Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes

Peter A. Crisp[a,b,1], Alexandre P. Marand[c], Jaclyn M. Noshay[a], Peng Zhou[a], Zefu Lu[c,d], Robert J. Schmitz[c], and Nathan M. Springer[a,1]

[a]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108; [b]School of Agriculture and Food Sciences, The University of Queensland, Brisbane, QLD 4072, Australia; [c]Department of Genetics, University of Georgia, Athens, GA 30602; and [d]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, 100081, China

Edited by Steven E. Jacobsen, University of California, Los Angeles, CA, and approved August 11, 2020 (received for review May 21, 2020)

The genomic sequences of crops continue to be produced at a frenetic pace. It remains challenging to develop complete annotations of functional genes and regulatory elements in these genomes. Chromatin accessibility assays enable discovery of functional elements; however, to uncover the full portfolio of cis-elements would require profiling of many combinations of cell types, tissues, developmental stages, and environments. Here, we explore the potential to use DNA methylation profiles to develop more complete annotations. Using leaf tissue in maize, we define ∼100,000 unmethylated regions (UMRs) that account for 5.8% of the genome; 33,375 UMRs are found greater than 2 kb from genes. UMRs are highly stable in multiple vegetative tissues, and they capture the vast majority of accessible chromatin regions from leaf tissue. However, many UMRs are not accessible in leaf, and these represent regions with potential to become accessible in specific cell types or developmental stages. These UMRs often occur near genes that are expressed in other tissues and are enriched for binding sites of transcription factors. The leaf-inaccessible UMRs exhibit unique chromatin modification patterns and are enriched for chromatin interactions with nearby genes. The total UMR space in four additional monocots ranges from 80 to 120 megabases, which is remarkably similar considering the range in genome size of 271 megabases to 4.8 gigabases. In summary, based on the profile from a single tissue, DNA methylation signatures provide powerful filters to distill large genomes down to the small fraction of putative functional genes and regulatory elements.

DNA methylation | chromatin accessibility | cis-regulatory elements

There is a rapidly growing knowledge of the genome structure and sequence for many organisms. However, to fully utilize this resource, it is critical to identify and annotate the functional elements within the genome. In particular, there are two major challenges in providing high-quality annotations of functional elements in complex eukaryotic genomes: correctly identifying functional genes and identification of cis-regulatory elements (CREs).

The challenge of identifying functional genes relates in part to the enigmatic concept of a gene. While classical definitions of a gene were commonly based on mutant phenotypes, it is clear that genetic redundancy or environment-specific phenotypic manifestations complicate our ability to identify phenotypes, even for functionally important genes. Genomics-based efforts to define gene models are often based on a combination of evidence of transcripts and/or ab initio predictions. Yet, gene models are best considered a hypothesis as to the existence of a gene (1). By and large, the majority of functional gene products can likely be captured based on identification of putative genes that are conserved in similar order among related species, often termed syntenic genes (1). However, there are also cases of functional genes that are created following gene duplication and/or transposition that are common in many plant genomes. One potential solution is to identify putative genes through genome-wide annotation and then to use chromatin features to filter the genes to highlight models that are more likely to retain function. These approaches have been applied in sorghum (2) and maize (3).

The problem of identifying potential CREs is even more challenging. In plants with large genomes, CREs can occur tens to hundreds of kilobase pairs (kb) from their target genes (4). These regulatory regions, including gene-distal (hereafter distal) CREs, have established roles in domestication and agronomic traits, for instance in *Zea mays* (maize) (5–11). Although only a handful of distal CREs have been characterized, recent studies suggest their prevalence in plants (4, 12–17). Yet, these regions do not necessarily produce easily detectable products (like transcripts) or have sequence features that can be identified, such as protein-coding potential. Several approaches that survey accessible chromatin (12, 13, 17) or interactions of intergenic regions with gene promoters (18–20) are providing major insights for the identification of putative CREs. However, many of these technologies are specific to the tissue or cell type that is assayed. A complete understanding of the potential CREs within a particular species would require profiling of chromatin accessibility and/or chromatin interactions in a wide variety of tissues, cell types, and conditions.

Although chromatin accessibility, histone modifications, and chromatin interactions often show substantial variation in different tissues (12, 14, 21), the majority of DNA methylation patterns are quite stable in plant species, especially during vegetative development (22–24) and in the face of environmental

Significance

Crop genomes can be very large, with many repetitive elements and pseudogenes. Distilling a genome down to the relatively small fraction of regions that are functionally valuable for trait variation can be like looking for needles in a haystack. The location of these regions is often not obvious, and current detection technologies are impractically expensive and intensive for many research projects. The unmethylated regions in a genome are highly stable during vegetative development and can reveal the locations of potentially expressed genes or cis-regulatory elements. This approach provides a framework toward complete annotation of genes and discovery of cis-regulatory elements using methylation profiles from only a single tissue.

Author contributions: P.A.C., R.J.S., and N.M.S. designed research; A.P.M., J.M.N., and Z.L. performed research; P.A.C., A.P.M., J.M.N., P.Z., and Z.L. analyzed data; and P.A.C. and N.M.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: p.crisp@uq.edu.au or springer@umn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2010250117/-/DCSupplemental.

stress (25–28). Likewise, prior reports have identified large DNA methylation "valleys" that are stable over development (29). There are well-characterized examples of specific changes in DNA methylation in endosperm tissues (30, 31) as well as in specific cell types in plant gametophytes (32–36). However, the majority of DNA methylation patterns are quite stable in different vegetative tissues, especially for DNA methylation in the CG and CHG contexts. In contrast, several studies have provided evidence for developmental or tissue-specific changes in CHH methylation (37–42). It should be noted, however, that the majority of these examples point to cases in which the level of CHH methylation at specific regions changes, but these regions often have some level of CHH methylation in all tissues (24, 38).

Prior studies have found that the majority of regions of chromatin accessibility are hypomethylated (4, 12, 13, 17). Here, we reverse the approach and identify the unmethylated regions (UMRs) of the maize, barley, sorghum, rice, and *Brachypodium* genomes and compare the genomic distribution of UMRs with tissue-specific chromatin accessibility and provide evidence for functional roles of UMRs. We demonstrate that unmethylated regions of the genome, particularly in plant species with large genomes, provide useful information for identification of functional genes and CREs. This improves annotation of complex crop genomes and provides clear hypotheses about the portions of the genome that likely contain functional elements.

Results

In general, the maize genome is highly methylated, with only a small portion of the genome lacking DNA methylation (43–45). Deep whole-genome bisulfite sequencing (WGBS) was performed on seedling leaf tissue of the inbred B73, which generated ~930 million reads, providing ~28× projected raw average coverage per strand (15.7× per cytosine average coverage following alignment and quality filtering; Dataset S1). DNA methylation levels in the

