Nascent transcription and the associated *cis*-regulatory landscape in rice

Jae Young Choi[1], Adrian E. Platts[2], Aurore Johary[3], Michael D. Purugganan[1,4*], and Zoé Joly-Lopez[3*]

[1] Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA

[2] Department of Horticulture, Michigan State University, East Lansing, MI, USA

[3] Département de Chimie, Université du Québec à Montréal, Montréal, Québec, Canada

[4] Center for Genomics and Systems Biology, NYU Abu Dhabi Research Institute, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

* Corresponding authors email: joly-lopez.zoe@uqam.ca (ZJL), mp132@nyu.edu (MDP)

Coauthor email list: jyc387@nyu.edu (JYC), plattsad@msu.edu (AEP), johary.aurore@courrier.uqam.ca (AJ).

Abstract

Background

Plant genomes encode transcripts that require spatio-temporal regulation for proper cellular function, and a large fraction of the regulators can be found in intergenic regions. In animals, distal intergenic regions described as enhancer regions are actively transcribed as enhancer RNAs (eRNAs); the existence of eRNAs in plants has only been fairly recently documented. In this study, we evaluated with high sensitivity the synthesis of eRNAs that arise at genomic elements both distal and proximal to genes by combining PRO-seq with chromatin accessibility, histone modification, and methylation profiles in rice.

Results

We found that regions defined as transcribed intergenic regions are widespread in the rice genome, and many likely harbor transcribed regulatory elements. In addition to displaying evidence of selective constraint, the presence of these transcribed regulatory elements are correlated with an increase in nearby gene expression. We further identified molecular interactions between genic regions and intergenic transcribed regulatory elements using 3D chromosomal contact data, and found that these interactions were both associated with eQTLs as well as promoting transcription. We also compared the profile of accessible chromatin regions to our identified transcribed regulatory elements, and found less overlap than expected. Finally, we also observed that transcribed intergenic regions that overlapped partially or entirely with repetitive elements had a propensity to be enriched for cytosine methylation, and were likely involved in TE silencing rather than promoting gene transcription.

Conclusion

The characterization of eRNAs in the rice genome reveals that many share features of enhancers and are associated with transcription regulation, which could make them compelling candidate enhancer elements.

Keywords

Gene regulation, *cis*-regulatory elements, functional genomics, enhancers, transcribed enhancers, PRO-seq, *Oryza sativa*, methylation, ATAC-seq, Pore-C, chromatin architecture, cis-eQTL

Background

The spatio-temporal regulation of gene expression is essential for coordinating development and adaptive responses to environmental change. Transcriptional regulatory DNA sequences encode information that leads to the recruitment of transcription factors (TFs) in a DNA sequence-dependent manner, allowing the control of the location and rates of chromatin decompaction, transcription initiation, and finally the release of RNA polymerase II (RNAPII) from pausing to productive elongation[1,2]. RNAPII can be recruited by regulatory elements both proximal and distal from the gene(s) they regulate. In animals, an important feature of many (if not all) active *cis*-regulatory elements (CREs), like enhancers, is the production of nascent transcripts by enhancers[3,4].

Enhancer RNAs (eRNAs) have been generally defined as bidirectional, largely unspliced and unpolyadenylated transcripts originating from putative enhancers, with lengths predicted to be less than 150 nucleotides (although some can be up to 2 kb long) [5,6]. In animals, eRNAs are predominantly localized in the nucleus and chromatin-bound fractions [7,8], and are considered a hallmark of active enhancers and a proxy in predicting the spatio-temporal activity of active CREs. Enhancer transcription is important during gene expression. For example, it may maintain chromatin accessibility to enable the binding of TFs and cofactors [9,10], it stimulates the catalytic activity of chromatin remodelers like histone acetyltransferases, it regulates the occupancy of TFs and coactivators [11,12], and it promotes the pause release of RNA polymerase II to productive elongation [13,14].

The instability of eRNAs, however, makes them difficult to detect with steady-state RNA-sequencing data, and their characterization depends on nascent RNA sequencing technologies that enable the measurement of transient RNA transcription at multiple stages and on a genome-wide scale. These methods include global run-on sequencing (GRO-seq) [15] and precision nuclear run-on sequencing (PRO-seq) [16]. These approaches have also successfully identified a diversity of RNA species, including long non-coding genes, upstream antisense RNAs, and eRNAs [17].

In plants, our understanding of eRNAs is still limited. There have been reports of eRNA transcription in maize and cassava [18], and in rice (*Oryza sativa*) [19], but they appear to be rare in other genomes such as Arabidopsis [20]. In maize and cassava, the intergenic regulatory elements from PRO-seq data appear to be enriched for expression quantitative trait loci (eQTLs) identified in kernels compared to a set of random intergenic regions and showed low levels of conservation, a pattern suggesting that these sequences evolve rapidly [18]. In rice, intergenic regions enriched for eRNA signatures had a marked enrichment for open chromatin, a generally asymmetric enrichment for H3K27ac histone modification, and weak but significant positive correlation with nearby gene expression. Similarly to maize and cassava, these sequences had low interspecies conservation but evidence of greater than two-fold excess of nucleotide sites under weak negative selection within *O. sativa* populations [19]. Evidence of recent selection on these sequences suggests the recent emergence of eRNA-producing intergenic regions within *O. sativa*, consistent with transcribed regulatory elements often being species specific [19,21]. Understanding the relationship between eRNAs and the broader class of transcribed regulatory elements in plants is an important challenge and may in part be aided by methods that associate specific genetic variants with agronomic traits. These often indicate that these variants are located in noncoding sequences and likely functional in gene regulation [22,23].

In plants, enhancer regions described so far appear to display specific characteristics. These include the presence of transcription factor (TF) binding motifs, chromatin accessibility, specific histone modifications, low DNA methylation, and evidence of physical interactions with their target genes that may be transient or stable [24,25]. The benefits of considering several of these signatures in parallel has increased our ability to differentiate enhancers from other types of CREs (e.g. silencers, insulators, TATA-box, etc.), as well as other types of regions like promoters, transcription start sites (TSSs) and coding regions [19,24,26–31]. In addition, coupling these signatures with massively parallel reporter assays have greatly contributed to further functional characterization of CREs [26,32,33]. For example, a large proportion of the DNA in regions with active CREs appears more hypomethylated compared to other intergenic regions [28,34–37]. The presence of histone modifications also signal the presence of either active CREs (e.g. H3K27ac, H3K18ac) [38,39], CREs with paused polymerases (enriched with H3K27me3 and low levels of an active histone mark such as histone acetylation) [26,40,41], or repressed CREs (low chromatin accessibility and and H3K27me3 levels) [26,42,43].

Despite the analysis of the various genomic and epigenomic features associated with plant enhancers, the presence of eRNAs are generally not considered. In this study, we used a combination of PRO-seq, complementary functional genomic datasets (DNA methylation profiles, histone modification profiles, transcriptomic profiles, chromatin accessibility), and 3D genome architecture to explore eRNAs and their genomic contexts in the Asian rice genome (*Oryza sativa*). We found that some intergenic eRNAs can be related to transposable element (TE) silencing but many share features of enhancers that suggest they harbor transcribed regulatory elements and are associated with transcription regulation and *cis*-expression quantitative trait loci (e-QTLs). This work continues efforts to understand important types of noncoding elements and is driven by the need to improve crop species by modulating gene expression to enhance plant system resilience [44].

Results and Discussion

Profile of nascent transcription in the rice genome

The goal of the study was to investigate the association between intergenic transcription and putative regulatory sequences in the context of DNA sequence composition, chromosomal environment, and evolutionary constraint. To examine genome-wide nascent transcription, particularly intergenic transcription including eRNAs, we generated PRO-seq data from rice leaf tissues of the *O. sativa* japonica cultivar Azucena grown under optimal conditions. We combined the newly generated PRO-seq data with previously published PRO-seq data that was also generated from plants grown under the same condition [19]. These PRO-seq datasets were combined and consisted of 17,519,424 reads that were aligned to the Azucena reference genome [45].

Because of the bi-directional nature of eRNAs [46,47] and based on previous nascent transcription studies in animals [e.g. [48–52]], we used bi-directional transcription activity as a first criterion for identifying candidate transcribed regulatory elements in the genome [7,53]. To do so, we used the machine learning algorithm dREG (discriminative Regulatory Element detection) that used mapped reads from our PRO-seq experiments (positive and negative strands) and used a support vector regression to recognize the characteristic pattern of divergent transcription at active transcribed regulatory elements (promoters, enhancers, and insulators) [50,54]. Genomic regions with significant bidirectional transcription (herein referred to as dREG peaks) showed an enrichment of PRO-seq reads in both DNA strands, and highly active regions were assigned a high dREG score (>1.0; Fig. 1a). In total we detected 69,898 dREG peaks across the genome, with dREG scores ranging from 0.33 and 1.56 (Fig. 1a; Additional file 1: Figure S1). The genome-wide median dREG score was 0.591 and the median size of a dREG peak was 390 bps (Additional file 1: Figure S1).

The majority of dREG peaks (~91.9%) were found in intergenic regions (Fig. 1b). We divided the intergenic dREG peaks into two classes depending on their distance to genic sequences: (i) proximal dREG peaks (dREG$_{proximal}$), which are intergenic dREG peaks that are <1 kb to predicted transcription start or end sites, and (ii) distal dREG peaks (dREG$_{distal}$), which are >1 kb away from predicted transcription start or end sites. Within the intergenic regions, we found approximately equal proportions of dREG$_{proximal}$ and dREG$_{distal}$ peaks (Fig. 1b). dREG$_{proximal}$ peaks had significantly higher dREG scores than dREG$_{distal}$ peaks (Mann Whitney U test, p-value <0.0001; Additional file 1: Figure S2).

Many cis-regulatory elements (CREs) that are presumed to be actively engaged have been shown to reside within accessible chromatin regions [26,55]. We examined the accessibility of dREG peak regions by comparing dREG peaks with open chromatin regions identified using Assays for Transposase-Accessible Chromatin with sequencing (ATAC-seq), and considered a conservative set of peaks that were within 100 bps of each other as overlapping peaks. Of the 69,898 dREG peaks across the genome, we found an overlap of 5,674 (8.1%) dREG peaks with the 23,961 detected accessible chromatin regions (herein referred to as ATAC peaks) as identified by ATAC-seq (Fig. 1c). We then examined any differences in overlap with ATAC peaks for $\mathrm{dREG}_{\mathrm{proximal}}$ and $\mathrm{dREG}_{\mathrm{distal}}$ peaks and found that $\mathrm{dREG}_{\mathrm{proximal}}$ peaks had twice as much overlap with ATAC peaks compared to $\mathrm{dREG}_{\mathrm{distal}}$ peaks (Additional file 2: Table S1).

