Title: **RNA polymerase mapping in plants identifies intergenic regulatory elements enriched in causal variants**

Roberto Lozano[1], Gregory T. Booth[2], Bilan Yonis Omar[3], Bo Li[4], Edward S. Buckler[1,5,6], John T. Lis[2], Dunia Pino del Carpio [1*], Jean-Luc Jannink [1,6,*]

Affiliations:

[1] Plant Breeding and Genetics, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA

[2] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

[3] Montpellier SupAgro, 34060 Montpellier Cedex 02, France

[4] State Key Laboratory of Plant Genomics and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Science, Beijing, China

[5] Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA

[6] United States Department of Agriculture, Agricultural Research Service (USDA-ARS) R.W. Holley Center for Agriculture and Health, Ithaca 14853, NY, USA

Present Address:
Gregory Booth
Department of Genome Sciences, University of Washington, Seattle, WA 98195
Dunia Pino del Carpio
Bayer Crop Science, Rotterdam, Netherlands

* Correspondence to: Jean-Luc Jannink (jeanluc.jannink@usda.gov)

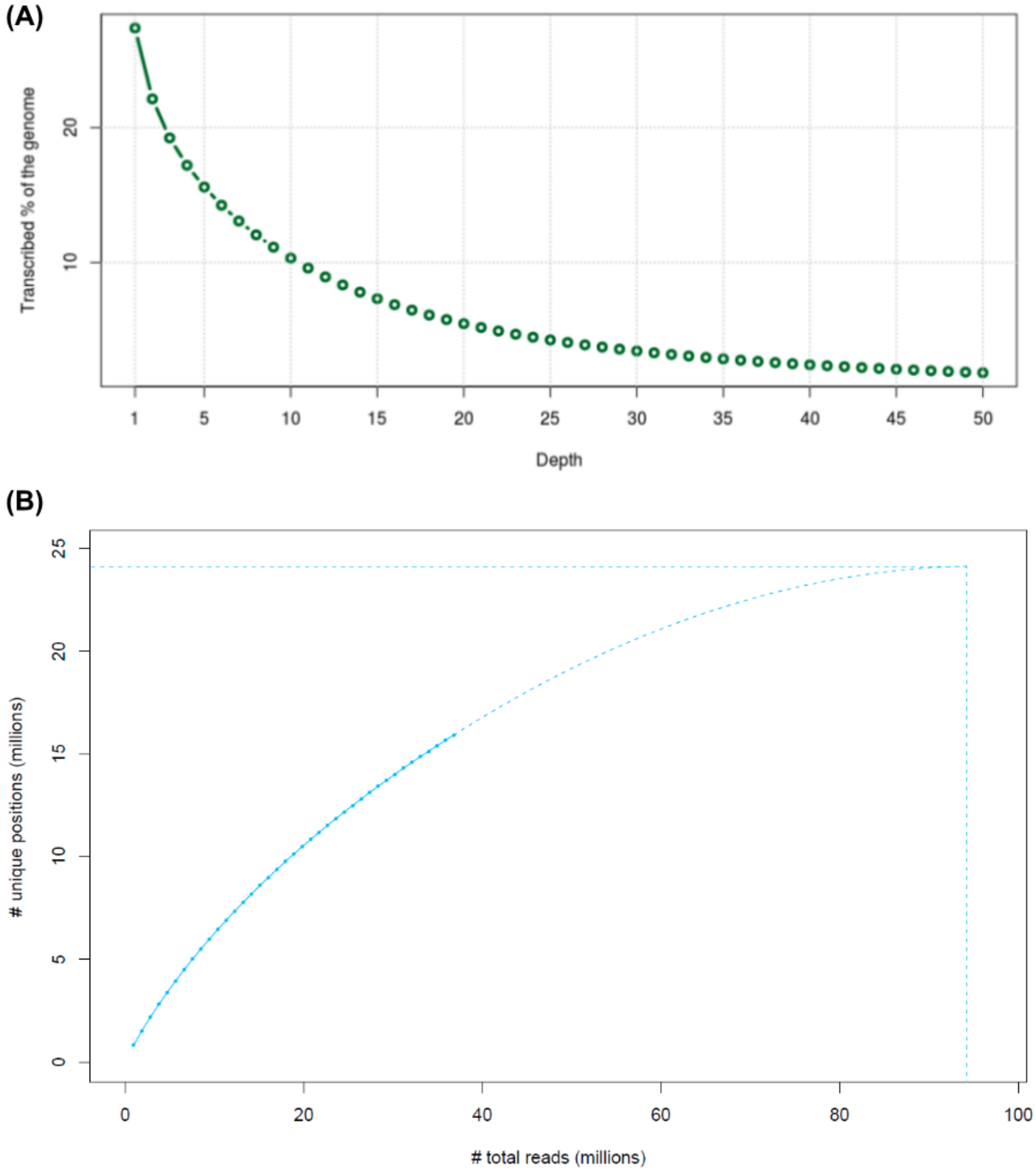