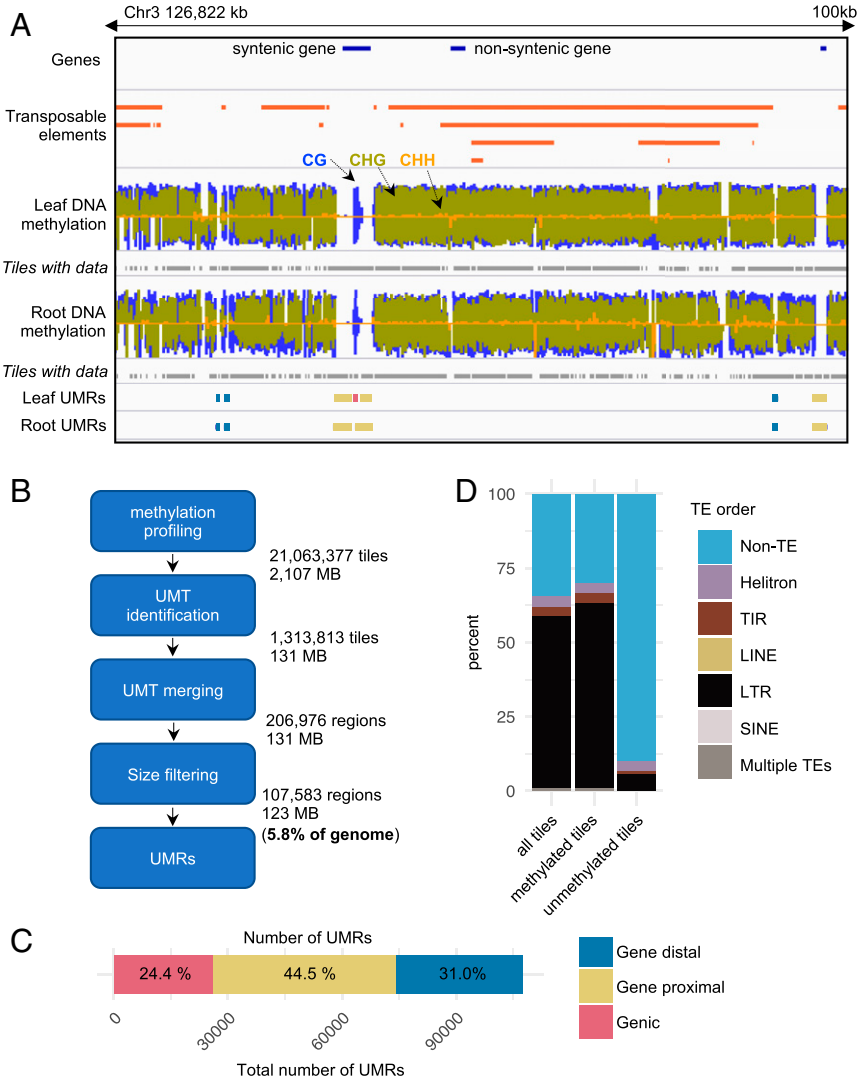

Fig. 1. Identifying unmethylated regions in the maize genome. (*A*) Example distribution of unmethylated regions (UMRs) in a 100-kb locus of the maize genome. DNA methylation context: blue, CG; green, CHG; orange, CHH. Under each methylation track, the 100-bp tiles with sufficient data to assess UMR status are shown in gray. (*B*) Pipeline overview of unmethylated region identification in maize leaf. (*C*) Genomic distribution of UMRs. Proximal UMRs are defined as those that overlap a 2-kb window upstream of the TSS or 2 kb downstream of the TTS (44.5%, $n = 47{,}910$), genic are entirely within the gene locus boundaries (24.4%, $n = 26{,}298$), and distal are >2 kb from a gene (31.0%, $n = 33{,}375$). (*E*) Percentage of UMTs (100-bp tiles) that overlap transposable elements (TEs). TIR, terminal inverted repeat; LINE, long interspersed nuclear element; LTR, long terminal repeat; SINE, short interspersed nuclear element; 10.7% overlap TEs, including 6.08% LTR, 3.50% Helitron, 1.02% TIR, 0.0388% LINE, 0.0272% SINE, and 0.0584% multiple TEs from different orders.

CG, CHG, and CHH context were determined for each 100-bp tile of the maize genome. While some regions lack cytosines in this context or could not be assessed due to lack of uniquely mapping reads, we were able to obtain DNA methylation estimates for 16.03 million tiles—~1.6 Gb—that contained at least two cytosines and an average of at least 5× coverage per cytosine per strand, representing 76.1% of the maize genome. A visual examination of a representative ~100-kb region containing two genes revealed that the majority of tiles are highly methylated; however, there are examples of unmethylated regions near syntenic genes and in distal regions (Fig. 1*A*). Across the maize genome, 8.19% of the 100-bp tiles with data—131 Mb—had very low (<10%) or no detectable DNA methylation in any sequence context, termed unmethylated tiles (UMTs; Fig. 1*B* and *SI Appendix*, Fig. S1*A*). In all analyses of unmethylated regions, we solely focus on tiles with data; tiles with missing data due to lack of cytosines or lack of coverage are not classified as unmethylated. We developed a framework to identify the unmethylated regions (UMRs) in a genome by first hierarchically categorizing each tile into one of six methylation domains (*Methods* and Dataset S2) and then merging

