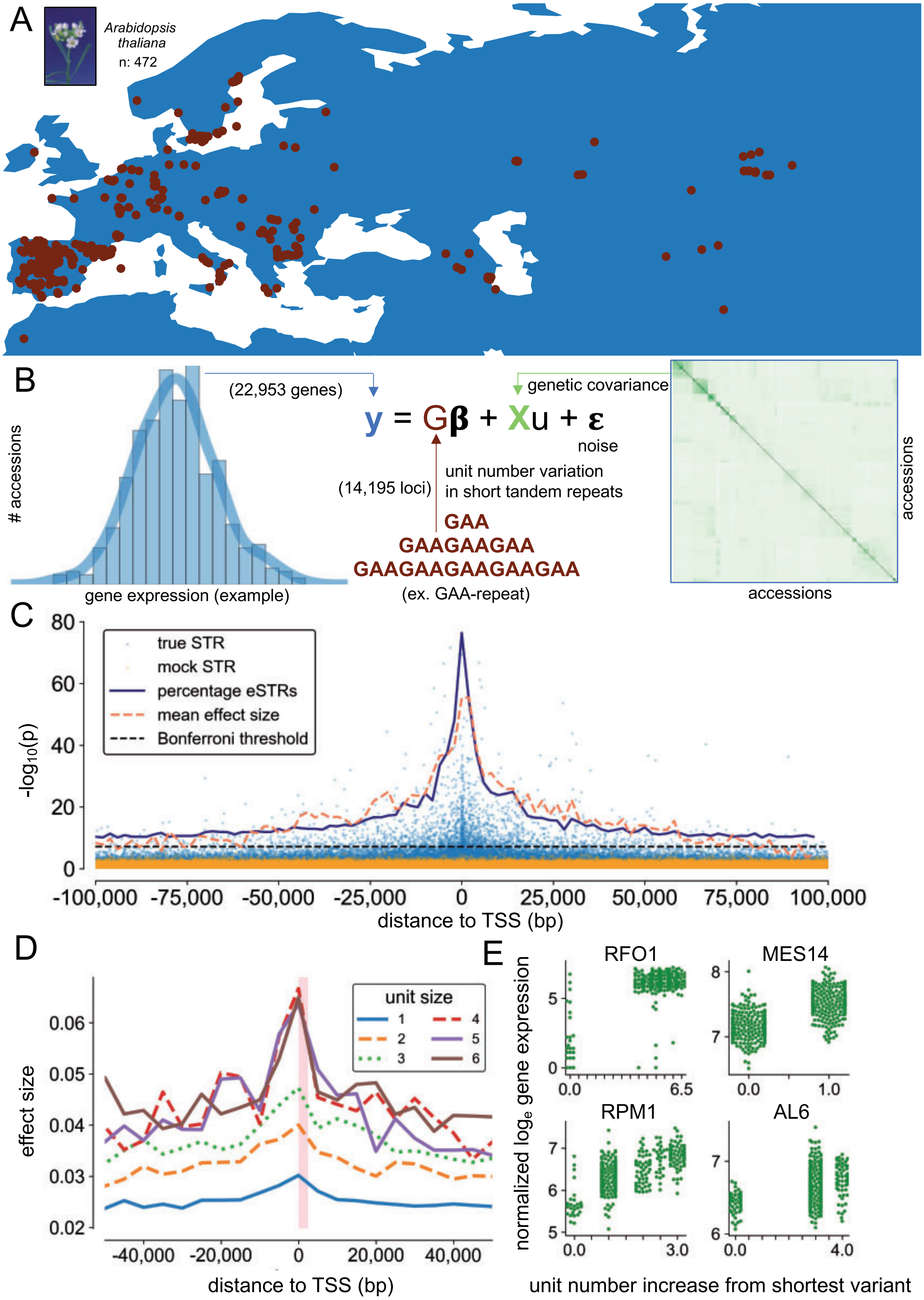


Figure 2 Results of modeling gene expression as a response to STR unit number variation. A, Distribution of *A. thaliana* accessions included in our modeling. B, Description of the model employed to test whether natural allelic variation in STRs could explain gene expression patterns. Quantile-normalized and $\log_e$-transformed gene expression values served as a response (y) in a linear model with unit number variation in STRs (G, with effect size β) as an explanatory variable. Models with and without G were compared using log-likelihood tests. In addition to ε, which captures noise, we also included a genetic covariance matrix based on intergenic, pruned SNPs (X), which captures expected variation in gene expression given the genetic covariance between individuals. See Supplemental Methods for further elaboration on the model and model validation. C, Results of modeling gene expression as a function of the number of units in STRs within 100 kb of the gene. Both the statistical significance

threshold and did not display an increase in effect with increasing proximity to gene TSSs (Supplemental Figure S5). This does not rule out the possibility that SNPs influence gene expression. However, it shows that potential true associations between SNPs and gene expression do not reach a sufficiently high effect size to obtain Bonferroni adjusted significance (adjusted $P < 5.6e^{-8}$) under this particular modeling framework and that large effects on gene expression by point mutations are too rare in comparison with STRs to display an effect that consistently increases with increasing proximity to the TSS. Henceforth, we define STRs with a predicted effect on gene expression as expression STRs (eSTRs). For every eSTR–gene association, we tested if common SNPs close to the eSTR could explain the correlation. Of the 2,306 eSTR–gene tests, 120 (5.2%) eSTRs had a nearby SNP that produced a $P < 0.05$ not corrected for multiple testing (Supplemental Data Set S8). This result indicates that only a small minority of eSTR–gene expression relationships can be attributed to nearby SNPs.