Figure S1. PRO-seq coverage in cassava. (A) The cumulative percentage of the genome covered by Pro-seq reads is plotted as a function of depth. The first dot in the graph should be read as: "27% of the genome was covered by at least one read," and the tenth dot as: "10% of the genome is covered by 10 or more reads." In total, we observed active transcription covering ~27% of the cassava genome in 6-week-old seedlings at 37 million mapped reads. This percentage was smaller than what was observed in Arabidopsis (~40%), which might be due to the difference in genome sizes. (B) Number of unique positions covered as a function of read depth as calculated by "bed-metric". This saturation curve shows that even at 37 million reads the sequencing depth is still insufficient to accurately estimate the portion of the genome that is transcribed.

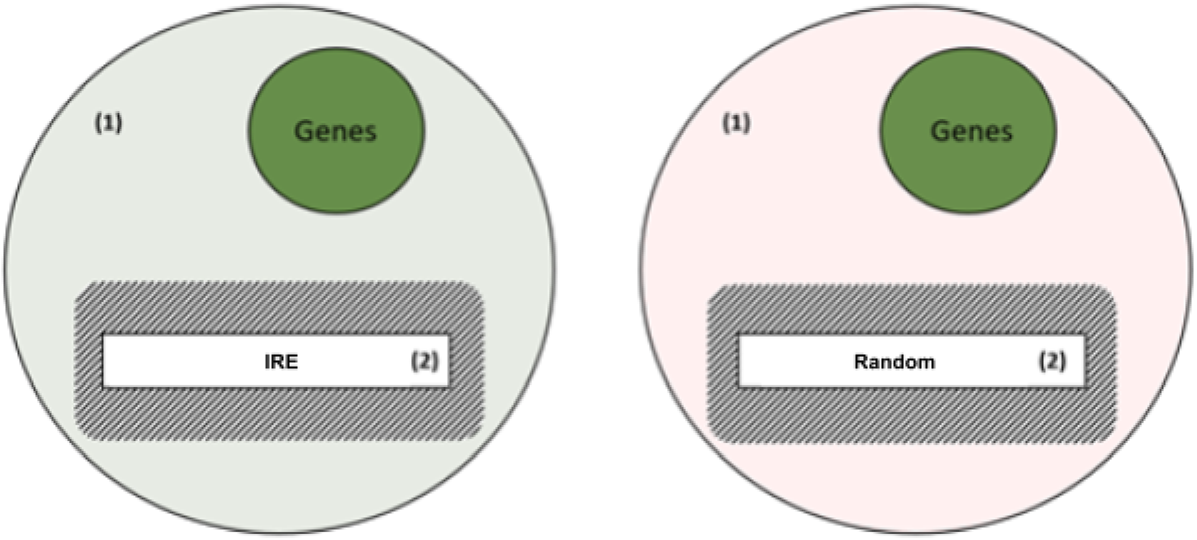


Figure S2. Diagram of the sets used for Genomic Partitioning. Each large circle represents the cassava genome. The first set (green) is composed of two disjoint regions, (2) includes just the regions annotated as IRE and (1) represents the rest of the genome (ROG). The second set (red) follows the same idea as the green one, however (2) now represents random intergenic regions with similar size and chromosome distribution as the IREs. In both sets the striped area represents markers in the ROG that were in high LD with SNPs located in the IRE/random partition, and that were eliminated from further analysis.