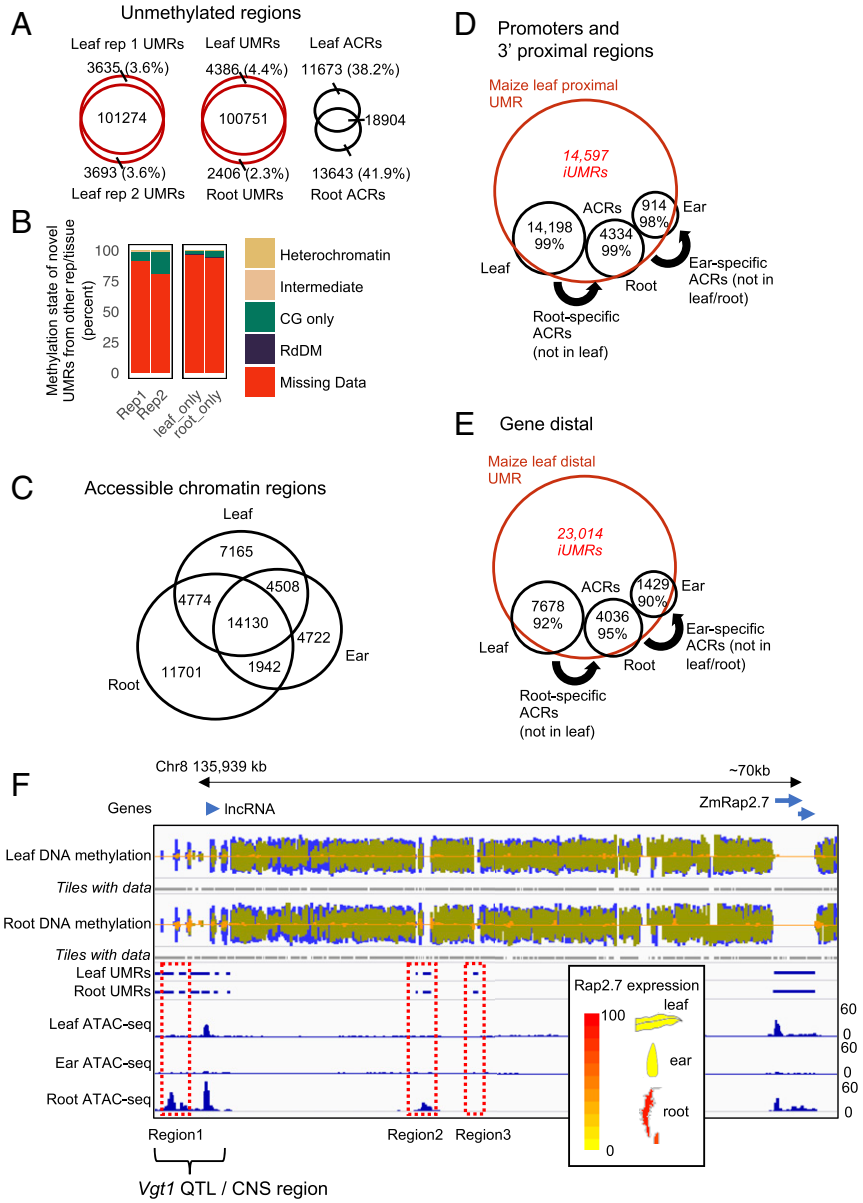

Fig. 2. Comparisons of unmethylated and accessible portions of the maize genome. (*A*) Overlap of unmethylated regions (UMRs) in two independent maize seedling leaf samples, "Rep 1" and "Rep 2," and a seedling root sample. The percentage of UMRs uniquely identified in one of the samples is listed in parentheses. (*Right*) Overlap of leaf and root ACRs is shown for comparison. (*B*) For UMRs uniquely identified in one of the samples in *A*, the methylation data (methylation domain) in the corresponding sample are displayed. For example, over 80% of the UMRs uniquely identified in Rep2 have missing data in Rep1. (*C*) Overlap of accessible chromatin regions (ACRs) from maize leaf ($n = 30,577$), root ($n = 32,547$), and ear ($n = 25,302$). (*D* and *E*) Gene-proximal (*D*) and gene-distal (*E*) UMRs from maize leaf capture the majority of both leaf ACRs as well as root- and ear-specific ACRs. Leaf UMRs were first overlapped with all leaf-proximal (*D*) or -distal (*E*) ACRs, and the percentage of ACRs overlapping is listed. Next, "root-specific" (not in leaf) ACRs were overlapped with leaf UMRs; then "ear-specific" (not in leaf or root) ACRs were overlapped with leaf UMRs. UMRs that do not overlap a leaf, root, or ear ACR are inaccessible (iUMRs) and listed in red. (*F*) Example leaf iUMRs that mark regions that become accessible in other tissues, including regions 1 and 2 (red boxes) upstream of rap2.7. Scale on the ATAC-seq tracks represents total read counts. (*Inset*) Relative gene expression FPKM of rap2.7 in leaf, root, and ear from the maize eFP browser.

adjacent unmethylated tiles into unmethylated regions. We restricted our analysis to 107,583 UMRs of at least 300 bp (detailed in *Methods*), accounting for 5.8% of the maize genome (Fig. 1*B*, *SI Appendix*, Fig. S1 *A–D*, and Dataset S3). Some unmethylated regions <300 bp could mark functional elements, although we have limited evidence for their functionality (*SI Appendix*, Fig. S1 *C* and *D*). UMRs include many examples within genes, in gene-proximal (within 2 kb) regions, and in distal regions at least 2 kb from the nearest gene (Fig. 1 *A* and *C*). A more detailed investigation of the types of features overlapping with UMRs revealed significant enrichment for syntenic genes and depletion within intergenic regions (*SI Appendix*, Fig. S1*E*). Only a small proportion (10.2%) of the unmethylated 100-bp tiles are found to overlap a variety of transposable elements (TE) from different orders (Fig. 1*D*), but, given the expectation that TEs are highly methylated, it was interesting to note that there are 20,232 UMRs in total (18.8%) that overlap maize TEs.

Comparisons of Unmethylated and Accessible Portions of the Maize Genome. There is evidence for enrichment of functional elements within accessible chromatin regions (ACRs) in maize (12, 13, 17). Several studies have found that these accessible regions tend to be hypomethylated (12, 13, 17). A comparison of the UMRs that are identified in B73 seedling leaf and root tissue using available WGBS data (Dataset S1) (46) reveals very few changes in UMRs between independent leaf samples or between tissues (Figs. 1*A* and 2*A*). There are ∼3 to 4% of UMRs identified solely in one of the samples; however, the vast majority of these were due to missing data in the other sample (Fig. 2*B*). Less than 0.1% of the UMRs from one tissue are classified as methylated in the other tissue, suggesting relatively infrequent changes in the UMRs among vegetative tissues in maize. Prior studies have found very few examples of major changes in CG or CHG methylation among vegetative tissues in maize or other plants (23, 24, 44). UMRs differ just as much between biological replicates of leaves as they do between two different tissues. In contrast, chromatin accessibility profiles (Dataset S4) from three distinct tissues show significant variability (Fig. 2*C*).

Given the different dynamics in tissue-specific chromatin accessibility and tissue-specific DNA methylation, we were interested in exploring if UMRs from a single tissue could capture and predict potential ACRs in multiple tissues or conditions. We assessed the overlap of the seedling leaf UMRs with ACRs from three different tissues, including leaf and ear ACRs identified by Ricci et al. (12) and ACRs identified in root (Dataset S3), for gene-proximal (Fig. 2*D*) and gene-distal regions (Fig. 2*E*). As expected, the vast majority of ACRs overlap with UMRs. Over 99% of the promoter ACRs and 92% of the distal ACRs identified in seedling leaf tissue overlap with a UMR defined in seedling leaf tissue. Interestingly, when we focus on ACRs that are found in root tissue (but not in leaf) or in ear tissue (but not in leaf/root), we find that the vast majority of these are unmethylated in leaf tissue as well, despite being inaccessible in leaf (Fig. 2 *D* and *E*). Examination of DNA methylation and ATAC-seq data for several UMRs that exhibit accessibility solely in nonleaf tissues supports the observation of tissue-specific ACRs that are stably unmethylated (Fig. 2*F* and *SI Appendix*, Figs. S2 and S3). In two cases of classic maize genes, *tb1* (8) and *ZmRap2.7* (5), with defined long-distance enhancers, we find that UMRs are stable in multiple tissues, including in leaf tissues, where these genes are not appreciably expressed. In contrast, ACRs at distal regulatory regions and gene-proximal regions only occur in tissues with expression for both genes (Fig. 2*F* and *SI Appendix*, Figs. S2 and S3). For both of these examples, the gene itself is unmethylated in tissues with or without expression (Fig. 2*F* and *SI Appendix*, Fig. S3). Combined, these observations suggest that UMRs defined on a single tissue may capture regions with potential for accessibility in a variety of cell types or tissues, thus providing a prediction of putative functionality.