Characteristics of STRs regulating gene expression

Overall, 0.34% of all tests we performed between natural allelic variation in STR length and the expression levels of genes produced a statistically significant result after Bonferroni correction. As we modeled all STR-gene pairs within 100 kb of one another, the low percentage of significant tests was not surprising. However, 1,718 unique eSTRs were involved in a total of 2,306 associations, and 78% of eSTRs were located in annotated genomic features. Interestingly, 935 of these were found in exons, 306 of which were protein coding. In total, 407 were located in introns and 376 were intergenic. And 2.3% of STRs within 2-kb regions upstream of genes were eSTRs, indicating that STRs in promoter sequence are much more likely to affect gene expression than STRs further away from genes. As evident from Figure 2A, the strongest effect was observed when the STR was located within the gene, reaching 4% eSTRs. For STRs located in the gene or its promoter, we investigated whether any specific STR motif was more common to eSTRs than STRs not associated with gene expression. For this purpose, we compared proportions of the DNA motifs in eSTRs and nonsignificant STRs using two-sided Fisher's exact test (Supplemental Data Set S9). $GAA_{(n)}$ and $AAG_{(n)}$ were significantly enriched in the protein-coding regions of eSTRs, which encode amino acid tracts of glutamates (poly-Es) and lysines. Interestingly, we also found $T_{(n)}$ to be enriched in protein-coding regions, as six different $T_{(n)}$ sites had associations with gene expression. On closer inspection, these $T_{(n)}$ sites were found in alternatively spliced transcripts, being located in the protein coding region of one transcript and in the intron of another. In promoters, we found fewer $T_{(n)}$ eSTRs than statistically expected, but Ts were enriched when present in combination with other nucleotides (e.g. $CTTTT_{(n)}$). $CT_{(n)}$-motifs were strongly enriched in introns, and a few motifs were also enriched or depleted in the untranslated regions (5′- and 3′-UTRs). The relative effect sizes of protein coding STRs (predominantly of unit size three) were lower compared with STRs with larger unit sizes but higher compared with homopolymer (unit size: 1) and dimer (unit size: 2) STRs (Figure 2D).

Next, we tested if the location of the STRs within the gene were predictive of being an eSTR, again using two-sided Fisher's exact test. We found that length variations in protein-coding STRs were 1.6 times more likely to affect gene expression than length variation in other STRs ($P = 0.001$), and STRs in 3′-UTRs were 0.6 times less likely to affect gene expression (Supplemental Data Set S9). No deviation from the expected ratios was observed for intronic STRs and STRs in the 5′-UTR. Together, these results indicate that protein-coding STRs, and especially protein-coding STRs with GAA (encoding poly-Es) as the repeated motif, are prime candidates for experimental verification. For a list of named genes with associated eSTRs, see Supplemental Data Set S10. The selected example associations shown in Figure 2E illustrate that the eSTRs we detected have both distal and local effects on the expression of genes involved in a variety of biological processes. These include *RESISTANCE TO P. SYRINGAE PV MACULICOLA 1*, which has an $A_{(n)}$-eSTR 4,644-bp upstream of the gene TSS and is involved in triggering plant resistance in response to a specific plant pathogen (Gopalan et al., 1996); *METHYL ESTERASE 14*, containing a coding $GGA_{(n)}$-eSTR varying only by one unit; *RESISTANCE TO FUSARIUM OXYSPORUM 1*, whose expression correlates with a $T_{(n)}$-eSTR 31,289-bp upstream of the gene TSS and is involved in responses to fungi (Diener and Ausubel, 2005); and *AL6*, encoding a TF known to bind methylated histone H3 and to be involved in regulating various plant responses, including root development upon phosphate (Pi) deficiency and seed germination (Lee et al., 2009; Chandrika et al., 2013; Molitor et al., 2014). The

and the effect size peak when STRs are in close proximity to the TSS. Note that *P*-values are $-\log_{10}$-transformed for clarity. Each blue dot shows the *P*-value resulting from a log-likelihood ratio test between models with and without STRs as an explanatory variable (665,364 tests). Orange dots show the p-values when modeling mock STR genotypes (665,330 tests), none of which reaches the Bonferroni threshold. The *x*-axis shows the distance in base pairs (bp) between the STR and the gene TSS. In 2,000-bp windows, the centered and standardized percentage of tests below the global Bonferroni threshold are shown as a dark blue line, and the centered and standardized mean effect size is shown as an orange dashed line. D, Effect sizes conditioned on unit sizes. The higher the unit size, the larger the effect observed. Average *A. thaliana* gene size is denoted as a pink rectangle. E, Example associations between the number of units in STRs and the expression of genes. The examples illustrate local effects of eSTRs, such as $A_{(n)}$-eSTR 4,644 bp upstream of *RPM1*. They also show that just a single unit increase in a protein coding eSTR can significantly influence expression levels, such as the GGA(n)-eSTR in *METHYL ESTERASE 14* (*MES14*). Distal effects of eSTRs are also present, such as a $T_{(n)}$-eSTR 31,289 bp upstream of the TSS of *RFO1*. Finally, *ALFIN-LIKE 6* (*AL6*) expression levels are influenced by the most overrepresented protein coding eSTR motif, $GAA_{(n)}$. A complete list of named genes influenced by eSTRs is available as Supplemental Data Set S10.

expression of *AL6* is associated with a protein-coding GAA$_{(n)}$-eSTR 1,827 bp downstream of the TSS.

Functional significance of protein-coding STRs in regulating gene expression

Evident from our empirical modeling, eSTRs cluster around the TSS (i.e. in putative promoters and 5′-UTR sequences) and likely exert effects on transcript abundance by altering promoter activity. Of the 11,426 tested STRs located from −2,000-bp upstream of a TSS to the start of the protein-coding sequence, we detected 329 eSTRs (2.8%). However, of the 2,433 tested protein-coding STRs, we detected 113 eSTRs, which yielded 4.4% significant tests. In other words, STRs most likely to produce effects are found in protein-coding sequences. We hypothesized that altered protein function due to length variation in STR-encoded amino acid tracts could potentially cause a feedback on promoter activity. We therefore sought to experimentally test this hypothesis. Given that the GAA$_{(n)}$ motif was the most overrepresented motif in protein-coding eSTRs (Supplemental Data Set S9), we focused our attention on functional investigation of the GAA-eSTR in *AL6*. The GAA-repeat encodes a poly-E tract that is located just upstream of a plant homeodomain (PHD)-type zinc finger domain in AL6 (Figure 3A). Our statistical modeling based on population-wide genome and transcriptome data suggests that the length variation in this repeat influences the expression of *AL6* itself. Accessions carrying a short poly-E tract have lower gene expression levels than accessions with a longer poly-E tract (Figure 2E).