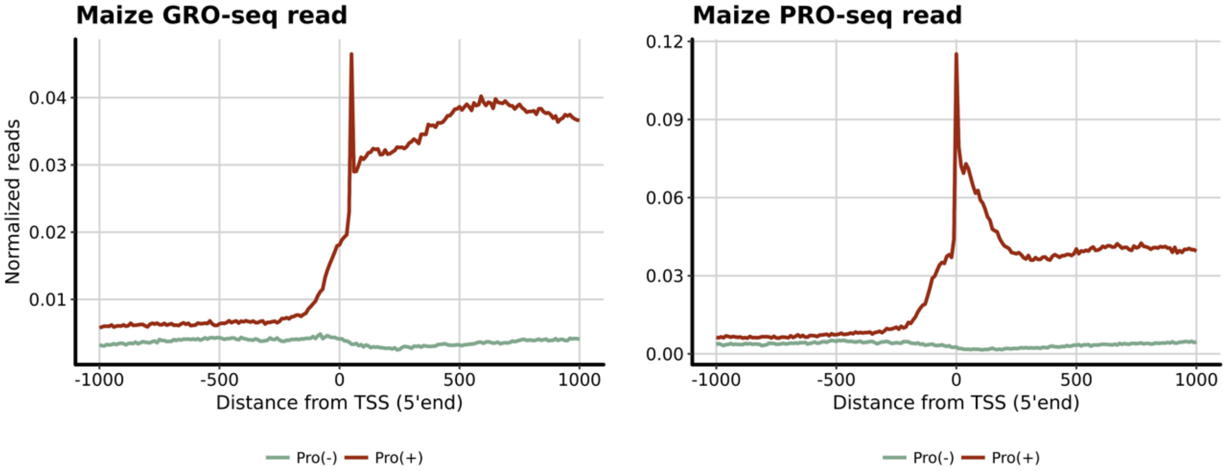
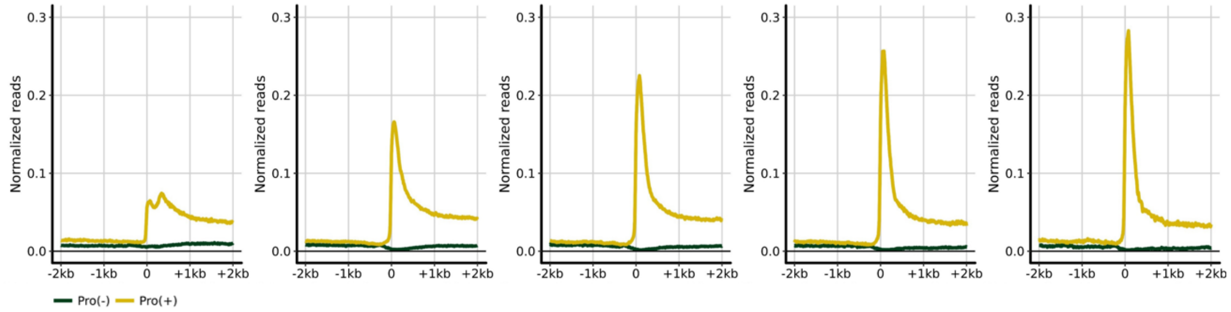


Figure S3. GRO-Maize vs PRO-seq profile around genes in the maize genome. Nascent transcription at the TSS site in maize as profiled with GRO-seq and PRO-seq. The spiky profile at the TSS is common across both libraries. The GRO-maize libraries were prepared in the absence of Sarkosyl (Erhard et al. 2015). Differences between libraries can be attributed to lack of Sarkosyl as this chemical is required to block the polymerase initiation (Hetzel et al. 2016).