UMRs Are Indicative of Expression Potential of Genes. To investigate accessibility dynamics of UMRs, UMRs defined on seedling leaf tissues were classified into two groups, accessible UMRs (aUMRs) and inaccessible UMRs (iUMRs), depending on the chromatin accessibility in seedling leaf tissue. In assessing the functional relevance of the aUMRs and iUMRs, we first focused on the UMRs found near gene transcription start sites (TSSs). There are 32,196 UMRs that overlap with the proximal region of maize genes (within 2 kb upstream or 1 kb downstream of the TSS) and 12,867 ACRs within these regions. Nearly all (>98%) of these ACRs overlap with a UMR (Fig. 3*A*). However, 60.7% of UMRs that are located near gene TSS do not overlap an ACR. Considering recent work that demonstrated DNA methylation levels near the ends of the genes could predict "expressibility" of genes (3), we hypothesized that genes that are actively expressed in the tissue used for documenting accessibility (seedling leaf) would be enriched for ACRs/aUMRs, while genes expressed in other tissues and silenced in seedling leaf would be enriched for iUMRs. To do so, we gathered B73 RNA-seq data across more than 240 different samples from tissues, conditions, and developmental stages, including seedling leaf RNA-seq data generated by Ricci et al. (12) (Dataset S5). Genes were classified as "leaf expressed" if they were detected at >1 count per million (cpm) in seedling leaf tissue. The remaining genes were classified as "other tissue" if they were detected in at least one of the other tissues (>1 cpm) or classified as "not expressed." We classified maize genes that are located in syntenic positions relative to other grasses (47) as syntenic (Fig. 3*B*) and the remainder as nonsyntenic (*SI Appendix*, Fig. S4). We then examined the methylation and accessibility of the promoters (defined as 2 kb upstream of TSS to 1 kb downstream of TSS) of these genes (Fig. 3*B*). In some cases, the lack of properly annotated TSSs for some gene models will lead to potential issues, as promoter proximal regions will not be accurately defined. We then examined the methylation and accessibilities patterns for each of the leaf-expressed, other-expressed, and not-expressed categories. First, examining leaf-expressed syntenic genes, we found that they are enriched for aUMRs in the promoter-proximal region (Fig. 3*B*). However, there are almost as many genes expressed in this tissue that contain an iUMR, and these may reflect examples in which the ACR region was too small to be effectively detected using ATAC-seq or expressed with limited accessibility (Fig. 3*B*). Very few genes with leaf expression lack UMRs and ACRs (Fig. 3*B*). Next, examining syntenic genes that are expressed in other tissues, we find these are less likely to contain an aUMR but frequently contain iUMRs (Fig. 3*B*). A total of 1,323 genes expressed in other tissues (21.8% of "other tissue-expressed") contained an aUMR in their promoters, representing genes that are possibly poised in leaf for expression, have unstable transcripts or high transcript turnover, or contain silencing *trans*-factors in their promoters precluding their activation. We also identified cases where genes expressed in other tissues have inaccessible but unmethylated promoters in leaf tissue that become accessible in other tissues, such as *NAC-transcription factor 114* (*nactf114*/*cuc3*; Fig. 3*C*). This gene is silent in leaf tissue but expressed in ear tissue (Fig. 3*D*), yet its promoter is already unmethylated in leaf. Last, genes that are never detected as expressed are much less likely to contain aUMRs or iUMRs and more likely to be nonsyntenic (*SI Appendix*, Fig. S4). Nonsyntenic genes are likely enriched for pseudogenes and transposon fragments compared to syntenic genes (2, 3); thus, these observations also suggest that UMRs and/or ACRs could be useful for discriminating true genes from other classes of gene.

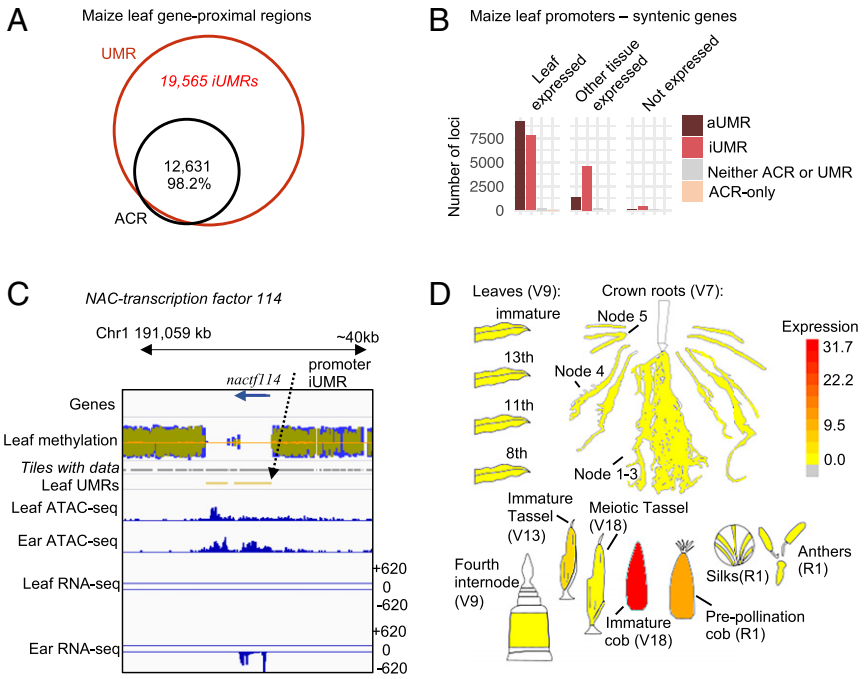

Fig. 3. Gene-proximal "promoter" unmethylated regions in maize. (*A*) Unmethylated regions (UMRs) in gene promoters overlapped with accessible chromatin regions also found in gene promoters (ACRs). (*B*) Relationship between expression of a gene and promoter accessibility and methylation for maize genes syntenic within the grasses. (*C*) Example of a leaf promoter iUMR that may mark a gene for expression in another tissue. The promoter of the *NAC-transcription factor 114* (*nactf114*) Zm00001d031463 is unmethylated but inaccessible in leaf (iUMR, black arrow). The promoter region becomes accessible in ear, and the gene is expressed in ear (RNA-seq). (*D*) The relative expression (FPKM) profile of *nactf114* in representative tissues from the maize eFP browser.

Leaf UMRs Are Enriched for Transcription Factor Binding Sites. The concept that unmethylated regions from a single tissue can reflect sites with regulatory potential in diverse developmental stages or tissues suggests that UMRs from a single tissue could predict potential transcription factor (TF) binding sites, even for TFs only expressed in other tissues. We tested this concept in two different ways. First, we used the combined DAP-seq profiles for 32 maize TFs (12, 48) (Dataset S6). While these DAP-seq enriched regions only represent ~1/10 of the genome, they account for 73% of the unmethylated regions and are enriched for both iUMRs and aUMRs (Fig. 4*A*). Relative to a set of randomized control regions, there is evidence for enrichment ($P < 0.001$) of both aUMRs and iUMRs within the regions identified by DAP-seq (Fig. 4*A*). Second, we used ChIP-seq data for five maize TFs, FASCIATED EAR4 (FEA4), KNOTTED1 (KN1), OPAQUE2 (O2), RAMOSA1 (RA1), and PERICARP COLOR1 (P1) (49–53). Notably, none of these TFs are highly expressed in seedling leaf tissues (Fig. 4*B* and *SI Appendix*, Figs. S5–S10), but are

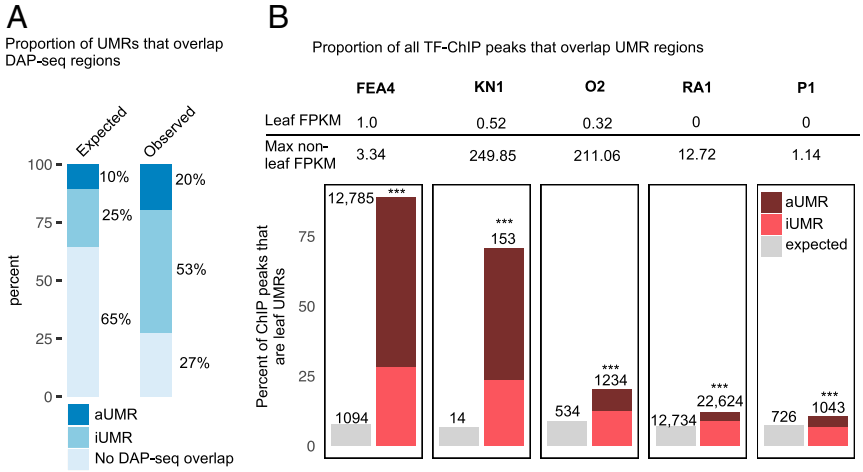

Fig. 4. Transcription factor binding site enrichment in UMRs. (*A*) The expected and observed proportion of UMRs that overlap transcription factor (TF) binding sites identified using DAP-seq. UMRs that overlap TF binding sites are divided into aUMRs and iUMRs. (*B*) Overlap of leaf UMRs with TF binding sites determined using ChIP-seq for TFs which are expressed and function in nonleaf tissues. The expression of each TF in leaf and nonleaf tissues is listed about each plot. The tissues with maximum expression for each tissue are as follows: *FEA4*, stem and SAM; *KN1*, immature cob; *O2*, endosperm; *RA1*, immature cob; *P1*, meiotic tassel. Expected ratios determined using a random set of regions of equal number and size to the relevant contrast. ***Overlap $P < 0.001$ based on 1,000 permutations of randomized control regions.