To verify that the activity of the *AL6* promoter (*pAL6*) was significantly altered by the length of the AL6 poly-E tract, we performed fluorescent *β-glucuronidase* (GUS) assays in *Nicotiana benthamiana* leaves. We transiently expressed the GUS enzyme under the control of the *AL6* promoter (*pAL6:GUS*) together with either the AL6 protein from the Col-0 accession with a poly-E tract of seven glutamates (AL6-7E: GAA$_{(7)}$/E$_{(7)}$) coupled to green fluorescent protein (GFP) or AL6 from accession CS77246 with a poly-E tract of three glutamates (AL6-3E: GAA$_{(3)}$/E$_{(3)}$) coupled to GFP in *N. benthamiana* leaves (AL6-E$_n$-GFP). No amino acid changes other than length variation in the poly-E tract were present in the proteins employed in the experiment (Supplemental Data Set S11). An estradiol-inducible 35S promoter was used to drive expression of both AL6-E$_n$-GFP variants to achieve comparable expression levels for both fusion proteins. Both AL6-7E-GFP and AL6-3E-GFP were found to be localized to the nucleus in *N. benthamiana* leaf cells (Figure 3B). Next, we measured the promoter activity of *pAL6* by performing a fluorescent GUS assay in leaves expressing either of the two AL6-E$_n$-GFP variants. We found that the promoter activity of *pAL6* (Supplemental Figure S6) was significantly higher in leaf tissues expressing AL6-7E-GFP compared with leaf tissues expressing AL6-3E-GFP (Figure 3C). The difference in promoter activity between samples was statistically significant (likelihood ratio test: $P =$ 0.006; see "Methods"; Supplemental Data Set S12). This result agrees with the population-wide expression patterns of *AL6* (Figure 2E).

Interestingly, AL6 has been shown to form protein–protein interaction with members of the Polycomb Repressing Complex 1 (PRC1), where the interaction is proposed to be important for silencing of genes during seed germination by regulating the chromatin state (Molitor et al., 2014). Among the members of PRC1, AL6 forms a complex with RING1A (Molitor et al., 2014). We therefore sought to test if length variation in the poly-E tract of AL6 had an impact on its protein–protein interaction with RING1A. The AL6 variants were expressed in fusions with the donor (GFP) fluorophore and RING1A to the acceptor (mCherry) fluorophore at their C-termini driven by an estradiol-inducible 35S promoter in *N. benthamiana* leaves. First, we established that both AL6-3E-GFP and AL6-7E-GFP colocalized to the nucleus with RING1A-mCherry in the same temporal and spatial manner (Figure 3D). Next, we investigated protein–protein interactions by performing Förster Resonance Energy Transfer (FRET)–Acceptor Photobleaching (APB). Protein–protein interaction in each sample was calculated as the GFP fluorescence after photobleaching of mCherry and was represented as the percentage of change in GFP emission (E^{FRET} [%]). We found higher E^{FRET} values in tissue expressing AL6-3E-GFP and RING1A-mCherry (6.5 ± 1.5, $n = 40$) compared with tissue expressing AL6-7E-GFP and RING1A-mCherry (5.2 ± 1.4, $n = 44$) (Figure 3E). The difference in E^{FRET} between samples was statistically significant (one-sided analysis of variance (ANOVA) *P*-value $=$ 0.0002, see "Methods"; Supplemental Data Set S12). Together, these experiments provide in vivo evidence for the role of repeat length variation in the regulation of gene expression as well as protein–protein interactions.

Discussion

In studies of human genomes and transcriptomes, a role for STRs as influencers of gene expression has been established (Gymrek et al., 2016; Quilez et al., 2016; Fotsing et al., 2019). In contrast, little is known about STR length variation as a driver of the regulation of gene expression in plants. Here, by performing genomic and transcriptomic statistical analysis of genome-wide, population-scale data, as well in vivo experiments, we demonstrated that natural allelic variation in STRs can regulate gene expression in *A. thaliana*.

It is important to note that since the high number of tests we performed required a proportionally strict significance threshold, the number of statistically significant associations we detected between gene expression variation and length variation in STRs represents a conservative estimate. In our modeling, we fitted STR variants as continuous variables in linear models, which are restricted to detecting positive or negative linear effects. Possible nonlinear effects could be tested if STR variants were treated as categorical variables. In total, 28% of tests we performed included only two variants, and under such scenarios, variants were treated as