Set 1: Pausing index (PI) < 2

Set 2: 2 < PI < 4

Set 3: 4 < PI < 6

Set 4: 6 < PI < 8

Set 5: 8 < PI < 10

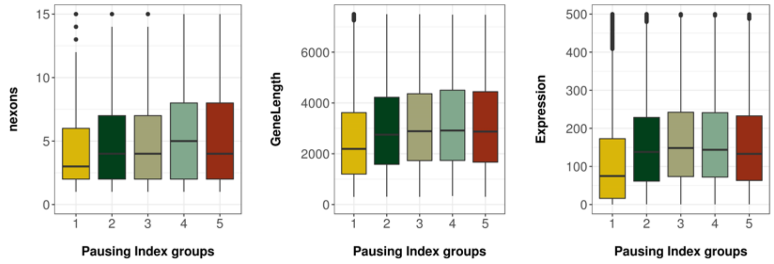


Figure S4. Categorization of expressed cassava genes according to their pausing index. Pausing Indices (PI) were calculated for the cassava genes that showed an expression > 0. See Supplemental table 1. Metagene plots were produced for each category. Pausing index was calculated as the average coverage in the promoter region (100 upstream of the Transcription Start Site (TSS) to 300 downstream of the TSS) divided by the average coverage of the gene body (300bp downstream of the TSS to the Polyadenylation site (PAS). Boxplots showing the distribution of number of exons (nexons), gene length and raw expression values across the 5 categories. Number of genes per set, set 1: 7891, set 2: 6823, set 3: 4984, set 4: 3089, set 5: 1745.

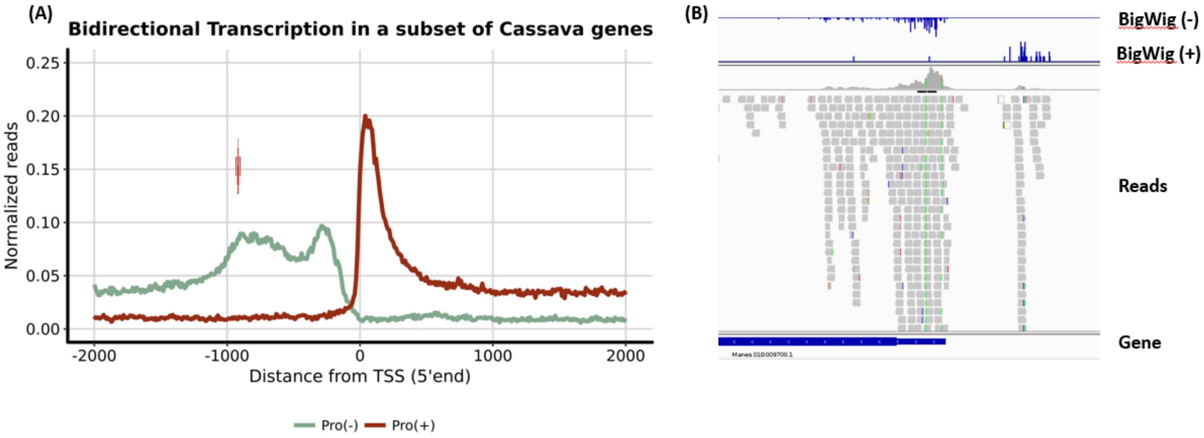


Figure S5. Bidirectional transcription of a subset of cassava genes. (**A**) A small subset of cassava genes (n = 800) shows this pattern of transcription around the TSS. We observe an extended peak (red arrow) that in some cases is due to genes being close to another gene that is transcribing in the opposite direction. **(B)** A clear example of a gene presenting divergent transcription (Manes.01G009700) as shown by IGV (Thorvaldsdottir, Robinson, and Mesirov 2013). The first panel shows the BigWig profiles around the TSS. The Bam files were plotted in the medium panel and the Gene model in the lower panel.