In addition to accessible chromatin regions, histone modifications can provide insight into the regulatory mechanisms of the CREs, whether or not they are contained within these accessible chromatin regions. To determine the chromatin features associated with dREG peaks, we used previously generated genome-wide maps of histone modifications and cytosine methylation [19] and compared their profile at ATAC peaks (Fig. 1d). We find that H3K27ac and H3K18ac histone marks were enriched around the dREG and ATAC peaks, (+/- 1.5kb). However, PRO-seq reads showed the expected enrichment of bi-directional transcription activity for dREG peaks but not for ATAC peaks. Moreover, the repressive H3K27me3 mark [56] was enriched within ATAC peaks but depleted within dREG peak regions (Fig. 1d). In addition, the levels of DNA methylation marks [37] were increased for both dREG and ATAC peak sequences, but the latter had higher density of DNA methylation than dREG peaks. We then compared epigenetic marks between $\mathrm{dREG}_{\mathrm{proximal}}$ and $\mathrm{dREG}_{\mathrm{distal}}$ peaks and whether they overlapped ATAC peaks or not, and found no overall differences in chromatin features (Fig. S3). Taken together, these results suggest there is a range of detected transcribed regulatory elements (represented by different dREG scores) in the rice genome and that overall they show relatively little overlap with accessible chromatin regions, as they display different signals in terms of their epigenetic architecture. Consistant with our previous work [19], the different combination of epigenetic marks and transcriptional signatures may play a more nunanced role in determining the chromatin state and functionality of a genomic region.

dREG peaks enriched for DNA methylation overlap with repetitive elements

DNA methylation is often found in inactive regulatory elements, where their target genes are repressed [57], which is why we did not expect to detect an enrichment of DNA methylation for actively transcribed candidate regulatory regions. For instance, when we compared epigenetic marks for transcriptionally active genes with annotated repetitive elements in the rice genome, the common repressive mark H3K27me3 and DNA methylation were enriched within repeat sequences or genes with low expression (Additional file 1: Figure S4). Methylation in plants is dependent on the RNA-directed DNA methylation (RdDM) pathway, wherein transcribed non-coding RNA molecules direct the addition of DNA methylation to specific DNA sequences that are largely associated with transcription repression [58,59]; thus the PRO-seq data in intergenic regions may, in part, be detecting silencing-related transcription. We therefore investigated whether cytosine methylation was mainly attributed to silencing of repetitive elements or was potentialy an inherent characteristic of plant TREs. We categorized dREG peaks into three repeat classes: (i) “without repeat” class, where the dREG peak region does not contain annotated repeat sequences, (ii) "intermediate repeat" class, where the dREG peak region overlaps with at least 1 bp of a repeat sequence, and (iii) "repeat" class, where the entire dREG peak region overlaps a repeat region. For both $\text{dREG}_{\text{proximal}}$ and $\text{dREG}_{\text{distal}}$ peaks, the majority of the peaks (>70%) were in the intermediate repeat class, which had significantly higher dREG scores than both repeat and without repeat class dREG peaks (Mann Whitney U test p-value < 0.001 and Fig. 2a).

DNA methylation in rice, as in other plants, occurs at three contexts: CpG, CHG, and CHH (where H indicates A, T, or C). We compared these cytosine contexts in $\text{dREG}_{\text{proximal}}$ and $\text{dREG}_{\text{distal}}$ peaks, and found that the $\text{dREG}_{\text{distal}}$ peaks had significantly higher levels of methylation across all three cytosine contexts (Mann Whitney U test p-value < 0.001 and Fig. S5). Also, for both $\text{dREG}_{\text{proximal}}$ and $\text{dREG}_{\text{distal}}$ peaks, regardless of repetitive content, CpG sites had the highest methylation level, while CHH sites had the lowest (Fig. 2b). Since a higher dREG score was associated with higher transcriptional activity (Fig. 1a), we plotted dREG peak score and methylation for both $\text{dREG}_{\text{proximal}}$ and $\text{dREG}_{\text{distal}}$ peaks. We found that overall methylation levels were negatively correlated with dREG scores ($p < 0.01$), although CHH methylation, previously associated with TE silencing [36,60–62], had significantly positive correlation ($p < 0.001$) with dREG scores (Fig. S5).

We then contrasted the methylation profiles of dREG peaks in the repeat and without repeat class, as the former is likely to represent epigenetic marks from chromosomal silencing. When we considered chromatin accessibility, we found that dREG peaks (proximal or distal) in the without repeat class were more accessible (Fig. 2c). Taken together, these results suggest that PRO-seq signals detected by dREG that overlap repetitive elements could be involved in silencing mechanisms (such as RDdM) rather than enhancer activity (candidate active CREs).

Evidence of selection within dREG peaks

One strategy known for finding candidate functional sequences in genomes is to look for evolutionary constraint across species [43,63–65]. While sequence conservation can identify conserved noncoding sequences (CNSs), other functional CREs, such as enhancers, have been shown to show sequence diversification and be more species specific [66,67], and therefore not readily detected by conservation-based methods. To infer functionality of our dREG peaks, we estimated the level of evolutionary constraint occurring within the dREG peak regions using two evolutionary-based approaches: phyloP [68] and fitCons [19,69]. Selection within dREG peaks were compared to selection in coding sequences and neutral regions of the genome (control) (Fig. 3).

The results showed an overall higher evolutionary constraint at $\text{dREG}_{\text{proximal}}$ peaks compared to $\text{dREG}_{\text{distal}}$ peaks. While the peaks had lower levels of purifying selection compared to coding sequences (Mann Whitney U test $p < 0.0001$), both $\text{dREG}_{\text{proximal}}$ and $\text{dREG}_{\text{distal}}$ peaks had higher phyloP values compared to random regions of the genome. The $\text{dREG}_{\text{proximal}}$ peaks had a significant negative correlation (p-value < 0.001) between dREG score and phyloP statistics, particularly in regions with dREG scores above 1.2 (Additional file 1: Figure S6). A possible explaination for this could be that some of the regions with higher dREG scores are detecting methylation-related activity. For instance, we detect a positive correlation between $\text{dREG}_{\text{proximal}}$ scores and CHH methylation, which suggests that not all dREG activity is related to transcription but may also be related to silencing activity (Additional file 1: Figure S5).

For fitCons, $dREG_{proximal}$ peaks had higher fitCons scores (ρ) than $dREG_{distal}$ peaks (Mann Whitney U test $p < 0.001$). For both $dREG_{proximal}$ and $dREG_{distal}$ peaks, we observed a significant positive correlation ($p < 0.001$) between dREG score and fitCons statistics. We noted that $dREG_{distal}$ peaks had fitCons scores that were significantly lower than $dREG_{proximal}$ and random control region (Fig. 3b). Moreover, lower dREG scores had lower fitCons scores, and lower dREG scores have been attributed to regions evolving more neutrally and within active repetitive elements (Additional file 1: Figure S6) [19,69].

To investigate why $dREG_{distal}$ peaks had lower fitCons scores, we looked to determine whether we could detect differences in evolutionary constraints at dREG peaks based on their repetitive element content. When we divided dREG peaks by repeat class we found that dREG peaks, regardless of the amount of repeat content, had significantly higher phyloP scores (Mann Whitney U test $p < 0.05$) than random regions of the genome (Additional file 1: Figure S7). For fitCons scores, compared to random genomic regions, $dREG_{proximal}$ peaks had significantly elevated ρ statistics for all repeat classes (Mann Whitney U test $p < 0.001$) but for $dREG_{distal}$ peaks, there was no significant difference for the without repeat class, while the intermediate repeat and repeat classes had significantly lower fitCons scores (Mann Whitney U test p-value < 0.0001). A similar trend in dREG peaks was previously observed in rice [19] and human populations [70]; in these cases, there was an excess of sites under weak negative selection was observed, suggesting a recent selection since the most recent common ancestor. Taken together, these results suggest dREG peaks display some level of conservation that is stronger in proximal peaks than distal peaks, although this distinction may be due to the increased occurrence of repetitive elements with the latter.

Functionality of dREG$_{\text{proximal}}$ peaks and gene expression

We explored whether candidate transcribed regulatory elements identified by the dREG algorithm could indeed be functional regulatory elements (active CREs). We first focused on dREG$_{\text{proximal}}$ peaks detected in the promoter region and examined whether their activity is associated with expression of nearby genes. The results showed that genes with a dREG$_{\text{proximal}}$ peak in its promoter region had significantly higher gene expression (Mann Whitney U test $p < 0.001$ and Fig. 4a). This higher level of gene expression was observed regardless of the repeat class associated with the peak.

Previous studies using PRO-seq data to define transcribed regulatory elements established a dREG score threshold of >0.8 for human enhancers and < 0.3 for non-functional or random transcriptional activity [54]. In plants, the only dREG threshold that has been explicitely used to characterize these regulatory elements was a dREG score of > 1.0 in the rice genome [19]. Based on this, we examined the relationship between dREG score and functional genomic activity by comparing the relationship between gene expression levels and dREG$_{\text{proximal}}$ peak scores. Results showed that genes with a dREG$_{\text{proximal}}$ peak and high dREG score (>0.8) had significantly higher gene expression than genes with low (<0.8) dREG score (Mann Whitney U test $p < 0.001$ and Fig. 4b). This was consistent even in the presence of repeat elements. We also examined the chromatin profiles surrounding dREG$_{\text{proximal}}$ peaks and discovered that peaks with scores higher than 0.8 had similar chromatin marks (Additional file 1: Figure S8), although with some unexpected increase in DNA methylation levels in all cytosine contexts. The methylation pattern in peaks with higher dREG scores appears different from those with lower scores. dREG peaks with higher scores display methylation in the center of the peak, in contrast to methylation being distributed throughput the lower dREG scored peaks. Methylation was therefore taken into consideration in our subsequent downstream analysis.

Identifying distal transcribed regulatory elements through bidirectional transcription activity

Distal transcribed regulatory elements associated with dREG$_{distal}$ peaks may signal RNA-producing enhancers, producing unstable transcripts in both directions [71]. Indeed, in animals, genomic coordinates of distal transcribed regulatory elements have often overlapped with enhancers that actively produce eRNAs, although in plants there is no evidence that these are indeed associated with enhancer activity.

To further characterize rice dREG$_{distal}$ peaks and identify potential functional roles in the rice genome, we examined long-range chromosomal contact interaction between genes and candidate CREs using the Pore-C sequencing method [72]. Pore-C couples chromatin conformation capture with long-read nanopore sequencing to detect genome-wide multi-way chromatin contacts. It has been shown to be highly effective at sequencing through repeat regions of the genome and is able to detect increased contact intensity with less sequencing reads than conventional Hi-C sequencing [72]**.** Using Pore-C sequencing on rice leaves, we generated 104 million concatamer reads that correspond to 290 million contacts across the rice genome (Additional file 2: Table S2).

Since Pore-C sequencing has not been applied in plants to profile chromatin activity, we first examined the functionality of the 3D genome architecture detected by this method. Using the Pore-C method, we detected 3,261 distinct topologically associated domains (TADs) which are localized chromosomal regions of high physical contact [73]. TADs had a median size of 80 kb and the TAD boundaries were enriched for transcription and active chromatin marks (Fig. 5a and Fig.S9). These results were consistent with previous *Oryza* Hi-C sequencing results [74] indicating that Pore-C sequencing was able to detect functional 3D contacts across the *Oryza* chromosome.

To determine whether dREG$_{\text{distal}}$ peaks could represent distal regulatory elements of one or multiple target genes, we used the Pore-C sequencing data to detect the formation of chromatin loops at 5 kbp resolution. Using the FitHiC2 algorithm [75], we detected 33,779 chromatin loops, where 42.7% (14,417) of those loops involved a gene (where the coding window had to contain more than 100 bp of coding sequences) and a noncoding region (Fig. 5b). These represent candidate gene-regulatory element loop interactions. We then visualized this loop interaction by conducting aggregate peak analysis (APA), which takes the contact map and measures its enrichment with respect to its local neighborhood (signal around detected chromatin loop formations) (Fig. 5c). When we compared the APA plots for all chromatin loops and candidate gene-regulatory element loops, the latter loops had strong enrichment of signals centered at the contact point. Specifically for the candidate gene-regulatory element loops, the central window (*i.e.* the contact point) had ~3.5 fold increased in contacts compared to the lower left windows (*i.e.* background contact levels) (Fig. 5c).