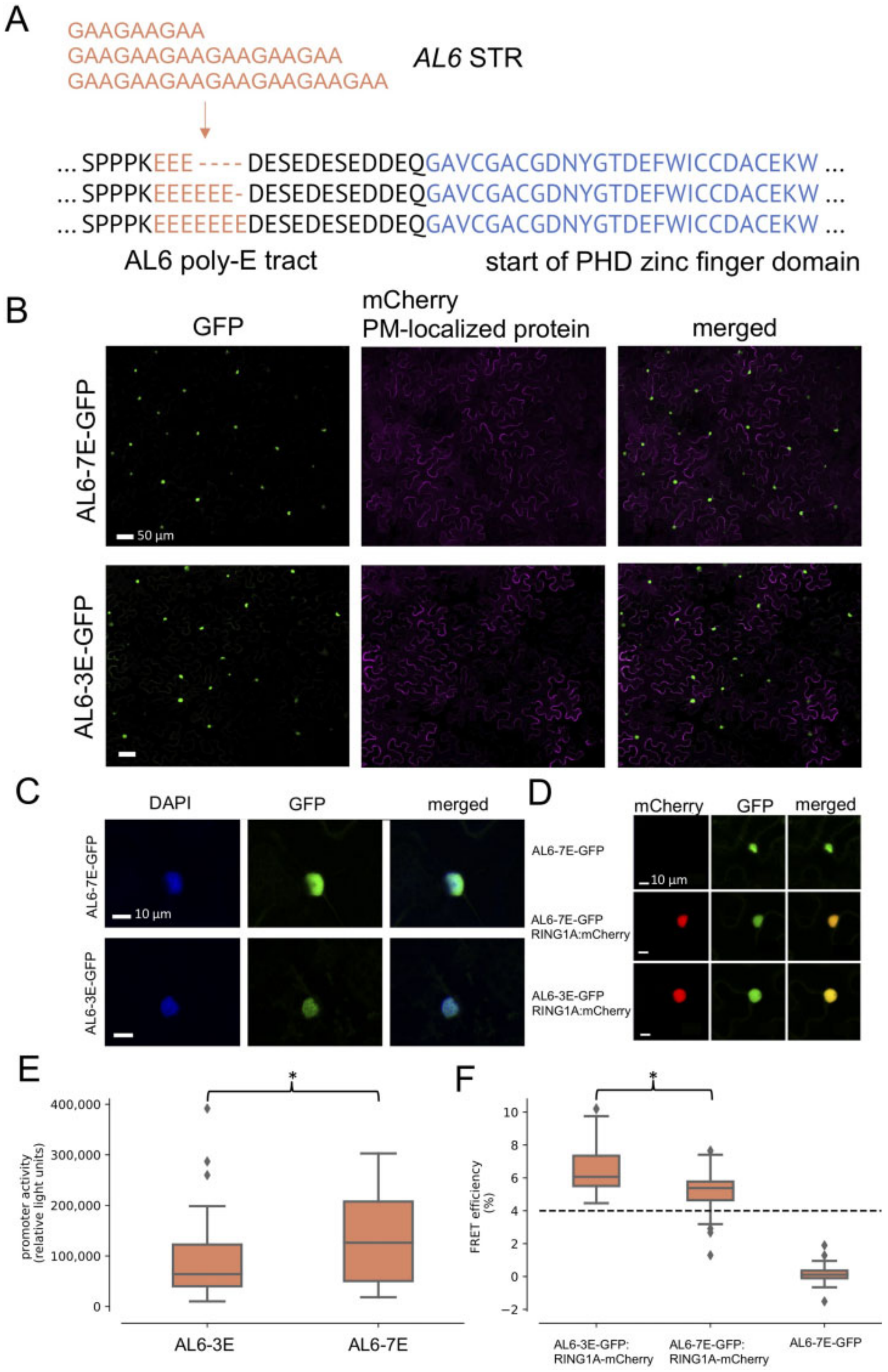


Figure 3 Functional relevance of the GAA-encoded poly-E tract in ALFIN-LIKE 6. A, The nucleotide and amino acid alignment shows three different natural variants of the AL6 glutamate tract (poly-E) present in the 472 *A. thaliana* accessions. The poly-E tract is located immediately upstream of the start of the PHD zinc finger DNA binding domain. B and C, Transient expression of AL6 from Col-0 (AL6-7E-GFP) and AL6 from accession CS77246 (AL6-3E-GFP) in *N. benthamiana* leaves. B, AL6-7E-GFP and AL6-3E-GFP localize to the nuclei in *N. benthamiana*. A protein known to localize to the plasma membrane (PM) was used to outline the PM of the cells. C, Staining with 1-µM 4′,6-Diamidine-2′-phenylindole dihydrochloride shows that AL6 localizes to cell nuclei. D, Expression of AL6-7E-GFP, AL6-3E-GFP, and RING1A-mCherry fusion proteins prior to FRET analysis. E, Boxplots show the promoter activity of *AL6* measured by a fluorescent GUS assay. There was a significant difference in *AL6* promoter activity in tissue expressing AL6-7E-GFP compared with tissue expressing AL6-3E-GFP. See also Supplemental Figure S2. F, Boxplots showing the results from FRET analysis of protein–protein interaction between AL6 and RING1A. AL6 with three repeated glutamates (AL6-3E) interacts significantly stronger with RING1A than AL6 with seven glutamates(AL6-7E). Expression of AL6-7E-GFP alone served as a negative control. The dashed line indicates the common threshold for significant interaction in FRET experiments (Bleckmann et al., 2010). In (E) and (F), the measurements below the upper whisker and above the lower whisker fall within the interquartile range × 1.5. Measurements above or below the whiskers are indicated with diamond symbols. The asterisks indicate a statistically significant difference between groups.

though they were categorical in the regression. For the remaining tests, we missed potential effects that are symmetric around zero. Another, possible obscuring factor for detecting true associations stems from our correction for relatedness. The genetic variation in *A. thaliana* is known to be closely intertwined with environmental variation (1001 Genomes Consortium, 2016). Although the natural samples were grown in chambers under identical conditions, gene expression differences that have been shaped by genetic adaptation to the natural samples' respective environment of origin were likely artificially minimized when we corrected for the genetic relatedness. This may have decreased our ability to detect associations. Taken together, our scoring of eSTRs should be regarded as a highly conservative estimate, and it suggests that the statistically significant associations we have found represent true eSTRs.

Our analyses showed that in light of STR length variation, fairly similar patterns arise in *A. thaliana* gene expression data as in human gene expression data, where a role for STRs as influencers of gene expression has been suggested and experimentally demonstrated (Gymrek et al., 2016; Quilez et al., 2016; Fotsing et al., 2019). Various molecular mechanisms have been proposed that can explain a causal association between length variation in an eSTR and gene expression, which differ between STRs with different unit sizes. Fotsing et al. (2019) proposed that length variation in homopolymer eSTRs may lead to nucleosome displacement, and length variable dimer eSTRs could alter TF binding affinities or the spacing between TF binding sites. Furthermore, GC-rich trimer, tetramer, pentamer, and hexamer eSTRs were proposed to influence DNA and RNA secondary structure and affect transcription. We note that the effect sizes of eSTRs discovered in our analysis are influenced by the unit size (Figure 2D), supporting the notion that the effects are driven by different mechanisms. However, there are some noteworthy differences between our results and reports from analyses of human transcriptomes. In general, STRs in the *A. thaliana* genome seem to cluster to an even higher extent near the TSS and within coding exons (Figure 1A, Supplemental Figure S1). Furthermore, we detected a comparatively much higher number of protein-coding eSTRs in the *A. thaliana* genome (306 in *A. thaliana* versus 11 in human). Specifically, length variation in $GAA_{(n)}$ motifs seems to be a strong driver of gene expression variation, predominantly present in protein-coding sequences as well as 5′-UTRs (Supplemental Data Set S9). We note that a study of the *A. thaliana* RNA–protein interaction landscape found $GAA_{(n)}$ to be the most common motif bound by mRNA binding proteins (Gosai et al., 2015). This finding suggests that the correlation observed between length variation in $GAA_{(n)}$ motifs and gene expression is caused by alterations on the transcript level, which are related to the interactions between RNA binding proteins and $GAA_{(n)}$ motifs. However, our experimental work shows that length variation in the $GAA_{(n)}$ motif in *AL6* affects the strength of protein–protein interactions, suggesting that length variation in such motifs could also be important at the protein level and could provide a more general mechanism to simultaneously affect protein interaction and gene expression regulation in *A. thaliana*. Our results suggest that such a mechanism, distinct from how STRs in promoter sequences are thought to influence gene expression, is more common to plants, which might provide an additional layer of fine-tuning required in a sessile organism.