expressed in other tissues or developmental stages. We were interested in assessing whether the binding sites for these TFs were unmethylated and inaccessible in the absence of their expression (note that the leaf tissue used for methylation profiling comprises a 5-cm section from the leaf and excludes all tissue from the shoot apex, including the meristem). For each TF, the number of ChIP-seq peaks that overlap with aUMRs and iUMRs is greater than expected by chance based on comparison to a set of randomly selected regions ($P < 0.001$; Fig. 4*B*). However, the relative enrichment is quite variable, as some TFs such as FEA4 and KN1 show major enrichments but there is less enrichment for O2, RA1, and P1 (Fig. 4*B*). This could reflect technical variation in the quality of the ChIP-seq datasets or may reflect differences in the potential for some TFs to bind methylated or unmethylated DNA. While some of the TF ChIP-seq peaks are found within aUMRs, there are many that are found within UMRs that are inaccessible in leaf tissue. Indeed, relative to the expected number of iUMRs, we find a significant enrichment ($P < 0.001$) for ChIP-seq binding peaks within UMRs for all of the five TFs (Fig. 4*B*). The observation that these binding sites are highly enriched for iUMRs suggests that the UMRs from this tissue can predict potential binding for these TFs in other developmental stages.

Evidence for Enrichment of Distinct Chromatin Features at Distal UMRs. Cloning of agronomically important QTLs has revealed several examples of important distal cis-regulatory regions that control expression of genes that are tens to hundreds of kb away (5–8). Recent studies in maize have found evidence for many putative distal CREs based on accessible chromatin, chromatin modifications, and three-dimensional chromatin interactions (4, 12, 17–19). There are many distal aUMRs and iUMRs that are located at least 2 kb from the nearest annotated gene (Fig. 1*B*). These capture the majority of distal ACRs identified by ATAC-seq, even when these ACRs are not found in seedling leaf tissue (Fig. 2*E*). We were interested in assessing whether the lack of DNA methylation at these distal regions was associated with unique chromatin profiles or function, especially for the iUMRs. The chromatin modifications from B73 seedling leaf tissue profiled by Ricci et al. (12) were used to compare the chromatin within and surrounding distal iUMRs and aUMRs with a set of random intergenic regions (Fig. 5*A*). Analysis of ATAC-seq data from seedling leaf tissue confirmed the lack of accessible chromatin at iUMRs (Fig. 5*A*). Both aUMRs and iUMRs exhibit altered profiles of many chromatin modifications relative to control regions both within the UMR and the flanking 1-kb regions. The most striking difference between aUMRs and iUMRs is observed for H3K4me1; aUMRs tend to have quite low levels of this modification and are depleted for this mark in flanking regions. In contrast, iUMRs show a strong enrichment for H3K4me1 (Fig. 5*A*). This histone covalent modification is intriguing because it is a characteristic mark of mammalian enhancers (54); in contrast, enhancers in plants so far notably lack a common chromatin signature (4, 12) such as the H3K4me1 mark in mammals. The majority of the other modifications examined exhibit similar trends for both iUMRs and aUMRs (Fig. 5*A*). For some modifications, such as H3K4me3, H3K27ac, K3K9ac, and H3K56ac, there are slightly stronger enrichments for the aUMRs (Fig. 5*A*). In other cases, such as H3K27me3, the profiles are similar but the iUMRs have stronger enrichments. Both aUMRs and iUMRs are deleted for H3K9me2, but the depletion is stronger within the UMR relative to flanking regions of aUMRs (Fig. 5*A*).

We proceeded to use several metrics developed by Ricci et al. (12) to investigate chromatin interactions and potential enhancer function for the iUMRs and aUMRs. We compared the proportion of iUMRs and aUMRs that overlap with HiC, H3K4me3-HiChIP, or H3K27me3-HiChIP loop edges (Fig. 5*B* and *SI Appendix*, Fig. S11). In each case, we compared these to an associated control set of randomized intergenic regions. The iUMRs show

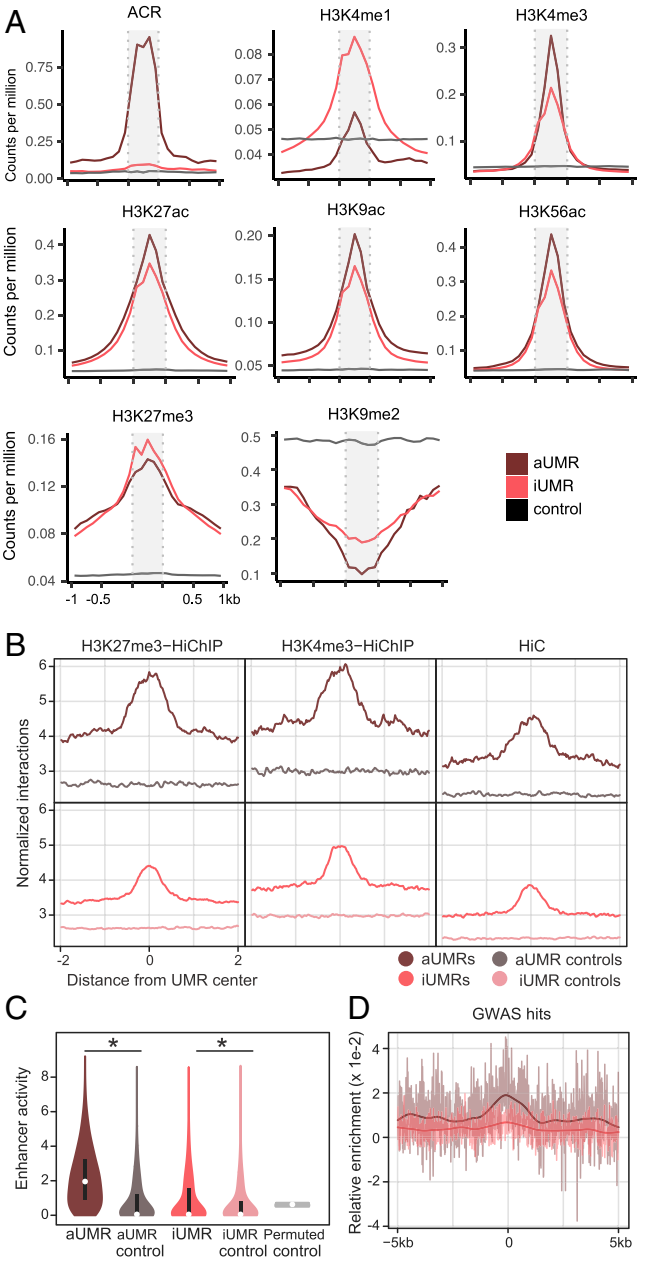

Fig. 5. The chromatin profile of gene-distal iUMRs. (*A*) The average enrichment of chromatin modifications over aUMR and iUMRs. Normalized read abundance in counts per million for ATAC-seq (ACRs) and ChIP-seq for histone modifications is averaged over UMRs and the 1 kb upstream and downstream. The UMR region is indicated by the shaded gray box. SE is overlaid in a lighter shade. (*B*) Metaplot averages of normalized interaction tags from H3K27me3 Hi-ChIP, H3K4me3 Hi-ChIP, and Hi-C in 4-kb windows centered on iUMRs, aUMRs, and their respective controls. (*C*) Distribution of enhancer activities (log2[RNA fragments per million/input fragments per million]) for aUMRs, iUMRs, their respective control regions, and averages from 10,000 Monte Carlo permutations of random intergenic regions (*$P <$ 5e-108). (*D*) Metaplot of relative enrichment of significant GWAS hits in 10-kb windows centered on iUMRs and aUMRs.