Next, we examined the potential functionality of candidate gene-regulatory element loops. As a first step, we processed a subset of candidate loops as probable gene-regulatory element interactions (see the “Methods” section for details). To do so, we (i) focused on chromatin loops that do not cross TAD boundaries, as transcription related chromatin loops co-localize within TAD domains in plants [76], (ii) removed loops where the noncoding anchor had peaks with CHH methylation, as these are more likely to represent transposable element silencing activity [37], and (iii) removed loops where the gene anchor overlapped multiple genes, as the resolution of our Pore-C data could not differentiate whether the noncoding region regulated one or multiple genes.

Following these filtering steps, we analyzed gene-regulatory element loops where the noncoding anchor had only an ATAC peak, only a dREG peak, or had both an ATAC and dREG peak (Fig. 1c). These loops were then compared to loops where the noncoding anchor had no detected ATAC peak nor dREG peak and represented control gene-noncoding sequence interactions (Additional file 1: Figure S10A). We found that for all post-filtered gene-noncoding sequence loops, and regardless of the annotation within the noncoding anchor, the majority of genes had contact with a single noncoding region (Additional file 1: Figure S10B). Furthermore, the average distance between the gene and the noncoding region was 65 kb to the dREG only peak, 45 kb to the ATAC only peak, 35 kb to both dREG and ATAC peaks, and 50 kb to a noncoding anchor with no annotation (Additional file 1: Figure S10B). We also found that genes that contacted a dREG peak had significantly higher gene expression than those that contacted a noncoding anchor with no annotation (Mann Whitney U test $p < 0.01$ and Fig. 5d). In contrast, genes that contacted an ATAC peak did not show any significant difference in expression. Finally, genes with both dREG and ATAC peaks had significantly elevated gene expression (Mann Whitney U test p-value < 0.001 and Fig. 5d). The median dREG scores for gene-contacting dREG peaks were 0.54 (Additional file 1: Figure S11). Gene ontology enrichment analysis on genes found in gene-non coding loops did not highlight specific pathways or function.

dREG$_{distal}$ peaks are targets of transcriptional regulation

As an orthogonal approach to investigate the gene regulatory functions of dREG peaks, we identified expression quantitative trait loi (eQTLs) across a panel of rice varieties and intersected those eQTLs with dREG peak regions. Using gene expression and SNP data from 216 rice varieties grown in well-watered field conditions [77] we detected 274,480 eQTLs after a 5% Bonferroni threshold. We intersected the dREG$_{distal}$ peaks with the significant eQTLs and found an overlap with 13,036 eQTLs.

To test the significance of the observed overlap, we generated a bootstrap distribution by randomly sampling the potentially non-functional regions of the genome, which were matched for size and total number of dREG$_{distal}$ peaks. Results showed that the observed number of overlap between eQTLs and dREG$_{distal}$ peaks (Fig. 6) was higher than the maximum number of overlaps in the bootstrap distribution across random regions, indicating dREG$_{distal}$ peaks are enriched for eQTLs. We also examined dREG$_{distal}$ peaks that were limited to those forming loops with genes and potentially involved in transcription (identified and analyzed in Fig 5d). Overlap of those filtered dREG$_{distal}$ peaks were also significantly enriched for eQTLs (Additional file 1: Figure S12). We repeated the process for distal ATAC peaks (ATAC$_{distal}$) that are intergenic and more than >1 kb away from predicted transcription start or end sites, and found a significant under representation for eQTLs (Additional file 1: Figure S13). Taken together, these results suggest that dREG$_{distal}$ peaks could be good candidate TREs that could impact gene activity by promoting transcription.

Summary: nascent transcription and the rice genome

The rice genome encodes a large number of transcripts that require spatio-temporal regulation and for which a large of *cis*-regulatory elements remain to be characterized. In this study, we took advantage of complementary functional genomics datasets to characterize with high sensitivity the synthesis of eRNAs that arise at both distal and proximal genomic elements. We considered a broad range of dREG scores (from 0.3 to >1.0) to explore a wider range of transcribed regulatory elements that could have distinct functional roles based on their functional genomic signatures. We find that dREG peaks (proximal and distal) that did not overlap with repetitive elements exhibited greater evolutionarily constraint and had a higher incidence of overlap with ATAC peaks. In addition, dREG$_{proximal}$ peaks with higher dREG scores correlated with increased transcription of nearby genes, and gene expression was higher for proximal peaks that did not overlap with repeat sequences (Figs. 2 and 4).

Studies in plant genomes, and particularly in maize, have relied mainly on accessible chromatin regions (e.g. ATAC peaks) as an indicator of the presence of active distal CREs [26,74]. In study by Lozano et al. [18] using dREG/PRO-seq data in maize, they found that 31% of their identified TREs co-located with a list of distal ATAC peaks characterized by Ricci et al. [26], as well as an overlap of 17% between their list of intergenic regulatory elements and CREs found to form CRE-gene loops. Our results suggests a much lower co-localization between dREG and ATAC peaks, with an overlap of about 3.7%. This could be due in part to differences between species in their 3D chromatin architecture, as has been reported with the growing number of chromosomal contact experiments in plants [78]. Our results suggest that a proportion of $\text{dREG}_{\text{distal}}$ peaks could be functional distal transcribed regulatory elements, as we detected a significant increase in gene expression in gene-noncoding loops that contain a dREG peak compared to gene-noncoding loops with no annotation or even with only an ATAC peak. However, since the loops with both an ATAC and a dREG peak is associated with the highest gene expression levels, we still have to determine which elements are functional, whether it requires the overlap between eRNA (i.e., bidirectional transcription) activity and chromatin accessibility, or whether the presence of eRNAs is sufficient.

One of the key observations in this study is the impact of the presence of repetitive elements overlapping the detected dREG peaks. For dREG peaks that partially or entirely overlapped with annotated repetitive elements, we detected a higher percentage of methylation across all three cytosine contexts. Interestingly, we observed an increase in CHH methylation, which occurs predominantly at TEs, and has been shown to be involved in the prevention of transposon jumping during development in Arabidopsis [79]; this methylation occurs through the plant-specific RdDM pathway that operates via non-coding RNA [80]. The fact that DNA methylation can positively or negatively impact transcriptional activity, for example by modulating binding affinity of TFs, makes it a confounding factor when characterizing candidate TREs. In plants, we argue that methylation profiles should always be considered with PRO-seq datasets to characterize the intergenic transcription signal. We also note that the dREG method uses a training set of mammalian transcribed regulatory elements, which may not be optimal for plant genomes. We cannot rule out that there may have been a level of RdDM transcription contamination in the set of transcribed regulatory elements, but this was taken into account when we filtered out dREG sites presenting methylation.

Beyond the potential presence of TE silencing mechanism associated with dREG peaks, we noted that dREG scores were significantly higher for the intermediate repeat class, but this class may contain different types of elements, as an overlap of as little as 1 bp with a repetitive element places a dREG peak in this category. An interesting perspective could be the presence of an overlap between dREG peaks and *cis*-regulatory elements derived from transposable elements, as a result of an evolutionary process called TE exaptation, or TE co-option [81–83]. In mammals, there are several reported instances of TEs providing CREs, including enhancers and repressive elements, and TEs have contributed an important fraction of TF binding sites across the genome (5–40% [84,85]). While in plants the contribution of TEs for CREs is less clear, future characterization of these overlapping regions could be an interesting avenue to identify potential cases of TE-derived CREs. Overall, this suggests that considerations, such as methylation levels and the potential differences with chromatin accessibility, have to be taken into account when addressing transcribed regulatory elements in plants.

Conclusions

In conclusion, we have characterized eRNA producing regions in the rice genome. We find that some of these share features of enhancers and are associated with transcription regulation, which makes them compelling candidate enhancer elements. While the production of eRNAs may be considered a key characteristic used for identifying enhancers in animal studies, there remains a debate as to whether every enhancer (in animal systems) produces eRNAs, even at low levels that are not detected by current methods [51]. In this study, we used the assumptions that eRNAs act principally in *cis* versus *trans*, due to the relative instability of eRNAs [8,67] and several studies have demonstrated eRNA-dependent transcriptional regulation of mRNAs produced from loci adjacent to the corresponding eRNA-producing regions [12,86,87]. Further characterization of eRNA producing regions in other plant genomes will help us better understand whether this assumption holds true for plants.

Materials and Methods

Plant material

Seeds of *O. sativa* landrace Azucena (IRGC 328; tropical Japonica), provided by the International Rice Research Institute (Los Baños, Philippines), were used for the functional genomic datasets. Seeds were incubated for 5 d at 50° C and germinated in water in the dark for 48 h at 30° C. These were subsequently sown on hydroponic pots suspended in 1× Peters solution and 1.8 mM FeSO4 (pH = 5.1–5.8) (JR Peters). Plants were grown for 15 d in growth chambers (12-h days; 30 °C/20 °C day/night; 300–500 μmol quanta m−2 s−1; relative humidity: 50–70%). Leaf tissue for library construction was collected from 17-d-old, young plants.

RNA-Seq

Total RNA was extracted using RNeasy Plant Mini kits (Qiagen), according to the manufacturer's instructions. RNA quality was determined by BioAnalyzer (Agilent). Contaminating DNA was removed from total RNA samples with Baseline-ZERO DNase (Epicentre), whereas ribosomal RNA was removed using a Ribo-Zero rRNA Depletion Kit (Epicentre). Strand-specific RNA-Seq libraries were synthesized using a Plant Leaf ScriptSeq Complete Kit (Epicentre). Libraries were sequenced for 2 × 100-bp reads on an Illumina HiSeq 2500. Two biological replicates were generated and a third replicate (SRA : SRX7082160; Bioproject : PRJNA586887) generated under the same conditions and used in a previous study [19] was used. The sequencing reads were adapter-trimmed and quality-controlled using BBTools (https://jgi.doe.gov/data-and-tools/bbtools/) bbduk program version 37.66 with option: minlen = 25 qtrim = rl trimq = 10 ktrim = r k = 25 mink = 11 hdist = 1 tpe tbo. Trimmed reads were aligned to the Azucena reference genome [45] (Bioproject PRJNA424001) using hisat2 version 2.2.1 [88] and estimated the read counts for each gene using featureCounts [89]. To normalize the variation existing between different samples, we applied the trimmed mean of M value (TMM) method [90] from the edgeR version 3.18.0 package [91] on each samples' gene expression values. For each gene the expression values were averaged across the three replicates.

DNA methylation

DNA was extracted using DNeasy Plant Mini kits (Qiagen) following the manufacturer's protocol. Extracted DNA was sheared into 350-bp fragments using an S220 Focused-ultrasonicator (Covaris). An Illumina TruSeq DNA Kit (Cat. No. FC-121-3001) was used to construct the library and a Zymo Lightning Kit (Cat. No. D5030) was used to perform the bisulfite treatment. KAPA Uracil Polymerase (Cat. No. KK2623) was used to amplify the library with 12 cycles. One biological replicate was generated and a second replicate, generated under the same conditions and used in a previous study [19] was used (SRA : SRX7082155; Bioproject : PRJNA586887). Libraries were sequenced using Illumina protocols for 2×100-bp reads on an Illumina HiSeq 2500. Raw bisulfite sequencing (BS-seq) reads quality controled using the program trim galore Ver. 0.6.6 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with default parameters. We used bismark version 0.16.3 [92] for mapping the BS-seq reads and deduplicating reads.