As STRs have a high mutation rate, selection could in principle operate rapidly to keep gene expression levels tuned according to biotic, environmental, and climatic changes, providing a mechanism by which plants could adapt to their surroundings over a short time span (Rando and Verstrepen, 2007). It is therefore intriguing that many of the genes whose expression levels are influenced by STRs function in responses to biotic or abiotic factors (Supplemental Data Sets S1 and S10). *RESISTANCE TO P. SYRINGAE PV MACULICOLA 1* (RPM1), a plant immune receptor that activates effector-triggered immunity by recognizing pathogen-released effectors alongside activated RPM1 INTERACTING PROTEIN 4, had an STR length variation localized upstream of the TSS that affected gene expression across the accessions included in this study (Grant et al., 1995; Mackey et al., 2002). The same was the case for *RESISTANCE TO FUSARIUM OXYSPORUM 1*, a receptor kinase that confers resistance to a broad spectrum of *Fusarium* races (Diener and Ausubel, 2005). Genome-wide time-series gene expression analysis of Bay-0 and Sha, two rapid-cycling spring annuals *A. thaliana* ecotypes grown in natural field environments, showed an enrichment for abiotic and biotic stress-inducible genes (Richards et al., 2012). It is therefore possible that the differences in gene expression driven by STR variation that we observe across the accessions play a significant role in adaptation to varying ecological factors, particularly those imposing stress on plants.

Genes whose expression influence the morphology of the plant are also important for local adaptation. In this respect, changes in root system architecture in response to variations in soil composition are imperative. Pi deficiency is a major growth-limiting factor in many natural ecosystems (Chiou and Lin, 2011). Plants have developed morphological and molecular responses to optimize Pi uptake and distribution (Raghothama, 1999; Ticconi and Abel, 2004). These adaptive responses are dependent on changes in the expression of several genes under low Pi conditions. The development of root hairs is altered in response to low Pi availability to increase the absorptive surface of the root, as observed in *A. thaliana*, where root hair density can increase by five-fold under suboptimal Pi concentrations (Ma et al., 2001). *AL6*, a member of the Alfin1-like homeodomain protein family, was identified as one of the TF genes whose expression correlated to a $GAA_{(n)}$-eSTR in the coding region of the gene. Our fluorescent GUS reporter assay (Figure 3E) indicated that differences in the GAA repeat influence the expression of *AL6*. AL6 acts as an upstream regulator of root hair formation during Pi starvation in *A. thaliana*, as mutants defective in this TF have shorter root hairs than wild-type plants under low Pi conditions and aberrant PI concentrations (Chandrika et al., 2013). Differences in expression as a result of eSTR variation in *AL6* in the natural accessions, where shorter repeats lead to lower promoter activity, likely contribute to local nutritional adaptation. AL6 control the transcription of a suite of genes critical for root hair elongation under low Pi conditions (Chandrika et al., 2013). Our results indicate that AL6 regulates its own expression either directly or indirectly. Interestingly, during seed development, AL6 interacts with RING1A, a component of PRC1, via its PHD zinc finger domain (Molitor et al., 2014) (Figure 3A). Association of this complex leads to a switch in the histone methylation pattern, causing a change from an active to a repressive transcriptional state in seed developmental genes during seed germination (Molitor et al., 2014). It is therefore interesting that our FRET measurements (Figure 3F) showed a difference in the protein–protein interactions between AL6 and RING1A depending on the length of the poly-E tract encoded by the $GAA_{(n)}$-eSTR in AL6. The addition of glutamates localized directly upstream of the PHD zinc finger domain (Figure 3A) leads to a weaker association with RING1A, as measured based on FRET efficiency (Figure 3F). A reduced interaction between AL6 and RING1A when the protein contains seven glutamates compared with three can explain the reduction in *AL6* promoter activity in the presence of AL6-3E-GFP compared with AL6-7E-GFP if a stronger protein–protein interaction between AL6 and RING1A leads to an increased repressive transcriptional state of the chromatin. It remains to be seen if *RING1A* has an effect of root hair elongation under low Pi conditions.

The presence of similar patterns among distant species suggests that the link between STRs and the fine-tuning of gene expression, regardless of the specific mechanisms, is common to eukaryotes. Future work should focus on the interaction between gene expression, abiotic and biotic stimuli, and length variation in STRs. Given the biological relevance of the genes with STRs, future experimental studies should focus on elucidating the significance of the expression patterns on biological activities such as plant organ development, hormone pathways, and stress responses.

Materials and methods

STR variant calling, filtering, and validation

We used HipSTR (Willems et al., 2017) to call STRs in all 1,135 sequenced accessions released by the 1001 genome project. Briefly, HipSTR performs genotyping of STRs by analyzing the alignment of sequencing reads to STRs detected in a reference genome (here the accession Col-0). If sufficient overlap of nonrepetitive flanking sequence is present, HipSTR is able to “call” the STR variant. First, we used BWA (mem) to align each accession’s reads to the TAIR10 reference genome (Li and Durbin, 2009). Next, we scanned the TAIR10 reference for STRs using Tandem Repeats Finder (Benson, 1999) and used the framework for building nonhuman HipSTR references as described at https://github.com/HipSTR-Tool/HipSTR-references/. The binary alignment map (BAM) files from BWA and the reference repeats serve as input to HipSTR, which performs PCR-stutter aware calling of variants.

```
./HipSTR
  –bams run1.bam, . . ., run1135.bam
  –fasta genome.fa
  –regions str_regions.bed
  –str-vcf str_calls.vcf.gz
```

We used the “vcf_melt” utility script of PyVCF (https://github.com/jamescasbon/PyVCF) to produce a data frame from the VCF built by HipSTR for further analysis. For each reference and alternative call, we extracted the number of units, combined for both alleles in cases of heterozygosity, within each STR (in-house Python scripts). The unit motif was defined by Tandem Repeats Finder in the initial STR detection step and does not necessarily represent the relevant reading frame (relevant if the STR is protein coding). From these unit counts, we constructed a matrix containing 1,135 rows (accessions) and 37,462 columns (STR sites). We reduced this matrix to 869 georeferenced and high-quality accessions following the rationale from a recent study (Ferrero-Serrano and Assmann, 2019). From this matrix, we first omitted all STR sites with >15% missing calls. From the resulting matrix, we omitted all accessions with >10% missing calls. This resulted in a matrix with 770 accessions. The gene expression dataset (Kawakatsu et al., 2016) contained 728 accessions, and 472 of these accessions were overlapping with the 770 in our variant calling dataset. The resulting variant calling dataset is available as Supplemental Data Set S2. The common STR variants (sites with allele frequencies >0.05) used in our gene expression modeling were compared with overlapping STR sites and accessions genotyped in an independent study (Press et al., 2018). For a vast proportion of sites, the mean squared error between our centered and standardized HipSTR calls and the centered and standardized calls from the study was zero, meaning that the relative lengths of the variants were identical to one another (Supplemental Figure S2 and Supplemental Data Set S4).