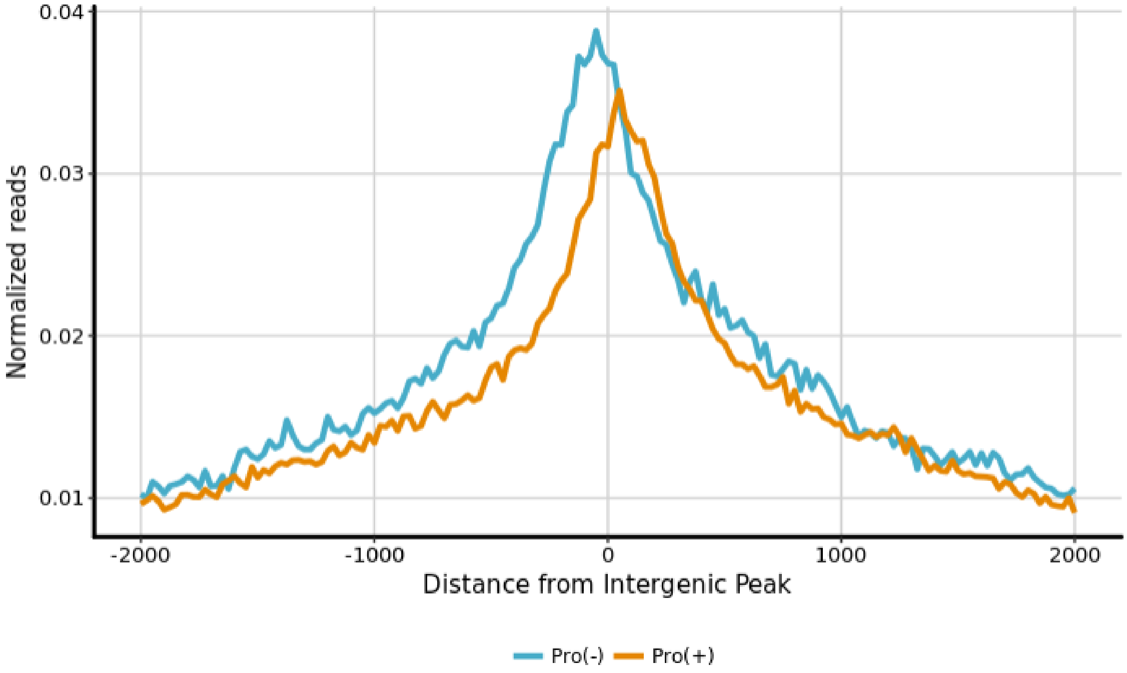


Figure S6. Cassava PRO-seq peaks identified in intergenic regions. To explore the possibility of finding enhancer RNAs (eRNAs) in cassava we used HOMER (Heinz et al. 2010) CHIP-seq peak caller to identify PRO-seq peaks far from genes (3kb distant from the 5'UTR and the 3' UTR of any gene). We identified ~2000 peaks in intergenic regions showing a bi-directional pattern similar to what was observed in mammals and other metazoans.