Chromatin accessibility

Intact nuclei were isolated using the plant nuclei isolation protocol described by Zhang and Jiang [38]. Nuclei quality was assessed using DAPI staining. Chromatin was fragmented and tagged following the standard ATAC-seq protocol [93]. Libraries were purified using Qiagen MinElute columns before sequencing and were sequenced as paired-end 51-bp reads on an Illumina HiSeq 2500 instrument._Sequencing reads were adapter trimmed and QC controlled using the script bbduk.sh version 38.90 (https://sourceforge.net/projects/bbmap/) with parameters: minlen=16 qtrim=rl trimq=20 ktrim=r k=19 mink=10 hdist=1 tpe tbo. Trimmed sequencing reads were aligned to the Azucena reference genome using Bowtie 2 version 2.4.2 [94] under option very-sensitive and with the parameter -X 1000. The Azucena reference genome included the chloroplast sequence (genbank ID: GU592207.1) to allow chloroplast originating ATAC-seq reads, and were subsequently removed using the script removeChrom.py from the Havard FAS informatics group (https://github.com/jsh58/harvard/blob/master/removeChrom.py). Peak calling were conducted using the program MACS2 version 2.2.7.1 [95] with the parameters: --nomodel -g 379627553 -f BED -q 0.05 --extsize 200 --shift -100 --keep-dup all -B. We used MACS2 to call peaks for each of the three replicate libraries and peaks that overlapped 50% in size between at least two replicates were chosen for downstream analysis. To determine ATAC peaks that overlapped dREG peaks we used bedtools closest function and peaks that were within 100 bp of each other were considered as overlapping peaks.

ChIP-Seq

Leaf tissue (2 g) was fixed in 1% formaldehyde (v/v) for 15 min, after which glycine was added to a final concentration of 125 mM (5 min incubation). Tissues were rinsed three times with de-ionized water before being flash frozen in liquid nitrogen. Chromatin extraction and chromatin shearing were performed using a Universal Plant ChIP-seq kit (Diagenode) following the manufacturer's instructions. Protease inhibitor cocktail (MilliporeSigma) was added to extraction buffer. Samples were sonicated for 4 min on a 30 s on/30 s off cycle using a Bioruptor Pico (Diagenode). Subsequent steps were performed as in the Universal Plant ChIP-seq kit protocol. Immunoprecipitation was done using anti-acetyl-histone H3 (Lys27) (H3K27ac; Cell Signaling Technology; Cat. No. 4353S; lot 1), anti-trimethyl-histone H3 (Lys27) (H3K27me3; MilliporeSigma; Cat. No. 07-449; lot 2919706), anti-trimethyl-histone H3 (Lys4) (H3K4me3; EMD Millipore; Cat. No. 07-473; lot 2746331) and anti-acetyl-histone H3 (Lys18) (H3K18ac; Cell Signaling Technology; Cat. No. 9675S; lot 1). The quality and fragment size of immunoprecipitated

DNA and input samples were measured using agarose gel electrophoresis and TapeStation 2200 (Agilent). Libraries were synthesized using a MicroPlex Library Preparation Kit (v.2; Diagenode). Libraries were sequenced as 2 × 50-bp reads on an Illumina HiSeq 2500 instrument. Two biological replicates were generated and a third replicate, generated under the same conditions and used in a previous study [19] was used (SRA : SRX7082158 H3K4me3; SRX7082157 H3K18Ac; SRX7082156 H3K27Ac; SRX7082153 H3K27me3; Bioproject : PRJNA586887).

Sequencing reads were adapter trimmed and QC controlled using the script bbduk.sh ver. 38.90 (https://sourceforge.net/projects/bbmap/) with parameters: minlen=16 qtrim=rl trimq=20 ktrim=r k=19 mink=10 hdist=1 tpe tbo. Trimmed reads were aligned to the Azucena reference genome [45] (Bioproject PRJNA424001) using Bowtie 2 version 2.4.2. [94] under option very-sensitive and with the parameter -X 1000.

PRO-Seq

Nuclei isolation was as described by Hetzel et al. [20], with some modifications. ~20 g of leaf tissue from 17-d-old plants was collected in 4 °C, placed in ice-cold grinding buffer and homogenized using a Qiagen TissueRuptor. Samples were filtered and pellets were washed twice, followed by homogenization, resuspension in storage buffer (10 mM Tris (pH 8.0), 5 mM MgCl2, 0.1 mM EDTA, 25% (v/v) glycerol and 5 mM DTT) and freezing in liquid nitrogen. Nuclei were stained with DAPI and loaded into a flow cytometer (Becton Dickinson FACSAria II). Around 15 million nuclei were sorted based on the size and strength of the DAPI signal, and subsequently collected in storage buffer. Nuclei were pelleted by centrifugation at 5,000g and 4 °C for 10 min, and resuspended in 100 µl storage buffer.

PRO-Seq was performed as described by Mahat et al. [16], generating strand-specific libraries with reads starting from the 3′ end of the RNA. Amplified libraries were assessed for quality on a TapeStation before sequencing with 1 × 50-bp reads on a HiSeq 2500. One biological replicate was generated and a second replicate, generated under the same conditions and used in a previous study [19] was used (SRA : SRX7082159; Bioproject : PRJNA586887).

The raw reads were then used on the proseq2.0 (https://github.com/Danko-Lab/proseq2.0) pipeline [50] that automatically pre-processes the reads, aligns to the reference genome, and generates output bigWig files for downstream PRO-seq peak calling analysis. To identify peaks of divergent transcription activity we used the bigwig file generated from the previous step as an input for the cloud computing version of the dREG algorithm (https://dreg.dnasequence.org/).

PoreC data generation and computational processing.

We generated PoreC libraries following the protocol of Choi et al. [96] and sequencing library was prepared using the Oxford Nanopore Technologies standard ligation sequencing kit SQK-LSK109. Sequencing was conducted on a GridION X5 and PromethION sequencer and the raw data were base-called by Oxford Nanopore Technologies basecaller Guppy version 4.4.0 (available on https://community.nanoporetech.com/) on high-accuracy mode. The Pore-C data analysis was conducted using the PoreC snakemake workflow developed by Oxford Nanopore Technologies (https://github.com/nanoporetech/Pore-C-Snakemake). Briefly, the pipeline first aligns the nanopore Pore-C chromosome contact sequence reads to the

Azucena genome using bwa-sw version 0.7.17-r1188 [97] with parameters -b 5 -q 2 -r 1 -T 15 -z 10 The alignment BAM file was processed with Pore-C tools (https://github.com/nanoporetech/pore-c) to filter spurious alignments, detect ligation junctions, and assign fragments that originated from the same chromosomal contacts. The workflow also generates cool and hic files that can be used for downstream analysis.

TAD and chromatin loop calling

The PoreC contact matrix generated from the previous analysis was normalized using the KR algorthim [98] with the computational suite HiCExplorer version 3.4 [99]. Using the normalized contact matrix the algorithm topdom [100] was used to call TADs as the method was shown recently to be a highly effiecnt and accurate method for detecting TADs [101]. The topdom analysis was conducted using a 5 kbp resolution contact matrix. PoreC contact matrix was also used to statistically determine the significant chromatin contacts using the program FitHiC2 [75]. Using the genomic distance between windows and their contact probability, FitHiC2 applies a spline fit to model an empirical null distribution and detect chromatin contacts as outliers to this null distribution. FitHiC2 was run with default parameters using the 5kbp resolution contact matrix, while setting the lower bound on the intra-chromosomal distance range (parameter -L) as 10 kbp and upper bound (parameter -U) as 1 Mbp. Candidate chromosome loops were filtered by selecting for window pairs that had a Benjamini-Hochberg procedure based false discovery rate threshold q-value < 0.05. Window pairs that had significant evidence of contact were then classified as whether it was a coding or noncoding window by defining a coding window as those that contained more than 100 bp (i.e. greater than 2% of the window) of coding sequences.

Azucena reference genome repeat sequence and gene annotation

Repetitive sequences in the Azucena genome were annotated using the EDTA program [102]. The Azucena reference genome lacked gene annotation. To annotate the gene we took the gene models from the Nipponbare reference genome, which arguably has the best gene models for rice, and lifted over the gene coordinates using the program liftoff [103].

Evolutionary analysis

To calculate phyloP scores we first generated whole genome alignments of wild rice (*O. nivara, O. rufipogon, O. punctata, O. glaberrima, O. barthii, O. brachyantha, O. glumaepatula, O. meridionalis*, and *Leersia perrieri*) [104]. The wild rice reference genomes were aligned to the repeat masked Azucena reference genome using LASTZ version 1.03.73 [105]. Alignment blocks were chained and filtered using the UCSC Kent utilities suite (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64.v287) to obtain a single chain the highest score to represent a single orthologous region of the reference genome. A final multi-genome alignment was generated using the aligner MULTIZ [106]. Using the multi-genome alignment four-fold degenerate sites were extracted using the phast version 1.3 package [107]. The four-fold degenerate sites were then used to build a phylogenetic tree using raxml version 8.2.12 [108] with the GTR gamma model. The topology obtained from the phylogenetic analysis and the four-fold degenerate site data was used to fit a phylogenetic neutral model with phylofit. Using the neutral model we estimated the per-base conservation score using the phylop program with mode CONACC and method LRT.

The fitCons score were obtained from Joly-Lopez et al. [19]. But because the fitCons score were calculated using the Nipponbare reference genome we converted those scores to Azucena reference genome coordinates, by aligning the Azucena reference genome to Nipponbare reference genome and using the program liftOver from the Kent utilities suite.

eQTL detection

Population whole genome resequencing and gene expression data were obtained from Groen et al. [77]. For genes that had multiple transcript expression profile, we chose the longest transcript to represent the expression level of that gene. We conducted eQTL analysis using the program MatrixeQTL [109]. To account for population structure we used plink [110] to calculate structure using polymorphism data and chose the first 5 principal components as covariates to the eQTL model. Resulting p-values for each SNP were filtered using Bonferroni correction and SNPs with adjusted p-value < 0.05 were considered significant eQTLs.

Gene ontology and motif enrichment

Gene ontology (GO) analysis of genes in gene-non coding loops was performed using BinGO [111] with the full list of GO terms (GO_Full) or using PANTHER [112] with the molecular functions, biological process and cellular component GO lists. Motif enrichment was determined using Homer (v.4.10; http://homer.ucsd.edu/homer/) findMotifs with the options -mset plants -len 6,7,8 enabled, and permuted sets of input sequences were used as controls.

Plotting functional genomic data

Enrichment of functional genomic reads around peaks of interest were plotted using deeptools [113], specifically the program computeMatrix. APA plots were generated using the program coolpup.py [114].

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. All raw sequencing data generated from this study are uploaded on the NCBI Sequence Read Archive. The functional genomic data RNA-seq, ChIP-seq, BS-seq, and PRO-seq that were generated in this study was uploaded under PRJNA586887 specifically with the identifiers XXXXXX-YYYYYYY. For PoreC the FAST5 files are available under accession numbers SRR13985185, SRR13985186, SRR13985187, SRR13985192, SRR13985193 and the FASTQ files are available under accession numbers SRR13985180, SRR13985181, SRR13985182, SRR13985190, SRR13985191.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported primarily by a grant from the Zegar Family Foundation (no. A16-0051-004), as well as some support from the National Science Foundation Plant Genome Research Program (no. IOS-1546218) and NYU Abu Dhabi Research Institute to MDP, and a grant from Natural Sciences and Engineering Research Council of Canada (NSERC) (no. RGPIN-2021-03302) to ZJL.