nearly the same level of enrichment as the aUMRs, suggesting that these regions are frequently making contacts with other regions and participate in chromatin looping. STARR-seq assays were performed by Ricci et al. (12) to assess the potential for ACRs to provide functional enhancer activity in maize leaf protoplasts. Given that many of the iUMRs are associated with genes that are

expressed in other tissues or ChIP-seq peaks for TFs that are not expressed in leaf tissue, we did not expect the same level of enrichment for enhancer activity in protoplasts from leaf tissue. While the iUMRs show substantially less enhancer activity based on leaf protoplast STARR-seq assays compared to aUMRs, we do still see significant enrichment relative to control regions (Wilcoxon rank-sum test P value <3.7e-108 and 10,000-permutations empirical P value <1e-4) of matched intergenic sites (Fig. 5*C*). We also assessed the frequency of GWAS hits at the iUMRs relative to aUMRs (Fig. 5*D*). While the aUMRs show significant increase for GWAS hits, there is less enrichment for iUMRs. Thus, aUMRs can more frequently be associated with regions linked to trait variation compared to iUMRs. This could be due to the fact that constitutively expressed genes (which will contain aUMRs) will be enriched for GWAS hits relative to tissue-specific expressed genes. Alternatively, some iUMRs may be nonfunctional, which would contribute to the lower GWAS enrichment. Overall, these analyses suggest that iUMRs and aUMRs have unique chromatin profiles relative to the other distal intergenic regions and that the iUMRs often participate in chromatin loops. However, these iUMRs are less often colocated with regions linked to trait variation.

The Utility of UMRs for Annotation and Discovery in Large Genomes. These analyses were initially focused on maize given the availability of other datasets that could be used to assess potential functions and roles of iUMRs. However, we predict that similar numbers of iUMRs and aUMRs would be identified in other cereals and grasses. We gathered DNA methylation and chromatin accessibility data for four other grasses (Dataset S1): barley (*Hordeum vulgare*) (55), sorghum (*Sorghum bicolor*) and *Brachypodium* (*Brachypodium distachyon*) (56), and rice (*Oryza sativa*) (57). These species vary substantially in genome size, with some species <500 Mb and others >4 Gb (Fig. 6*A*). The genome size that could be assessed for DNA methylation varied, with >1.5 Gb for maize and barley and <400 Mb for rice and *Brachypodium*. Despite these major differences in total genome size and the size of the genome for which DNA methylation could be profiled, we find roughly similar amounts of UMRs across all profiled species (Fig. 6*A* and Datasets S7–S10). This

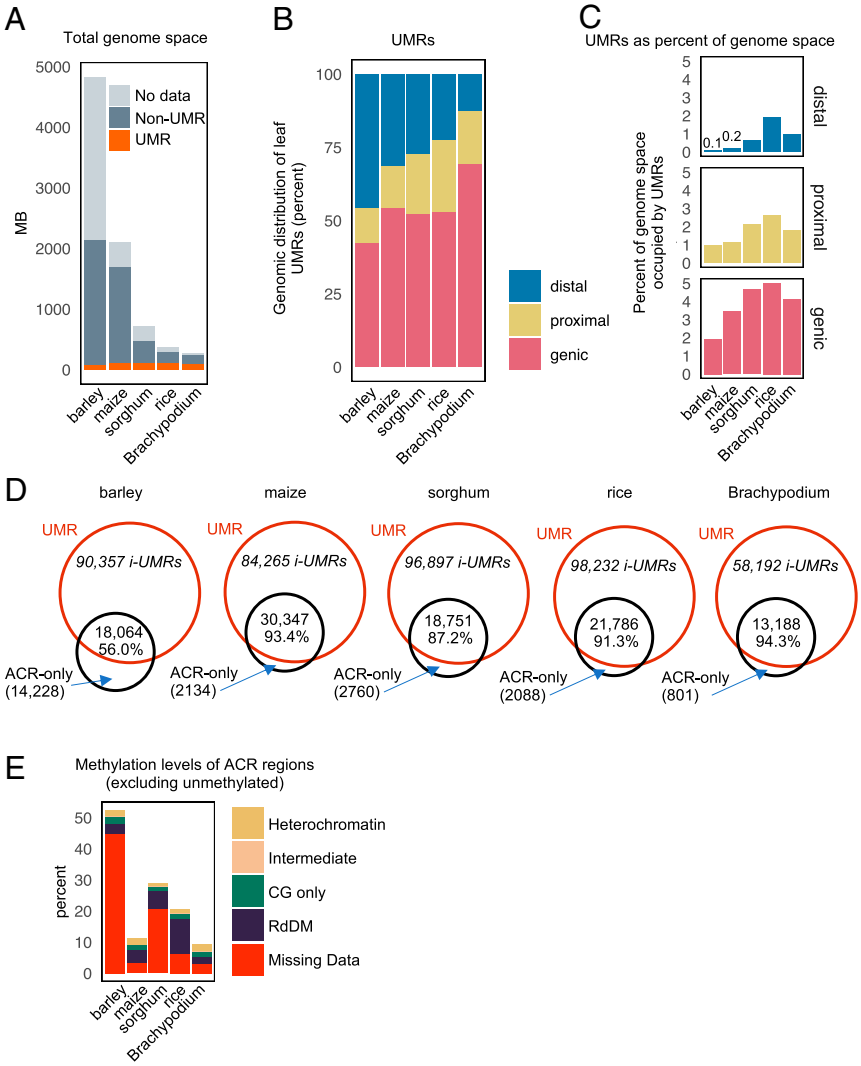

Fig. 6. Similarity between the absolute size and features of the unmethylated portion of cereal and grass genomes. (*A*) Proportion (MB) of UMRs in five grass genomes. (*B*) The genomic distribution of UMRs in each genome. (*C*) The proportion of genic, proximal, and distal regions of each genome comprised of UMRs. Larger genomes, such as maize and barley, have a much larger intergenic space, and hence intergenic/distal UMRs are a much smaller fraction. (*D*) Overlap between UMRs and ACRs in each species. Percentages refer to the percentage of ACRs that are unmethylated and captured by UMR profiling (not corrected for missing data). (*E*) The methylation profile (distribution of methylation domains) of ACR regions in each species, excluding unmethylated regions. For example, over 43.6% of the ACRs in barley overlap regions with missing methylation data, explaining the relative low overlap in *D*.

suggests that the total genome space of UMRs is relatively constant despite dramatic changes in overall genome size. This is consistent with the finding that the total accessible space (ACR) is similar in genomes of different sizes (4). The distribution of genic, proximal, and distal UMRs varies between species but is related to genome size (Fig. 6*B* and *SI Appendix*, Fig. S11). The large genomes have more examples of distal UMRs and relatively fewer proximal UMRs compared to small genomes. This likely reflects the higher gene density in smaller genomes and reduction in the amount of genome classified as distal intergenic. If we assess the proportion of all genic, proximal, and distal space in each genome that is classified as UMR, we find that the amount of genic space within UMRs is quite similar for all species (Fig. 6*C*). In contrast, the proportion of distal space that is classified as UMR is much lower in species with large genomes (Fig. 6*C*). This highlights the potential of DNA methylation to reveal the subset of potentially functional intergenic space, particularly in large genomes.