Gene expression modeling

To test if gene expression, Y, is a function of the number of units present in an STR, G, we needed to perform regression with Y as response and G as an explanatory variable. Gene expression is generally assumed to show a negative binomial distribution, which is an overdispersed version of the Poisson distribution for count data of independent events (Robinson et al., 2010). When the expected values are high, the negative binomial distribution has a standard deviation that increases in proportion to the expected value, as does the log-normal distribution. Also, for high expected values, a large number of outcomes are possible, making a continuous approximation feasible. Thus, the log-normal distribution can be a good approximate distribution for gene expression data, meaning that after log-transformation, regression models using normally distributed residuals can be utilized. This is important, as we modeled gene expression data from a highly genetically divergent population, and thus needed to control for relatedness and population structure.

Normal distribution-based tools are available for performing such analysis. We used the Python package "limix" (https://github.com/limix/limix), which performs regression on the following model: $y = \beta_0 1 + X\beta + \sigma\varepsilon + s\delta$, where y is the vector over the accessions of $\log_e$-transformed gene expressions, β_0 is the intercept, 1 is the unit vector, β is the set of regression coefficients (which is transposed in the equation), X is the explanation variable matrix (one column running over the accessions for each explanatory variable), σ is the standard deviations of independent noise, ε is a vector of independent standard normally distributed noise, s is a scaling factor for the relatedness covariance, and $\delta\sim(0, \sum)$ is a random effect due to relatedness, having covariance matrix $\sum$. Apart from the last term, this follows the form of a standard linear regression. The last term allows the relatedness to be taken into account by having closely related accessions share the same noise term. The independent noise term is there because gene expressions can vary even between highly related individuals or even the same individual measured at different times. The explanation variable matrix X is for our analysis simply the STR that is being examined, although extra explanation variables (such as SNPs in STR analysis) are also possible. The significance of the STR was evaluated by comparing models with and without the STR as an explanatory variable. The model log-likelihood assesses the "goodness of fit" of each model, and a log-likelihood ratio test can test whether the differences between models are statistically significant. A low *P*-value indicates that a model that treats differences in gene expression as a response to natural allelic variation in the STR is a better model than a model that does not take STR variation into account. As we modeled hundreds of thousands of gene–STR pairs, we had to adjust the *P*-value to maintain a 5% type II error rate. For this purpose, we used the Bonferroni correction, that is, dividing the chosen significance threshold of 0.05 by the number of tests we performed (665,364), resulting in a much stricter significance threshold of $7.5e^{-8}$. In our case, y is $\log_e$-transformed RNAseq data produced by Kawakatsu et al. (2016). We retrieved the RNAseq data from the NCBI GEO Accession GSE80744. To avoid spurious associations due to outliers (with a higher probability of being erroneous) in the STR calls, we kept only common variants (minor allele frequency > 0.05), requiring that every variant included in the model should be present in at least 24 of the 472 accessions. Note that the sample size for each regression (gene–STR pair) thus depends both on the number of common STR variants present in X and the number of nonzero measurements in *y*. The result is that the samples size varies from $n = 16$ to 470. However, 99.6% of tests had a sample size > 100 (Supplemental Figure S3). Although it is possible to model all STRs versus all genes and not to restrict modeling to STR–gene pairs within 100 kb of each other, we argue that the gain in "completeness" is counterbalanced by the time, CPU, and memory requirements needed to analyze and work with such a large dataset, which would require approximately 482,000,000 tests. Further discussion, including simulations to support our choice of model, can be found in the Supplemental Methods.

Mock STR control

For every STR site, accessions were given a random STR variant from the pool of all variants that occurred across the population in that particular site. These mock genotypes were modeled using the exact same approach as described for real STR genotypes.

Modeling SNPs

As mentioned in the main text, we tested, on average ~29 STRs per gene. To treat SNPs similarly, we first drew 300,000 unique SNPs randomly from the 10,709,949 SNPs present in the 1,135 sequenced natural *A. thaliana* samples available at 1001genomes.org. Next, we calculated the allele frequency for these 300,000 SNPs and omitted rare variants (i.e. variants with allele frequency <0.05) as we did prior to STR modeling. Of these 300,000 SNPs, 51,437 were common based on this definition. Next, we retrieved the 100 closest SNPs to the TSS for every gene, prioritizing SNPs upstream of the TSS before choosing downstream SNPs. This choice made the distribution of SNPs in relation to the TSS skewed toward promoter sequences, and more similar to how STRs are distributed, which should facilitate fair comparison. As in our treatment for STRs, we omitted every SNP more than 100 kb in any direction from the TSS. The remaining 46,767 SNPs were modeled as described for STRs, testing all of these SNPs against the expression of genes within 100 kb of the SNP. In addition to this, we separately modeled the closest (common) SNP to STRs that produced a significant association with gene expression (2,306 SNP-gene tests) in the same manner as previous gene expression modeling.

STR distributions

We used the "distplot" function of the Python package "seaborn" with regular options for obtaining and plotting kernel density estimates of the distances between STRs and gene TSSs.

Cloning and transient expression of proteins

Arabidopsis thaliana accession CS77246 was ordered from Nottingham Arabidopsis Stock Center. The DNA sequences encoding RING1A from Col-0, AL6 from Col-0 (seven GAAs), and AL6 from accession CS77246 (three GAAs) were cloned in frame with expression vectors containing an 35S estradiol-inducible promoter and a C-terminal fluorescent molecule of GFP or mCherry (Bleckmann et al., 2010) using the Invitrogen Gateway cloning system. Primers used for cloning of *AL6* and *RING1A* are listed in Supplemental Table

S1. The *AL6* promoter (pAL6) was defined as the region 1,500-bp upstream of the *AL6* start codon and amplified from Col-0 genomic DNA using the primers 5′-TTCACAAACGATGTCGCCGG′-3 and 5′-TTCACAAACGATGTCGCCGG′-3. pAL6 was cloned into the R4pGWB633 vector (Nakamura et al., 2009; Tanaka et al., 2013) containing the GUS gene, creating the pAL6:GUS construct. Plasmids were transformed into *Agrobacterium tumefaciens* C58 and further used for transient expression in *N. benthamiana* leaves following a previously described protocol (Butenko et al., 2014). The clones were verified by sequencing, and no other mutations were present that altered the amino acid sequences (Supplemental Data Set S11).