Contributions. JYC, ZJL, and MDP designed the idea and participated in revision and discussion; ZJL, JYC, and AEP did the data collection; JYC and ZJL executed data analysis; AJ helped in data analysis; ZJL and JYC drafted the manuscript with significant participation of MDP. All author(s) read and approved the final manuscript.

Acknowledgements. We would like to thank David Xiaoguang Dai, Priyesh Rughani, Scott Hickey, Eoghan Harrington, and Sissel Juul from Oxford Nanopore Technologies for their assistance on the PoreC data generation. We thank the New York University Center for Genomics and Systems Biology GenCore Facility for sequencing support.

Corresponding author. Correspondence to Michael D. Puruggangn (mp132@nyu.edu) or Zoé Joly-Lopez (joly-lopez.zoe@uqam.ca).

Additional files

Additional file 1: Supplementary figures

Supplemental Fig 1. Distribution of genome-wide dREG scores.

Supplemental Fig 2. dREG scores for $\mathrm{dREG}_{\mathrm{proximal}}$ and $\mathrm{dREG}_{\mathrm{distal}}$ peaks.

Supplemental Fig 3. Epigenetic marks for $\mathrm{dREG}_{\mathrm{proximal}}$ and $\mathrm{dREG}_{\mathrm{distal}}$ peaks.

Supplemental Fig 4. Epigenetic marks for coding and repetitive sequences in the rice genome.

Supplemental Fig 5. DNA methylation levels for $\mathrm{dREG}_{\mathrm{proximal}}$ and $\mathrm{dREG}_{\mathrm{distal}}$ peaks.

Supplemental Fig 6. Scatter plot for dREG peak regions' score and evolutionary conservation scores.

Supplemental Fig 7. Evolutionary scores for $\mathrm{dREG}_{\mathrm{proximal}}$ and $\mathrm{dREG}_{\mathrm{distal}}$ peaks.

Supplemental Fig 8. Chromatin profiles of $\mathrm{dREG}_{\mathrm{Proximal}}$ peaks that are binned by dREG scores.

Supplemental Fig 9. Epigenetic marks surrounding TAD boundaries.

Supplemental Fig 10. Distribution of loops detected by Pore-C for $\mathrm{dREG}_{\mathrm{distal}}$ peaks.

Supplemental Fig 11. Distribution of dREG scores for the dREG peaks contacting a gene.

Supplemental Fig 12. Enrichment of eQTLs within $\mathrm{dREG}_{\mathrm{distal}}$ peaks identified in Figure 5D.

Supplemental Fig 13. Enrichment of eQTLs within ATAC peaks.

Additional file 2: Supplementary tables

Supplemental Table 1. Total number and proportion of genic, distal, and proximal dREG peaks that overlapped a ATAC peak region.

Supplemental Table 2. Pore-C summary statistics.

References

1. Jonkers I, Lis JT. Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Bio. 2015;16:167–77.

2. Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends Genet. 2015;31:426–33.

3. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biol. 2021;22:108.

4. Erik A, O. DC, Kristoffer V-S, Robin A, Berit L, Finn D, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science [Internet]. 2015;347:1010–4. Available from: https://doi.org/10.1126/science.1259418

5. Field A, Adelman K. Evaluating Enhancer Function and Transcription. Annual Review of Biochemistry [Internet]. 2020;89:213–34. Available from: https://doi.org/10.1146/annurev-biochem-011420-095916

6. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol. 2020;27:521–8.

7. Consortium TF, Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

8. Core LJ, Waterfall JJ, Gilchrist DA, Fargo DC, Kwak H, Adelman K, et al. Defining the Status of RNA Polymerase at Promoters. Cell Reports. 2012;2:1025–35.

9. Mousavi K, Zare H, Dell'Orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, et al. eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. Mol Cell. 2013;51:606–17.

10. Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol Cell. 2013;51:310–25.

11. A. SA, J. AB, Xiong J, Benoit M, M. HN, Eric GY, et al. Transcription factor trapping by RNA in gene regulatory elements. Science [Internet]. 2015;350:978–81. Available from: https://doi.org/10.1126/science.aad3346

12. Rahnamoun H, Lee J, Sun Z, Lu H, Ramsey KM, Komives EA, et al. RNAs interact with BRD4 to promote enhanced chromatin engagement and transcription activation. Nat Struct Mol Biol [Internet]. 2018;25:687–97. Available from: https://doi.org/10.1038/s41594-018-0102-0

13. W. ZZ, Rahul R, M. GJC, M. SD, R. CA, Sunney XX. Spatial organization of RNA polymerase II inside a mammalian cell nucleus revealed by reflected light-sheet superresolution microscopy. Proceedings of the National Academy of Sciences [Internet]. 2014;111:681–6. Available from: https://doi.org/10.1073/pnas.1318496111

14. Schaukowitch K, Joo J-Y, Liu X, Watts JK, Martinez C, Kim T-K. Enhancer RNA Facilitates NELF Release from Immediate Early Genes. Mol Cell. 2014;56:29–42.

15. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science [Internet]. 2013;339:950. Available from: http://science.sciencemag.org/content/339/6122/950.abstract

16. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc [Internet]. 2016;11:1455–76. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27442863 https://www.ncbi.nlm.nih.gov/pmc/PMC5502525/

17. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, et al. A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. Cell. 2011;145:622–34.

18. Lozano R, Booth GT, Omar BY, Li B, Buckler ES, Lis JT, et al. RNA polymerase mapping in plants identifies intergenic regulatory elements enriched in causal variants. Koning"] ["D -J de, editor. G3 Genes Genomes Genetics [Internet]. 2021;11:jkab273. Available from: https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkab273/6364897

19. Joly-Lopez Z, Platts AE, Gulko B, Choi JY, Groen SC, Zhong X, et al. An inferred fitness consequence map of the rice genome. Nat Plants. 2020;6:119–30.

20. Hetzel J, Duttke SH, Benner C, Chory J. Nascent RNA sequencing reveals distinct features in plant transcription. Proc National Acad Sci [Internet]. 2016;113:12316–21. Available from: http://www.pnas.org/content/113/43/12316.abstract

21. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. Cell. 2015;160:554–66.

22. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

23. Zheng XM, Chen J, Pang HB, Liu S, Gao Q, Wang JR, et al. Genome-wide analyses reveal the role of noncoding variation in complex traits during rice domestication. Sci Adv [Internet]. 2019;5:eaax3619. Available from: http://advances.sciencemag.org/content/5/12/eaax3619.abstract

24. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet [Internet]. 2014;15:272–86. Available from: http://www.nature.com/articles/nrg3682

25. Weber B, Zicola J, Oka R, Stam M. Plant Enhancers: A Call for Discovery. Trends Plant Sci [Internet]. 2016;21:974–87. Available from: http://dx.doi.org/10.1016/j.tplants.2016.07.013

26. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, et al. Widespread Long-range Cis-Regulatory Elements in the Maize Genome. Nat Plants. 2019;5:1237–49.

27. Marand AP, Zhang T, Zhu B, Jiang J. Towards genome-wide prediction and characterization of enhancers in plants. Biochimica Et Biophysica Acta Bba - Gene Regul Mech. 2017;1860:131–9.

28. Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, Gent JI, et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol. 2017;18:137.

29. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet [Internet]. 2020;21:71–87. Available from: http://dx.doi.org/10.1038/s41576-019-0173-8

30. Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA, et al. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. 2014;15:R41. Available from: https://doi.org/10.1186/gb-2014-15-3-r41

31. Pang B, Snyder MP. Human cells 101. Nat Genet [Internet]. 2020;52:254–63. Available from: http://www.nichd.nih.gov/publications/pubs/fragileX/sub3.cfm

32. Sun J, He N, Niu L, Huang Y, Shen W, Zhang Y, et al. Global Quantitative Mapping of Enhancers in Rice by STARR-seq. Genom Proteom Bioinform. 2019;17:140–53.

33. Jores T, Tonnies J, Dorrity MW, Cuperus JT, Fields S, Queitsch C. Identification of Plant Enhancers and Their Constituent Elements by STARR-seq in Tobacco Leaves. Plant Cell. 2020;32:2120–31.

34. Zilberman D, Coleman-Derr D, Ballinger T, Henikoff S. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature [Internet]. 2008;456:125–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2877514&tool=pmcentrez&rendertype=abstract

35. Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, et al. The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs . Plant Physiol. 2013;162:116–31.

36. Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res. 2013;23:628–37.

37. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016;17:194.

38. Zhang W, Jiang J. Genome-Wide Mapping of DNase I Hypersensitive Sites in Plants BT - Plant Functional Genomics: Methods and Protocols. Stepanova"] ["Jose M Alonso and Anna N, editor. Methods Mol Biology [Internet]. 2015;1284:71–89. Available from: https://doi.org/10.1007/978-1-4939-2444-8_4

39. Fenley AT, Anandakrishnan R, Kidane YH, Onufriev AV. Modulation of nucleosomal DNA accessibility via charge-altering post-translational modifications in histone core. Epigenet Chromatin. 2018;11:11.

40. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature Methods [Internet]. 2012;9:215. Available from: https://doi.org/10.1038/nmeth.1906 http://10.0.4.14/nmeth.1906 https://www.nature.com/articles/nmeth.1906#supplementary-information

41. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011;470:279–83.

42. Huang D, Petrykowska HM, Miller BF, Elnitski L, Ovcharenko I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Res. 2019;29:657–67.

43. Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants [Internet]. 2019;5:1250–9. Available from: http://dx.doi.org/10.1038/s41477-019-0548-z

44. Wing RA, Purugganan MD, Zhang Q. The rice genome revolution: from an ancient grain to Green Super Rice. Nat Rev Genet [Internet]. 2018;19:505–17. Available from: https://doi.org/10.1038/s41576-018-0024-z

45. Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, et al. A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci Data. 2020;7:113.

46. Santa FD, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. Plos Biol. 2010;8:e1000384.

47. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.

48. Wang J, Zhao Y, Zhou X, Hiebert SW, Liu Q, Shyr Y. Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. Bmc Genomics. 2018;19:1–18.

49. Arenas-Mena C, Miljovska S, Rice EJ, Gurges J, Shashikant T, Wang Z, et al. Identification and prediction of developmental enhancers in sea urchin embryos. Bmc Genomics. 2021;22:751.

50. Chu T, Wang Z, Chou S-P, Danko CG. Discovering Transcriptional Regulatory Elements From Run-On and Sequencing Data Using the Web-Based dREG Gateway. Curr Protoc Bioinform [Internet]. 2019;66:e70. Available from: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.70

51. Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, et al. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Gene Dev. 2018;32:42–57.

52. Vihervaara A, Mahat DB, Guertin MJ, Chu T, Danko CG, Lis JT, et al. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat Commun. 2017;8:255.

53. Melgar MF, Collins FS, Sethupathy P. Discovery of active enhancers through bidirectional expression of short transcripts. Genome Biol. 2011;12:R113.

54. Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements with GRO-seq. Nat Methods. 2015;12:433–8.

55. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet [Internet]. 2019;20:207–20. Available from: http://www.nature.com/articles/s41576-018-0089-8

56. Du Z, Li H, Wei Q, Zhao X, Wang C, Zhu Q, et al. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in Oryza sativa L. Japonica. Mol Plant [Internet]. 2013;6:1463–72. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1674205214602203

57. Lee BH, Rhie SK. Molecular and computational approaches to map regulatory elements in 3D chromatin structure. Epigenetics & Chromatin [Internet]. 2021;14:14. Available from: https://doi.org/10.1186/s13072-021-00390-y

58. Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Bio. 2018;19:489–506.

59. Matzke MA, Kanno T, Matzke AJM. RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol. 2014;66:1–25.

60. Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc National Acad Sci. 2015;112:14728–33.

61. Guo W, Wang D, Lisch D. RNA-directed DNA methylation prevents rapid and heritable reversal of transposon silencing under heat stress in Zea mays. Plos Genet. 2021;17:e1009326.

62. Martin GT, Seymour DK, Gaut BS. CHH Methylation Islands: A Nonconserved Feature of Grass Genomes That Is Positively Associated with Transposable Elements but Negatively Associated with Gene-Body Methylation. Genome Biol Evol. 2021;13:evab144.

63. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.

64. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet [Internet]. 2013;45:891–8. Available from: http://www.nature.com/ng/journal/v45/n8/full/ng.2684.html\nhttp://www.nature.com/ng/journal/v45/n8/pdf/ng.2684.pdf

65. Reynoso MA, Kajala K, Bajic M, West DA, Pauluzzi G, Yao AI, et al. Evolutionary flexibility in flooding response circuitry in angiosperms. Science. 2019;365:1291–5.

66. Carelli FN, Liechti A, Halbert J, Warnefors M, Kaessmann H. Repurposing of promoters and enhancers during mammalian evolution. 2018;9:4066. Available from: https://doi.org/10.1038/s41467-018-06544-z

67. Andersson R, Andersen PR, Valen E, Core LJ, Bornholdt J, Boyd M, et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun [Internet]. 2014;5:5336. Available from: https://doi.org/10.1038/ncomms6336

68. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.

69. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet [Internet]. 2015;47:276–83. Available from: http://www.nature.com/doifinder/10.1038/ng.3196

70. Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, et al. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. Nat Ecol Evol. 2018;2:537–48.

71. Core LJ, Martins AL, Danko CG, Waters C, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46:1311–20.

72. Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, Behr JM, et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol. 2022;1–12.

73. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5:eaaw1668.

74. Liu C, Cheng Y-J, Wang J-W, Weigel D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. Nat Plants. 2017;3:742–8.

75. Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc. 2020;15:991–1012.

76. Deschamps S, Crow JA, Chaidir N, Peterson-Burch B, Kumar S, Lin H, et al. Chromatin loop anchors contain core structural components of the gene expression machinery in maize. Bmc Genomics. 2021;22:23.

77. Groen SC, Ćalić I, Joly-Lopez Z, Platts AE, Choi JY, Natividad M, et al. The strength and pattern of natural selection on gene expression in rice. Nature. 2020;578:572–6.

78. Ouyang W, Xiong D, Li G, Li X. Unraveling the 3D Genome Architecture in Plants: Present and Future. Mol Plant. 2020;13:1676–93.

79. Creasey KM, Zhai J, Borges F, Ex FV, Regulski M, Meyers BC, et al. miRNAs trigger widespread epigenetically-activated siRNAs from transposons in Arabidopsis. Nature. 2014;508:411–5.

80. Erdmann RM, Picard CL. RNA-directed DNA Methylation. Plos Genet. 2020;16:e1009034.

81. Joly-Lopez Z, Bureau TE. Exaptation of transposable element coding sequences. Curr Opin Genet Dev. 2018;49:34–42.

82. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet [Internet]. 2017;18:71–86. Available from: http://www.nature.com/doifinder/10.1038/nrg.2016.139

83. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Genetics [Internet]. 2007;41:331–68. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2167627&tool=pmcentrez&rendertype=abstract

84. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008;18:1752–62.

85. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24:1963–76.

86. Tsai P-F, Dell'Orso S, Rodriguez J, Vivanco KO, Ko K-D, Jiang K, et al. A Muscle-Specific Enhancer RNA Mediates Cohesin Recruitment and Regulates Transcription In trans. Mol Cell. 2018;71:129-141.e8.

87. Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, et al. RNA Exosome-Regulated Long Non-Coding RNA Transcription Controls Super-Enhancer Activity. Cell [Internet]. 2015;161:774–89. Available from: https://doi.org/10.1016/j.cell.2015.04.034

88. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.

89. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

90. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25–R25.

91. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

92. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.

93. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biology. 2015;2015:21.29.1-21.29.9.

94. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods [Internet]. 2012;9:357. Available from: https://doi.org/10.1038/nmeth.1923 http://10.0.4.14/nmeth.1923 https://www.nature.com/articles/nmeth.1923#supplementary-information

95. Zhang Y, Liu T, Meyer C a, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol [Internet]. 2008;9:R137. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2592715&tool=pmcentrez&rendertype=abstract

96. Choi JY, Lye ZN, Groen SC, Dai X, Rughani P, Zaaijer S, et al. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. Genome Biol. 2020;21:21.

97. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: Genomics. 2013;

98. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis [Internet]. 2012;33:1029–47. Available from: https://doi.org/10.1093/imanum/drs019

99. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9:189.

100. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44:e70–e70.

101. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19:217.

102. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:1–18.

103. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37:btaa1016.

104. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet [Internet]. 2018;50:285–96. Available from: https://doi.org/10.1038/s41588-018-0040-0

105. Harris RS. Improved pairwise alignment of genomic DNA. The Pennsylvania State University;

106. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, et al. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. Genome Res. 2004;14:708–15.

107. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011;12:41–51.

108. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

109. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.

110. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genetics. 2007;81:559–75.

111. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. 2005;21:3448–9. Available from: https://doi.org/10.1093/bioinformatics/bti551

112. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. 2019;47:D419–26. Available from: https://doi.org/10.1093/nar/gky1038

113. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44:W160–5.

114. Flyamer IM, Illingworth RS, Bickmore WA. Coolpup.py: versatile pile-up analysis of Hi-C data. Bioinformatics. 2020;36:2980–5.

Figure Legend

Figure 1. Chromosomal features associated with dREG peaks in the rice genome. (**a**) PRO-seq read counts for positive-sense (top) and negative-sense of dREG peaks. (**b**) Classification of dREG peaks according to the genomic regions it was located. Genic: within coding sequence regions; proximal: within 1 kbp of genic sequences; and distal: more than 1 kbp away from genic sequences. (**c**) Overlap between dREG and ATAC peaks. (**d**) Enrichment of functional genomic sequencing reads 1.5 kbp upstream and downstream of dREG (left) and ATAC (right) peak regions.

Figure 2. Repetitive sequence characteristics within proximal (top row) and distal (bottom row) dREG peak regions. (**a**) Median dREG scores for dREG peaks classified into three repeat classes: (Left) Repeat: entire dREG peak region is a repetitive sequence; (Middle) Intermediate repeat: dREG peak regions that are not classified as “repeat” or “no repeat” class; (Right) No repeat: no repeat sequence was annotated in dREG peak region. The numbers in the box plot represent the count of dREG peaks in each category. (**b**) Percentage of methylated cytosine for the three different cytosine contexts CpG, CHG, and CHH sites (where H is A, T, or C nucleotide). (**c**) ATAC-seq read counts centered at dREG peak for the repeat classes No repeat and Repeat.

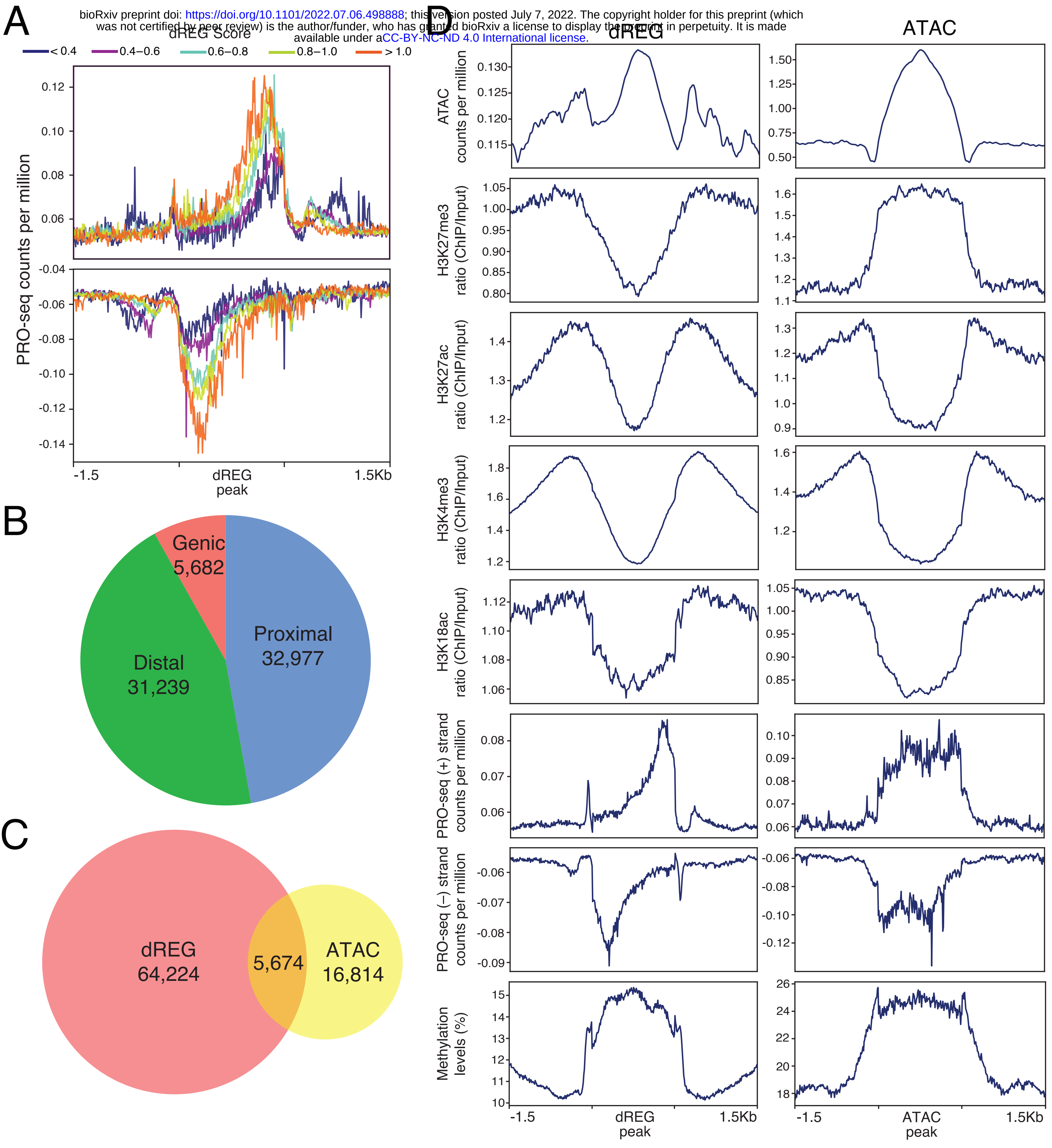
Figure 3. Evolutionary scores. (**a**) phyloP and (**b**) fitcons for dREG peaks and comparison to coding sequence regions or random regions of the genome.