A comparison of the UMRs and ACRs for each species revealed substantial overlap, the same as was observed in maize (Fig. 6*D*). In most species, the vast majority of ACRs occur within UMRs. The proportion of overlap is the smallest for barley. However, when we assessed ACRs failing to overlap with UMRs, we found that, for all species, the vast majority of these represent either unmethylated tiles that did not meet the criteria for UMRs or had missing data (Fig. 6*E*). There are very few examples of ACRs in any of the species that are classified as having high levels of heterochromatic DNA methylation. The observation that some ACRs are not captured within the classified UMRs due to missing data for DNA methylation highlights the importance of deep-coverage methylation datasets for use in annotation of UMRs. The barley methylome dataset is only ~4.6× (Dataset S1), and this results in a substantial amount of genomic space that does not have sufficient sequencing depth for accurate classification of the DNA methylation state.

Discussion

The annotation of genomes remains a difficult problem, especially the discovery of putative regulatory elements. Documenting tissue- or cell-specific expression levels and chromatin states has been successfully applied to improve annotations using ENCODE-like approaches (58–61). However, generating comprehensive atlases of expression or chromatin in many cell types and conditions can be experimentally challenging and costly. Here, we suggest that identification of the unmethylated portions of crop genomes from a single tissue can help provide fairly complete catalogs of potential regulatory elements and expressed genes across many developmental stages. The advantage to this approach is that it can be performed on a single, easily harvested tissue type. Prior reports in both plants (29, 38) and animals (62, 63) have identified very large unmethylated regions, tens of kb in size, termed valleys or canyons. In contrast, we focused on UMRs marking putative regulatory regions, which are smaller in size, 1.1 kb on average with the majority less than 1 kb. With our higher-coverage datasets, for example, in maize and *Brachypodium*, around 94% of ACRs overlapped UMRs, and it is not clear if this overlap could be increased further. A small percentage of ATAC-seq peaks had some apparent level of methylation (either CG only, heterochromatin, or RdDM). This may in part be due to technical differences between WGBS and ATAC-seq data generation and analysis. In some cases, the specific boundaries of the ACR may be in the middle of a 100-bp tile used for classifying methylation, and a tile could include part of a region with high methylation and part with low methylation. As we noted in our comparison of species (Fig. 6), it is important to generate a relatively deep-coverage dataset of DNA methylation to maximize the amount of the genome that is confidently classified as methylated or unmethylated. Even with deep coverage, monitoring both UMRs and ACRs within highly repetitive regions remains challenging. In maize, we are only able to profile methylation levels for 76% of the genome with a relatively deep-coverage dataset and stringent coverage filter. By lowering the required coverage threshold to ×3, we increased coverage to 82% of the genome, which was sufficient to identify UMRs. Around 6% of the genome lacks a sufficient number of cytosines, and the remainder is too repetitive to allow unique mapping using short reads. We recommend ~10× average cytosine coverage per strand after quality filtering and alignment as a prudent target for analysis of unmethylated regions.

It is worth noting that this conceptual framework—using UMRs from a single tissue to develop a catalog of potential regulatory elements and expressed genes—relies upon the stability of CG and CHG methylation in different cell types. While CG methylation can be quite variable in different cell types for mammals (64), the CG and CHG methylation patterns are dramatically more stable in plants. There are examples of dynamic CHH methylation in different tissues in plants (37–42), and there is also evidence that some specific cell types undergo substantial changes in the methylation during reproduction (32–34). However, the bulk of the genome exhibits an often underappreciated consistency in the patterns of DNA methylation among different vegetative tissues (22–24).

One application of the UMR framework is to identify genes that have potential for expression. In maize and other crop genomes, it is difficult to discriminate transposon-derived gene fragments from true genes (1, 65). Annotation of genes requires a balancing between quality and comprehensiveness (65). In many cases, the desire to have a relatively complete set of putative genes results in numerous pseudogenes being included in gene annotations. Prior work has shown that applying machine learning to the patterns of context-specific levels of DNA methylation can classify genes with potential for expression (3). Here we show that the majority of genes that are detected as expressed (in a panel of >200 samples) contain an unmethylated region close to their annotated TSS. The presence of a UMR within the promoter of a putative gene can be used to indicate the potential for expression of the gene. Several other studies have implemented different approaches to use DNA methylation data to augment gene annotations (2, 3).

UMRs can also be used to discover potential regulatory elements. UMRs are enriched for TF-binding sites based on DAP-seq or ChIP-seq datasets. It is especially noteworthy that this enrichment for TF binding is observed even though the UMRs are defined in a tissue type where most of these TFs are very low-expressed or silent. There are also many examples of tissue-specific ACRs that are equally unmethylated across multiple tissues. This suggests that, in plant genomes, the majority of regions with potential to be TF binding sites in some tissue, developmental stage, or environment are stably unmethylated. While the application of chromatin-accessibility assays or TF-binding assays in a specific tissue can provide a very high-quality representation of the active regulatory elements in that tissue, here we show that it is possible to rapidly develop a far more complete set of potential regulatory elements through the analysis of DNA methylation profiles from a single tissue.

The utility of a methylation filter to focus on unmethylated regions may be variable across different species. In species with relatively small genomes, for example <500 Mb, and limited intergenic space, the filtering power of focusing on unmethylated regions is likely diminished. In a species like *Arabidopsis thaliana*, most genes are arranged in close proximity to other genes, and the amount of the genome that could be masked as methylated is relatively small (66). In contrast, in species such as barley or maize with large genomes and low gene density, the ability to focus on the unmethylated portions of the genome provides a powerful framework to distill an enormous genome down to a

relatively small fraction of genomic space: highly enriched regions valuable for regulation or manipulation of plant traits.

Methods

WGBS Data. Whole-genome bisulfite sequencing samples are listed in Dataset S1. For deep-coverage maize leaf data generated in this study, DNA was extracted from leaves of 2-wk-old V2 glasshouse-grown maize B73 plants using the DNeasy Plant Mini kit (Qiagen). Six biological replicates were sampled for sequencing and later combined into a single data set. Details of library preparation are provided in the *SI Appendix*. WGBS data generated in this study are available under accession code GSE150929. Additional maize seedling leaf (SRR8740851) and seedling root (SRR8740850) samples were described previously (46) and downloaded from SRA PRJNA527657. WGBS data for other species were downloaded from SRA, including barley leaf (*H. vulgare* accession Morex) SRR5124893 (55), sorghum leaf (*S. bicolor* accession code BTx623) SRR3286309 (56), rice leaf (*O. sativa* accession Nipponbare) SRX205364 (57), and *Brachypodium* leaf (*B. distachyon* accession code Bd21) SRX1656912 (56). Sequencing reads were processed as detailed in the *SI Appendix* and aligned with bsmap v2.74 (67) to the respective genomes. WGBS pipelines are available on GitHub (https://github.com/pedrocrisp/springerlab_methylation).