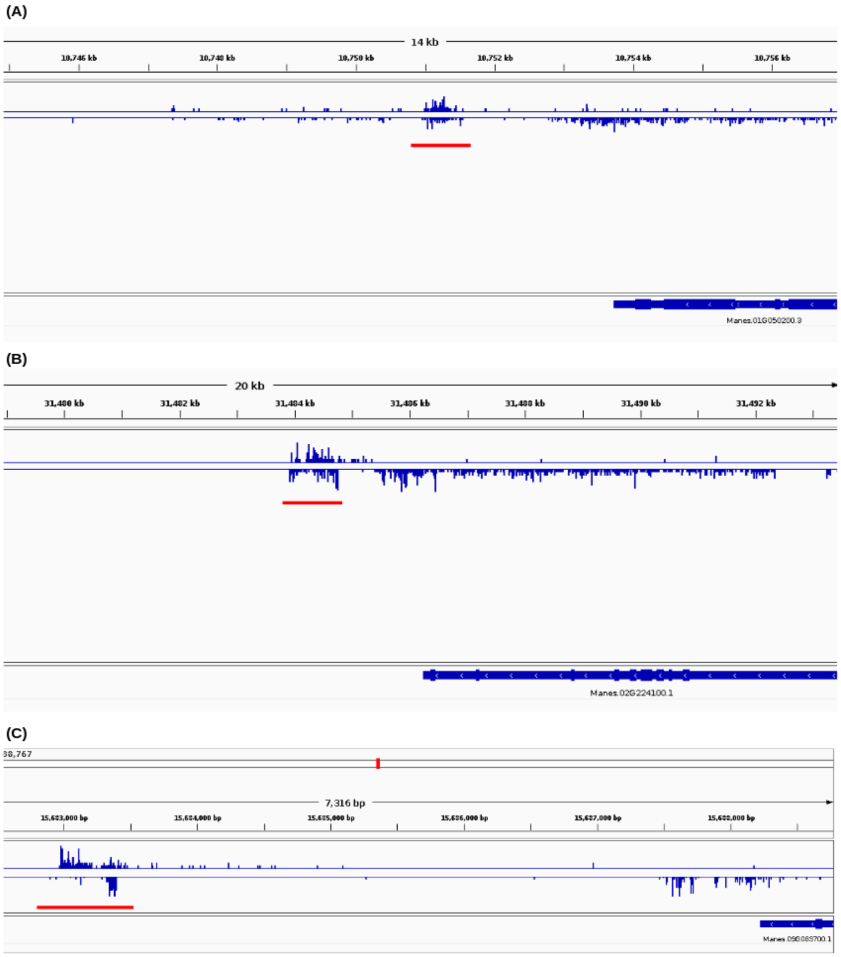


Figure S7. Transcription of individual elements identified by PRO-seq/dREG in cassava. Bigwig tracks for the minus and plus strand show transcription in both strands at elements identified by dREG (underscored in red). In contrast to nascent transcription of genes that shows a primarily single stranded transcription. The regions shown are located in chromosome 1 (A), chromosome 2 (B), and chromosome 9.

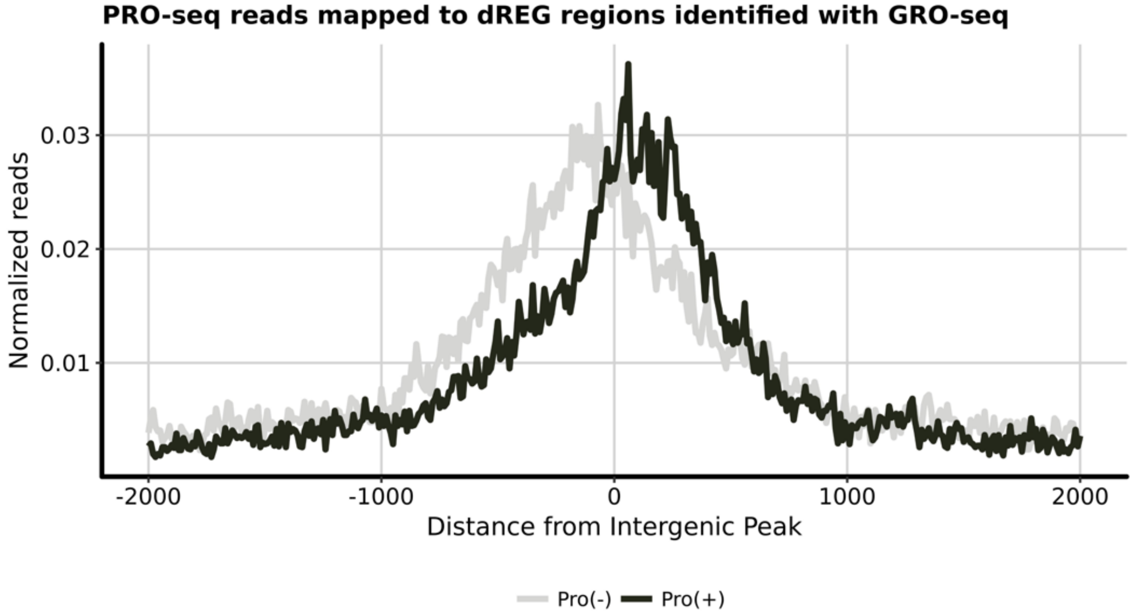


Figure S8. Mapping PRO-maize reads to IRE identified using GRO-maize. The transcription at these candidates seems to be conserved across different techniques as the PRO-seq data generated in this study on maize seedling shows clear signs of transcription in the same regions. PRO-seq data in maize was not used to identify IREs using dREG because the sequencing depth was low.