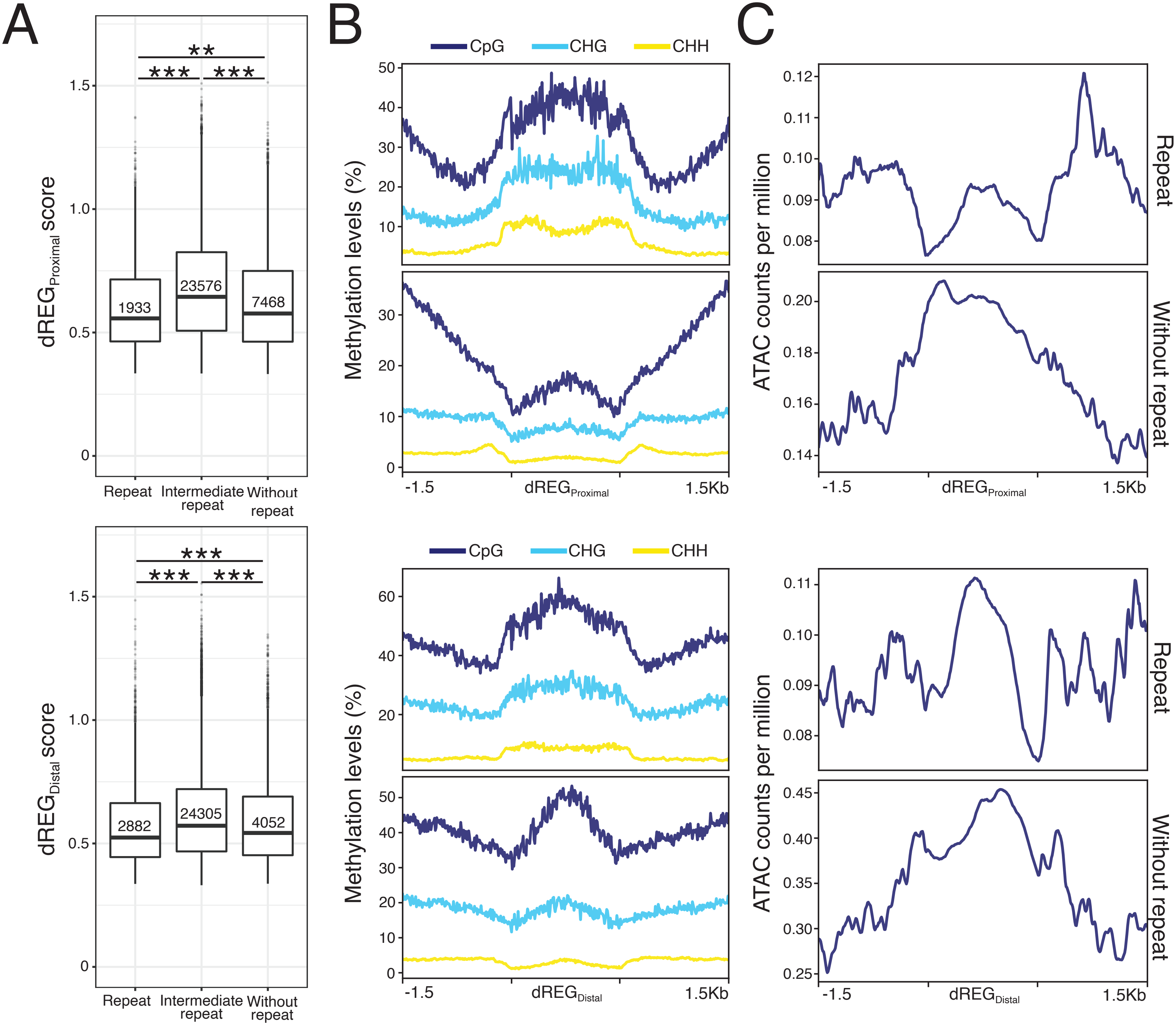
Figure 4. Functional characteristics of dREG$_{\mathbf{Proximal}}$ peaks that are detected at 5 prime untranslated regions of genes. (**a**) Expression levels (shown as Reads per kilo base per million mapped reads, RPKM) for genes with and without dREG$_{\text{Proximal}}$ peaks. Genes with dREG$_{\text{Proximal}}$ peaks were divided by repeat class and their expression levels were compared to the genes without dREG$_{\text{Proximal}}$ peaks. Numbers in boxplot represent sample size of genes. (**b**) Gene expression levels for genes with dREG$_{\text{Proximal}}$ peaks divided by dREG score and repeat classification. Asterisk (*) indicate significant differences after all pairwise comparisons using the Mann-Whitney U test. Numbers in boxplot represent sample size of genes.

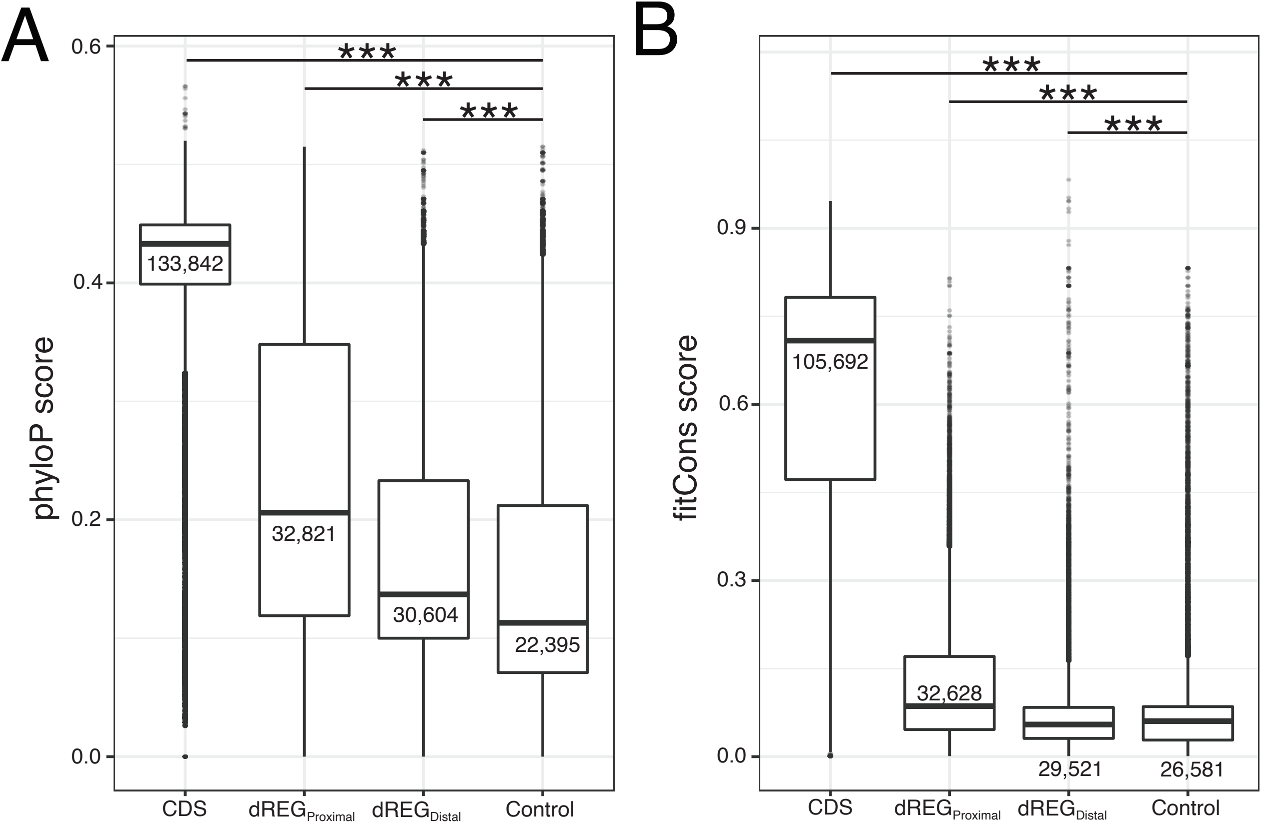
Figure 5. Chromatin features and functionality of dREG$_{\mathbf{Distal}}$ peaks. (**a**) PRO-seq read count enrichment surrounding TADs. Shown are 10 kbp upstream and downstream of TADs with the TAD scaled to 5 kbp. PRO-seq read counts were averaged in 100 bp windows. (**b**) Proportion of gene-gene, gene-noncoding, and noncoding-noncoding loops that were detected using Pore-C sequencing. (**c**) Aggregate Peak Analysis (APA) plots showing the aggregated Pore-C contacts around chromatin loops identified in all chromosomes (left) and only between a noncoding-gene loop (right). The plot is a pile-up of 25 kbp upstream and downstream of loop anchors (centered in each axis) of every identified loop. Color represents log2 fold enrichment of the observed aggregated matrix over a normalization matrix that was aggregated from randomly shifted controls regions across the chromosome. (**d**) Gene expression level for genes that are contacting a noncoding anchor with either an ATAC peak, dREG peak, or has no annotation. Asterisk

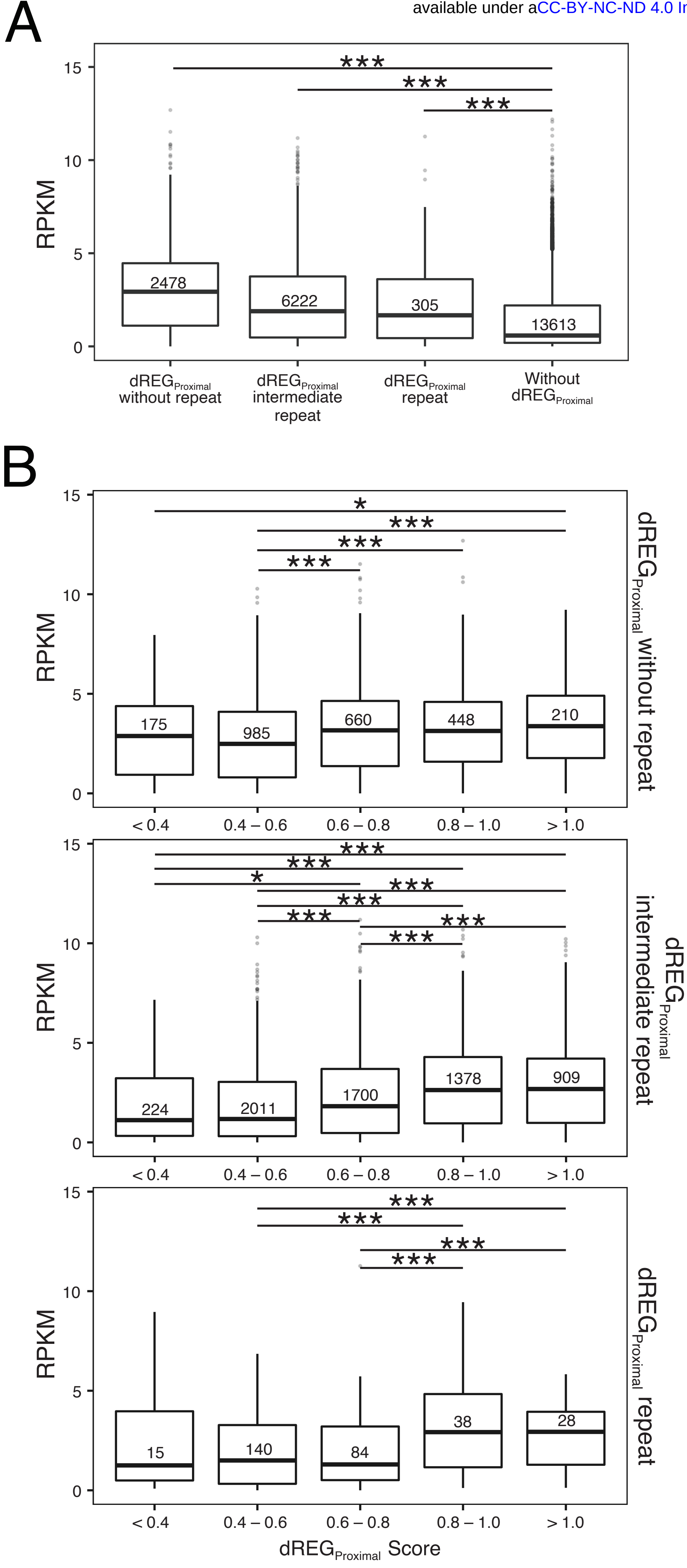
(**) indicates a p-value < 0.01 and *** indicates a p-value < 001 after an FDR-corrected Mann-Whitney U test.