Förster Resonance Energy Transfer–Acceptor PhotoBleaching

We performed FRET—APB measurements to investigate if AL6 proteins containing different numbers of repeated glutamates would have differences in binding properties to RING1A. FRET—APB was performed as described in Bleckmann et al. (2010) on *N. benthamiana* leaves transiently expressing the proteins of interest. A ZEISS LSM880 Airyscan microscope with a Plan-Apochromat 20×/0.8 WD = 0.55 M27 objective, an optical zoom of 5×, frame size of 256 × 256 pixels, and scan speed of 629 ms per frame was used for all measurements. Frame size, laser-power, and gain were kept constant throughout all measurements. FRET efficiency (E_{FRET}) was measured based on the increase in GFP fluorescence intensity after photobleaching of the acceptor mCherry using the ZEISS FRET measurement option ($E_{FRET} = (GFP_{after} - GFP_{before})/GFP_{after} \times 100$). All measurements were performed 10–15 times, and each experiment was repeated three times. The donor only sample (GFP) was used as a negative control. FRET–APB measurements were performed at the NorMIC Imaging platform.

Fluorescent GUS assay

We performed a fluorescent GUS assay to examine if the promoter activity of *pAL6* was significantly altered by expressing AL6 containing different numbers of repeat units. *Nicotiana benthamiana* leaves coinfiltrated with *pAL6:GUS* and either *AL6-7E-GFP* (7 GAAs) or *AL6-3E-GFP* (3 GAAs) were cut into leaf disks and incubated in an 10-µM estradiol solution overnight to induce gene expression. After induction, the leaf disks were individually transferred to wells in a 96-well plate containing 100-µL reaction mixture (10 mM EDTA [pH 8.0], 0.1% SDS, 50 mM sodium Pi [pH 7.0], 0.1% Triton X-100, and 1 mM 4-MUG [M9130-Sigma]) as described previously (Blázquez, 2007) and incubated at 37°C for 6 h. The reaction was stopped by adding 50 µL of stop reagent (1 M sodium carbonate) to each well. Fluorescence was detected using a Wallac 1420 VICTOR2 microplate luminometer (PerkinElmer) at an excitation wavelength of 365 nm and a filter wavelength of 430 nm. Each experiment was repeated three times.

Statistical analysis of experiments

For both the GUS and FRET experiments, individual measurements ($n = 10$–15) were performed on three different days (Supplemental Data Set S12). As such, we included measurement day as a random factor in linear mixed-effect models, which are linear models that take into account the notion that measurements can be dependent. First, we tested the significance of measurement day in the models. For the GUS experiments, measurement day was a significant explanatory variable and was included in model comparisons with and without repeat length as an additional explanatory variable. For the FRET experiments, measurement day was not a significant explanatory factor and was not included. We tested repeat length as an explanatory variable by comparing it to a model without repeat length (null model) using the Chi-squared log-likelihood-ratio test, as implemented in the “anova” function from the R “lme4” package (Bates et al., 2015). See Supplemental Data Set S12 for the experimental data analyzed and the R commands employed.

Accession numbers

The Illumina sequencing reads of 1,135 accessions are available through the National Center for Biotechnology Information (NCBI) Sequencing Read Archive under accession number SRP056687. The RNAseq data of 728 accessions are available through the NCBI Sequencing Read Archive under accession number GSE80744. Accession numbers of named candidate genes identified in this study are available in Supplemental Data Set S10.

Supplemental data

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

Supplemental Data Sets, additional data materials, as well as Python scripts required for figures and modeling are available at: https://doi.org/10.5061/dryad.fttdz08sg.

Supplemental Figure S1. The distribution of short tandem repeats (STRs) in relation to transcription start sites (TSSs).

Supplemental Figure S2. Validation of short tandem repeat variant calls with an independent study.

Supplemental Figure S3. Sample sizes used in modeling gene expression.

Supplemental Figure S4. Short tandem repeat sites tested per gene.

Supplemental Figure S5. Comparison of short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs).

Supplemental Figure S6. Representative images of measurements shown in Figure 3E.

Supplemental Table S1. Primers used for cloning of *AL6* and *RING1A*.

Supplemental Methods S1. Relatedness, simulations, and histochemical GUS assay.

Supplemental Methods Figure S1. Histogram of log10 gene expression averages.

Supplemental Methods Figure S2. Comparison of negative binomial and linear regression test results on simulated data.

Supplemental Data Set S1. Gene Ontology enrichment of genes with STRs in close proximity.

Supplemental Data Set S2. Diploid STR unit number counts in the sampled population.

Supplemental Data Set S3. Genetic relatedness of samples (covariance matrix).

Supplemental Data Set S4. Comparison of scored STR lengths to an independent study.

Supplemental Data Set S5. Modeling results, mock STR genotypes (665,330 tests).

Supplemental Data Set S6. Modeling results, true STR genotypes (665,364 tests).

Supplemental Data Set S7. Modeling results, SNPs (893,372 tests).

Supplemental Data Set S8. Modeling results, SNPs close to eSTRs (2,306 tests).

Supplemental Data Set S9. Counts used in Fisher's Exact tests.

Supplemental Data Set S10. Named genes affected by eSTRs.

Supplemental Data Set S11. Sequenced *AL6* cDNA from Col-0 and natural sample CS77246.

Supplemental Data Set S12. GUS and FRET experimental data and statistical analysis.

Acknowledgments

We thank the NorMic Imaging platform for the use and technical support. Most computational work was performed on the Abel Supercomputing Cluster (Norwegian Metacenter for High-Performance Computing (NOTUR) and the University of Oslo), operated by the Research Computing Services group at USIT, the University of Oslo IT Department. We thank the 1001 Genomes Consortium for making available the DNA sequencing data and also the Ecker lab for making available transcriptome sequencing data used in this study.

Funding

This work was funded by grants from the Research Council of Norway (RCN Grant 251076) to K.S.J. In addition, V.O.L. and M.A.B. were supported by the Research Council of Norway (RCN Grant 230849).

Conflict of interest statement. None declared.