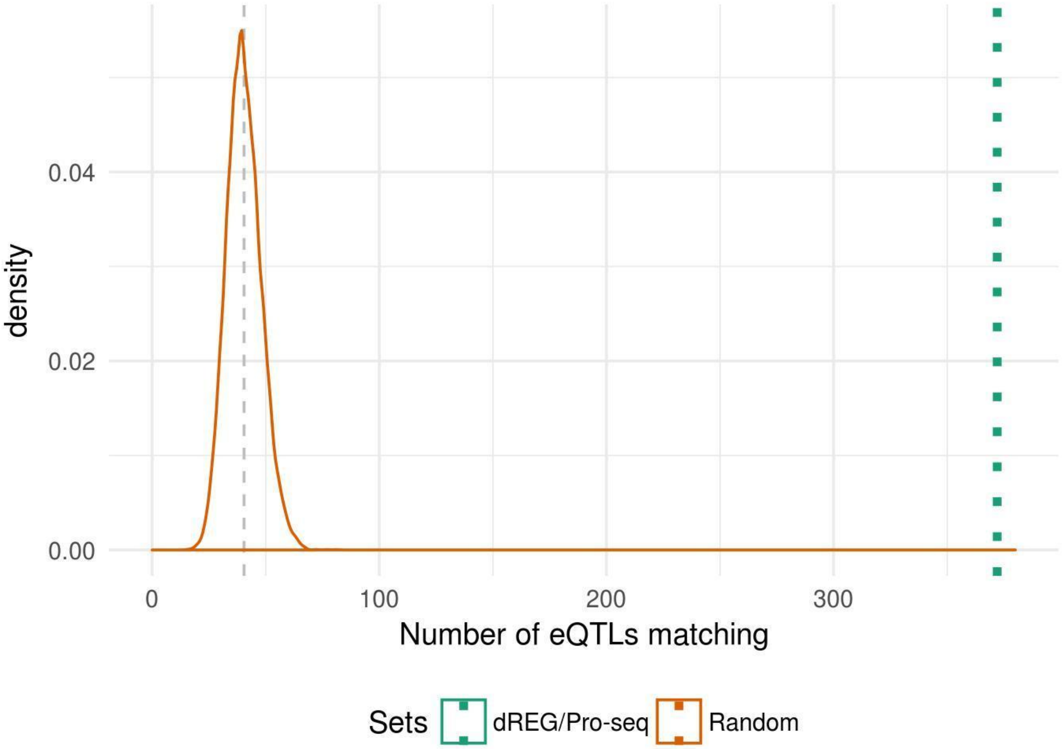


Figure S9. eQTL enrichment in sequences identified as IRE. To further characterize the IRE candidates identified by dREG in maize and explore whether they house more phenotypically-relevant SNPs, we compared our results to a list of previously identified expression Quantitative Trait Loci (eQTLs) in maize kernels. We created a null distribution of the overlap using 10,000 random sets of 4,135 random intergenic regions with the same size as the IRE candidates in orange. The dotted green lines show the observed overlap (8.9% of the total IRE).

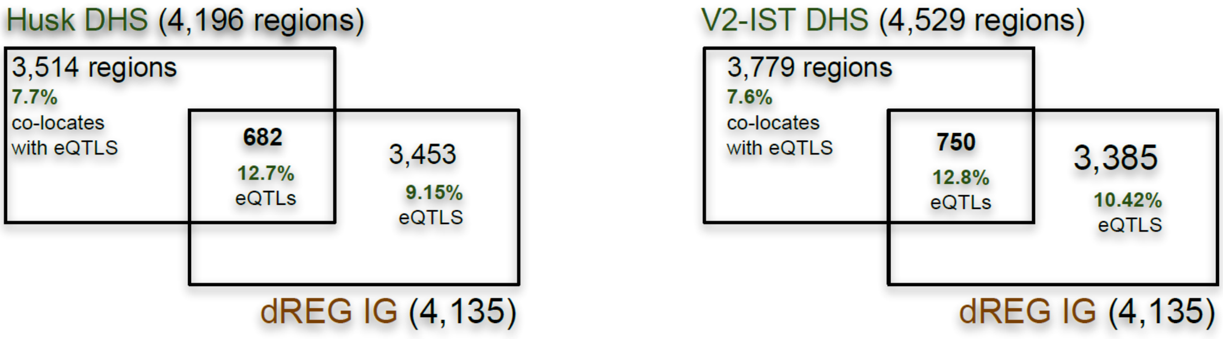


Figure S10. Comparison between open chromatin sites and candidate regions identified by dREG. We compared the levels of eQTL enrichment among the categories defined by the intersection of open chromatin sites and IRE candidates identified using dREG.