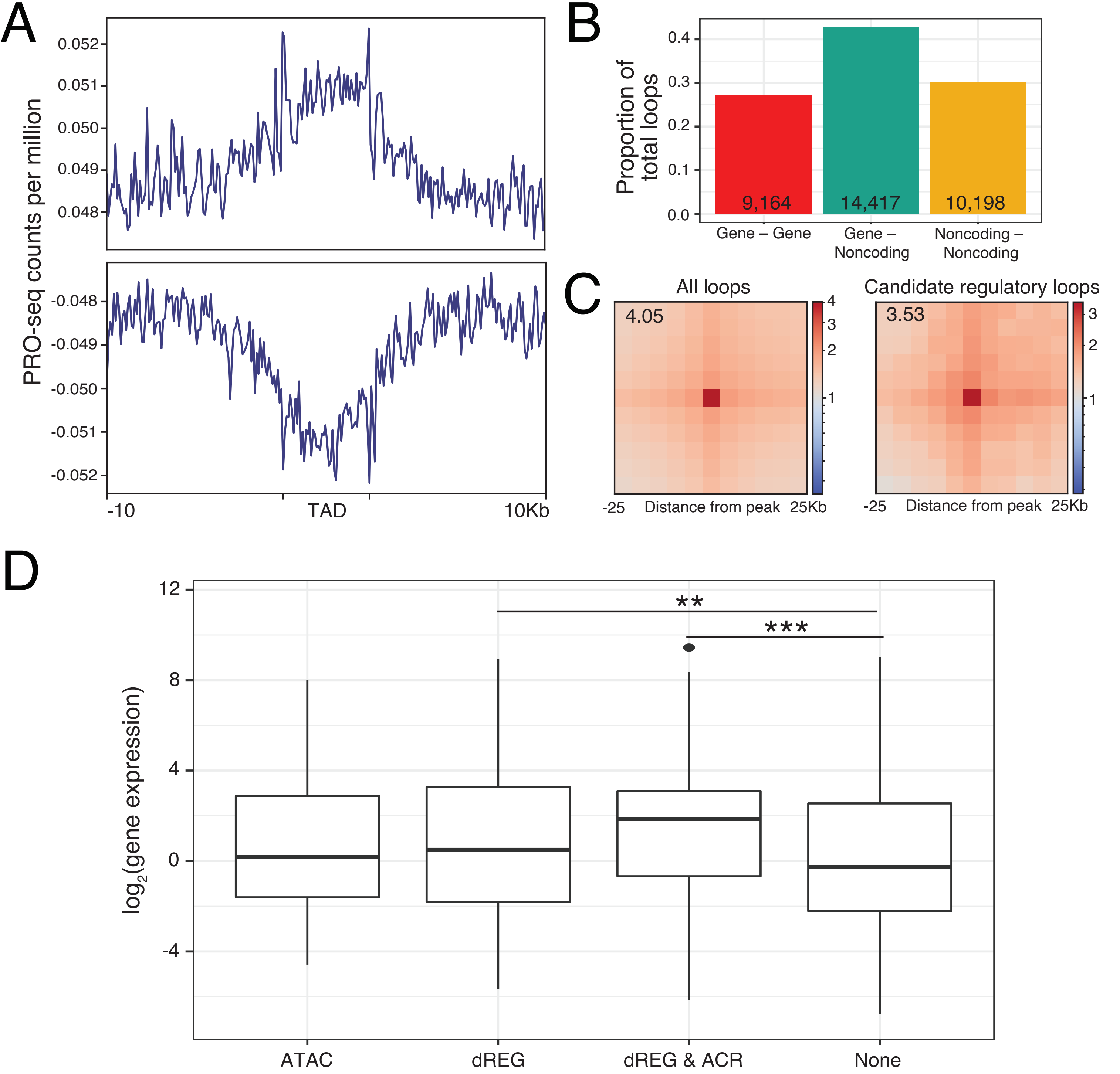
Figure 6. Enrichment of eQTLs within dREG$_{\mathbf{distal}}$ peaks. Histogram shows the bootstrap distribution of the total number of eQTLs overlapping a random region of the genome that are matched for size and total number of dREG$_{distal}$ peaks. The red dotted line shows the total number of eQTLs overlapping the dREG$_{distal}$ peaks.

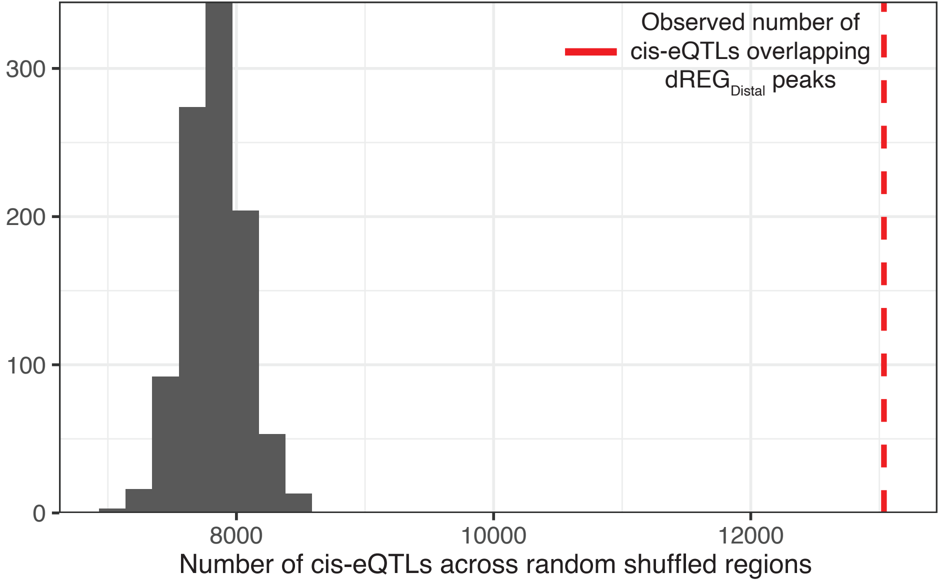

A
dREG Score
< 0.4
0.4–0.6
0.6–0.8
0.8–1.0
> 1.0
PRO-seq counts per million
0.12
0.10
0.08
0.06
-0.04
-0.06
-0.08
-0.10
-0.12
-0.14
-1.5
dREG peak
1.5Kb
B
Genic 5,682
Proximal 32,977
Distal 31,239
C
dREG 64,224
5,674
ATAC 16,814
D
dREG
ATAC
ATAC counts per million
0.130
0.125
0.120
0.115
1.50
1.25
1.00
0.75
0.50
H3K27me3 ratio (ChIP/Input)
1.05
1.00
0.95
0.90
0.85
0.80
1.6
1.5
1.4
1.3
1.2
1.1
H3K27ac ratio (ChIP/Input)
1.4
1.3
1.2
1.3
1.2
1.1
1.0
0.9
H3K4me3 ratio (ChIP/Input)
1.8
1.6
1.4
1.2
1.6
1.4
1.2
H3K18ac ratio (ChIP/Input)
1.12
1.10
1.08
1.06
1.05
1.00
0.95
0.90
0.85
PRO-seq (+) strand counts per million
0.08
0.07
0.06
0.10
0.09
0.08
0.07
0.06
PRO-seq (−) strand counts per million
-0.06
-0.07
-0.08
-0.09
-0.06
-0.08
-0.10
-0.12
Methylation levels (%)
15
14
13
12
11
10
26
24
22
20
18
-1.5
dREG peak
1.5Kb
-1.5
ATAC peak
1.5Kb


A
B
C
**
***
***
1.5
1.0
0.5
0
$\text{dREG}_{\text{Proximal}}$ score
1933
23576
7468
Repeat
Intermediate repeat
Without repeat
***
***
***
$\text{dREG}_{\text{Distal}}$ score
2882
24305
4052
Repeat
Intermediate repeat
Without repeat
CpG
CHG
CHH
Methylation levels (%)
50
40
30
20
10
0
-1.5
$\text{dREG}_{\text{Proximal}}$
1.5Kb
CpG
CHG
CHH
Methylation levels (%)
60
40
20
50
40
30
20
10
0
-1.5
$\text{dREG}_{\text{Distal}}$
1.5Kb
ATAC counts per million
0.12
0.11
0.10
0.09
0.08
Repeat
0.20
0.18
0.16
0.14
Without repeat
-1.5
$\text{dREG}_{\text{Proximal}}$
1.5Kb
ATAC counts per million
0.11
0.10
0.09
0.08
Repeat
0.45
0.40
0.35
0.30
0.25
Without repeat
-1.5
$\text{dREG}_{\text{Distal}}$
1.5Kb


A
0.6
0.4
0.2
0.0
phyloP score
***
***
***
133,842
32,821
30,604
22,395
CDS
dREG$_{Proximal}$
dREG$_{Distal}$
Control
B
0.9
0.6
0.3
0.0
fitCons score
***
***
***
105,692
32,628
29,521
26,581
CDS
dREG$_{Proximal}$
dREG$_{Distal}$
Control


A
***
***
***
15
10
5
0
RPKM
2478
6222
305
13613
dREG$_{Proximal}$ without repeat
dREG$_{Proximal}$ intermediate repeat
dREG$_{Proximal}$ repeat
Without dREG$_{Proximal}$
B
*
***
***
***
175
985
660
448
210
dREG$_{Proximal}$ without repeat
***
***
*
***
***
***
***
***
224
2011
1700
1378
909
dREG$_{Proximal}$ intermediate repeat
***
***
***
***
15
140
84
38
28
dREG$_{Proximal}$ repeat
< 0.4
0.4 – 0.6
0.6 – 0.8
0.8 – 1.0
> 1.0
dREG$_{Proximal}$ Score


A
PRO-seq counts per million
0.052
0.051
0.050
0.049
0.048
-0.048
-0.049
-0.050
-0.051
-0.052
-10
TAD
10Kb
B
Proportion of total loops
0.4
0.3
0.2
0.1
0.0
9,164
14,417
10,198
Gene – Gene
Gene – Noncoding
Noncoding – Noncoding
C
All loops
4.05
4
3
2
1
-25
Distance from peak
25Kb
Candidate regulatory loops
3.53
3
2
1
-25
Distance from peak
25Kb
D
log2(gene expression)
12
8
4
0
-4
**
***
ATAC
dREG
dREG & ACR
None



Observed number of cis-eQTLs overlapping dREG$_{\text{Distal}}$ peaks
300
200
100
0
8000
10000
12000
Number of cis-eQTLs across random shuffled regions