References

1001 Genomes Consortium (2016). 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell **166**: 481–491

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw **67**: 1–48

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27**: 573–580

Blázquez M (2007) Quantitative GUS activity assay in intact plant tissue. CSH Protocols **2007**: db.prot4688

Bleckmann A, Weidtkamp-Peters S, Seidel CAM, Simon R (2010) Stem cell signaling in Arabidopsis requires CRN to localize CLV2 to the plasma membrane. Plant Physiol **152**: 166–176

Butenko M, Wildhagen M, Albert M, Jehle A, Kalbacher H, Aalen RB, Felix G (2014) Tools and strategies to match peptide-ligand receptor pairs. Plant Cell **26**: 1838–1847

Bryan AC, Zhang J, Guo J, Ranjan P, Singan V, Barry K, Schmutz J, Weighill D, Jacobson D, Jawdy S et al. (2018) A variable polyglutamine repeat affects subcellular localization and regulatory activity of a Populus ANGUSTIFOLIA protein. G3 **8**: 2631–2641

Chandrika NNP, Sundaravelpandian K, Yu S-M, Schmidt W (2013) ALFIN-LIKE 6 is involved in root hair elongation during phosphate deficiency in Arabidopsis. New Phytol **198**: 709–720

Chiou T-J, Lin S-I (2011) Signaling network in sensing phosphate availability in plants. Annu Rev Plant Biol **62**: 185–206

Diener AC, Ausubel FM (2005) RESISTANCE TO FUSARIUM OXYSPORUM 1, a dominant Arabidopsis disease-resistance gene, is not race specific. Genetics **171**: 305–321

Dubin MJ, Zhang P, Meng D, Remigereau M, Osborne EJ, Casale FP, Drewe P, Kahles A, Jean G, Vilhjálmsson B et al. (2015) DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. eLife **4**: e05255

Ferrero-Serrano Á, Assmann SM (2019) Phenotypic and genome--wide association with the local environment of Arabidopsis. Nat Ecol Evol **3**: 274–285

Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M (2019) The impact of short tandem repeat variation on gene expression. Nat Genet **51**: 1652–1659

Gemayel R, Vinces MD, Legendre M, Verstrepen, KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet **44**: 445–477

Gopalan S, Bauer DW, Alfano JR, Loniello AO, He SY, Collmer A (1996) Expression of the Pseudomonas syringae avirulence protein AvrB in plant cells alleviates its dependence on the hypersensitive response and pathogenicity (Hrp) secretion system in eliciting genotype-specific hypersensitive cell death. Plant Cell **8**: 1095–1105

Gosai SJ, Foley SW, Wang D, Silverman IM, Selamoglu N, Nelson ADL, Beilstein MA, Daldal F, Deal RB, Gregory BD (2015) Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus. Mol Cell **57**: 376–388

Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes RW, Dangl JL (1995) Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. Science **269**: 843–846

Gymrek M (2017) A genomic view of short tandem repeats. Curr Opin Genet Dev **44**: 9–16

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet **48**: 22–29

Gymrek M, Willems T, Reich D, Erlich Y (2017) Interpreting short tandem repeat variations in humans using mutational constraint. Nat Genet **49**: 1495–1501

Jung J-H, et al. (2020) A prion-like domain in ELF3 functions as a thermosensor in Arabidopsis. Nature **585**: 256–260

Kawakatsu T,Huang SC,JupeF,SasakiE,Schmitz RJ,Urich MA,Castanon R,Nery JR,Barragan C,He Y et al. (2016)Epigenomic diversity in a global collection of Arabidopsis thaliana accessions. Cell **166**: 492–505 10.1016/j.cell.2016.06.044

Lee WY, Lee D, Chung W-I, Kwon CS (2009) Arabidopsis ING and Alfin1-like protein families localize to the nucleus and bind to H3K4me3/2 via plant homeodomain fingers. Plant J **58**: 511–524

Legendre M, Pochet N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res **17**: 1787–1796

Li Y-C, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol **21**: 991–1007

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760

Long TA, Brady, SM, Benfey PN (2008) Systems approaches to identifying gene regulatory networks in plants. Annu Rev Cell Dev Biol **24**: 81–103

Mackey D, Holt BF 3rd, Wiig A, Dangl JL (2002) RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis. Cell **108**: 743–754

Ma Z, Bielenberg DG, Brown KM, Lynch JP (2001) Regulation of root hair density by phosphorus availability in *Arabidopsis thaliana*. Plant Cell Environ **24**: 459–467

Molitor AM, Bu Z, Yu Y, Shen W-H (2014) Arabidopsis AL PHD-PRC1 complexes promote seed germination through H3K4me3-to-H3K27me3 chromatin state switch in repression of seed developmental genes. PLoS Genetics **10**: e1004091

Nakamura S, Nakao A, Kawamukai M, Kimura T, Ishiguro S, Nakagawa T (2009) Development of Gateway binary vectors, R4L1pGWBs, for promoter analysis in higher plants. Biosci Biotechnol Biochem **73**: 2556–2559

Press MO, McCoy RC, Hall AN, Akey JM, Queitsch C (2018) Massive variation of short tandem repeats with functional consequences across strains of *Arabidopsis thaliana*. Genome Res **28**: 1169–1178

Press MO, Queitsch C (2017) Variability in a short tandem repeat mediates complex epistatic interactions in *Arabidopsis thaliana*. Genetics **205**: 455–464

Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ (2016) Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res **44**: 3750–3762

Raghothama KG (1999) Phosphate acquisition. Annu Rev Plant Physiol Plant Mol Biol **50**: 665–693

Rando OJ, Verstrepen KJ (2007) Timescales of genetic and epigenetic inheritance. Cell **128**: 655–668

Richards CL, Rosas U, Banta J, Bhambhra N, Purugganan MD (2012) Genome-wide patterns of Arabidopsis gene expression in nature. PLoS Genetics **8**: e1002662

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One **8**: e54710

Srivastava S, Avvaru AK, Sowpati DT, Mishra RK (2019) Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics **20**: 153

Tanaka Y, Shibahara K, Nakagawa T (2013) Development of gateway binary vectors R4L1pGWB possessing the bialaphos resistance gene (bar) and the tunicamycin resistance gene as markers for promoter analysis in plants. Biosci Biotechnol Biochem **77**: 1795–1797

Tang H, Kirkness, EF,Lippert, C,Biggs, WH,Fabani, M,Guzman, E,Ramakrishnan, S,Lavrenko, V,Kakaradov, B,Hou, C,et al.. (2017) Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. Am J Hum Genet **101**: 700–715 10.1016/j.ajhg.2017.09.013

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res **13**: 2129–2141

Ticconi CA, Abel S (2004) Short on phosphate: plant surveillance and countermeasures. Trends Plant Sci **9**: 548–555

Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, et al. (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Research **47**: 10994–11006

Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y (2017). Genome-wide profiling of heritable and de novo STR variations. Nat Methods **14**: 590–592

Zheng H, Xie W (2019) The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol **20**: 535–550