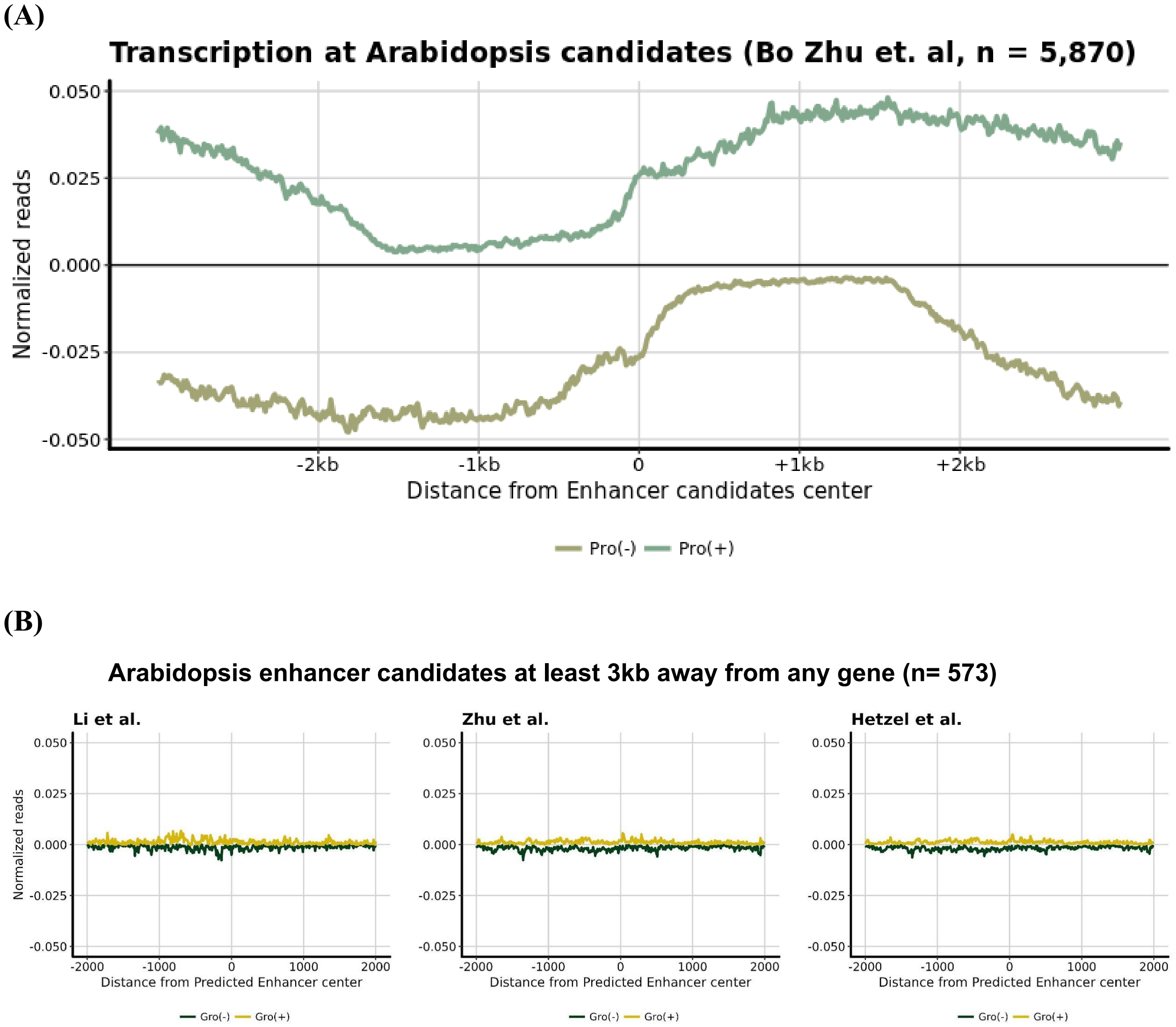


Figure S11. Mapping GRO-seq reads to Enhancer candidates in Arabidopsis. (A) Bo Zhu et al. (B. Zhu et al. 2015) identified 5,870 (File S5) intergenic enhancer candidates present in both leaf and flower tissues based on DNase-seq data. We aligned the Arabidopsis GRO-seq to these regions and did not find a clear pattern as observed in maize or cassava. Reads mapping to the negative strand [Gro (-)] are shown with negative values. The transcription of this enhancer candidates might have been altered by their proximity to nearby genes. **(B)** We chose a subset of the enhancers candidates which were at least 3,000bp away from the nearest gene (n = 573, File S6). Transcription around these regions showed only background expression in all three Arabidopsis GRO-seq datasets analyzed; Hetzel et al, Li et al and Zhu et al.