Identification of Unmethylated Regions. To identify unmethylated regions, each 100-bp tile of the genome was classified into one of six domains or types, including "missing data" (including "no data" and "no sites"), "RdDM," "heterochromatin," "CG-only," "unmethylated," or "intermediate," according to the hierarchy in Dataset S2 (also see *SI Appendix*, Fig. S1). Briefly, tiles were classified as missing data if tiles had less than two cytosines in the relevant context or if there was less than 5× coverage for maize or less than 3× coverage when comparing the different grass species (owing to lower coverage in some species); RdDM if CHH methylation was greater than 15%; heterochromatin if CG and CHG methylation was 40% or greater; CG-only if CG methylation was greater than 40%; unmethylated if CG, CHG, and CHH were less than 10%; and intermediate if methylation was 10% or greater but less than 40%. A breakdown of the proportion of tiles for each domain is detailed in Dataset S1. Following tile classification, adjacent unmethylated tiles (UMTs) were merged. To capture and combine any unmethylated regions that were fragmented by a short interval of missing data (low coverage or no sites), any merged UMT regions that were separated by missing data were also merged so long as the resulting merged region consisted of no more than 33% missing data. Regions less than 300 bp were removed (*SI Appendix*, Fig. S1 *C* and *D*) because they were depleted of accessible regions, with 99.5% inaccessible. Some very short UMRs could be functional, but thus far, we have limited evidence for functionality. The remaining regions were defined as unmethylated regions (UMRs). We classified UMRs within 2,000 bp of the annotated TSS as gene-proximal or -distal if greater than 2,000 bp as described in Ricci et al. (12); however, where a UMR overlapped both the gene locus and the gene-proximal region, it was hierarchically classified as proximal.

ATAC-Seq Data. For maize root ATAC-seq data generated in this study, *Z. mays* (accession B73) was grown in soil for around 6 d at 25 °C under 16 h light–8 h dark. The staple roots were harvested and were used for experiments. ATAC-seq was performed as described previously (68) and described in detail in the *SI Appendix*. ATAC-seq raw reads were aligned as described before (4, 12), and a summary is outlined in the *SI Appendix*. ATAC-seq data generated in this study are available under GEO accession code GSE152046. ATAC-seq data for maize leaf and ear were as described in Ricci et al. (12), and the coordinates of accessible chromatin regions were downloaded from the GEO archive, accession code GSE120304. ATAC-seq data for barley (*H. vulgare* accession Morex), sorghum (*S. bicolor* accession code BTx623), rice (*O. sativa* accession Nipponbare), and *Brachypodium* (*B. distachyon* accession code Bd21) are as described in Lu et al. (4), and the coordinates of accessible chromatin regions were downloaded from the GEO archive, accession code GSE128434.

Expression Data. RNA-seq expression data for maize leaf and ear (12) as well as 247 samples for other maize tissues (69–78) were downloaded from NCBI Sequence Read Archive and processed as described in Zhou et al. (79) and detailed in the *SI Appendix*. Synteny classifications (i.e., syntenic and non-syntenic) and assignment to maize subgenomes were obtained from a previous study based on pairwise whole-genome alignment between maize and sorghum, downloaded from Figshare (Schnable et al., 2019; DOI 10.6084/m9.figshare.7926674.v1) (47). The eFP browser expression data were downloaded from bar (80) hosted on Maize GDB incorporating the maize expression datasets (76, 81). The expression of fea4, kn1, o2, ra1, and p1 in leaf was evaluated considering the samples: pooled leaves V1, topmost leaf V3, tip of stage 2 leaf V5, base of stage 2 leaf V5, tip of stage 2 leaf V7, base of stage 2 leaf V7, immature leaf V9, 13th leaf V9, 11th leaf V9, 8th leaf V9, 13th leaf VT, and 13th leaf R2.

Transcription Factor DAP-Seq and ChIP-Seq. DAP-seq profiles for 32 maize TFs (12, 48) (Dataset S6) were downloaded from the SRA, as were ChIP-seq for five TFs: KNOTTED1 (KN1) (49), RAMOSA1 (RA1) (50), fasciated ear4 (FEA4) (51), Opaque2 (O2) (53), and PERICARP COLOR 1 (p1) (52). Sequencing data were downloaded from NCBI using the SRA Toolkit and processed using the nf-core ChIP-seq pipeline (82) (detailed description provided in the *SI Appendix*). Randomized control regions were generated using bedtools shuffle, and comparison with UMRs evaluated with 1,000 permutations using regioneR. Pipeline scripts, QC files, and peak calling results and annotation are available at GitHub (https://github.com/orionzhou/chipseq).

Analysis of Histone Modifications. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) data for H3K3me1, H3K3me3, H3K27ac, H3K9ac, H3K56ac, H3K27me3, and H3K9me2 chromatin modification as reported by Ricci et al. (12) were downloaded from the GEO database, accession GSE120304, and processed as described in the *SI Appendix*.

Analysis of Hi-C, Hi-ChIP, STARR-Seq, and GWAS Data. Raw and processed Hi-C, Hi-ChIP, and STARR-seq data were acquired from Ricci et al. (12) and analyzed as detailed in the *SI Appendix*. Genomic positions of significant GWAS hits were obtained from Wallace et al. (83), and GWAS hits and SNPs in the NAM founder lines were downloaded from cyverse. Relative GWAS enrichment per bin was estimated as detailed in the *SI Appendix*.

Data Availability. ATAC-seq (fastq) data have been deposited in GEO (accession no. GSE152046).

ACKNOWLEDGMENTS. This work was funded by grants NSF IOS-1934384 to N.M.S., NSF IOS-1802848 to N.M.S., and NSF IOS-1856627 to R.J.S. P.A.C. is the recipient of an Australian Research Council Discovery Early Career Award (project number DE200101748) funded by the Australian Government. J.M.N. was supported by a Hatch grant from the Minnesota Agricultural Experiment Station (MIN 71-068). A.M. was supported by an NSF Postdoctoral Fellowship in Biology (NSF DBI-1905869). R.J.S. is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts. The Minnesota Supercomputing Institute at the University of Minnesota provided computational resources that contributed to this research.

1. J. C. Schnable, Genes and gene models, an important distinction. *New Phytol.*, 10.1111/nph.16011 (2019).
2. A. Olson *et al.*, Expanding and vetting sorghum bicolor gene annotations through transcriptome and methylome sequencing. *Plant Genome* **7**, 1–20 (2014).
3. R. C. Sartor, J. Noshay, N. M. Springer, S. P. Briggs, Identification of the expressome by machine learning on omics data. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 18119–18125 (2019).
4. Z. Lu *et al.*, The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* **5**, 1250–1259 (2019).
5. S. Salvi *et al.*, Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 11376–11381 (2007).
6. M. Louwers *et al.*, Tissue- and expression level-specific chromatin looping at maize b1 epialleles. *Plant Cell* **21**, 832–842 (2009).
7. G. Xu *et al.*, Complex genetic architecture underlies maize tassel domestication. *New Phytol.* **214**, 852–864 (2017).
8. A. Studer, Q. Zhao, J. Ross-Ibarra, J. Doebley, Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).
9. C. Huang *et al.*, *ZmCCT9* enhances maize adaptation to higher latitudes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E334–E341 (2018).
10. L. Liu *et al.*, KRN4 controls quantitative variation in maize Kernel row number. *PLoS Genet.* **11**, e1005670 (2015).
11. L. Zheng *et al.*, Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in Zea mays. *J. Exp. Bot.* **66**, 3917–3930 (2015).
12. W. A. Ricci *et al.*, Widespread long-range cis-regulatory elements in the maize genome. *Nat. Plants* **5**, 1237–1249 (2019).
13. E. Rodgers-Melnick, D. L. Vera, H. W. Bass, E. S. Buckler, Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3177–E3184 (2016).
14. K. A. Maher *et al.*, Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell* **30**, 15–36 (2018).
15. A. M. Sullivan *et al.*, Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Rep.* **8**, 2015–2030 (2014).

16. B. Weber, J. Zicola, R. Oka, M. Stam, Plant enhancers: A call for discovery. *Trends Plant Sci.* **21**, 974–987 (2016).
17. R. Oka *et al*., Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* **18**, 137 (2017).
18. Y. Peng *et al*., Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nat. Commun.* **10**, 2632 (2019).
19. E. Li *et al*., Long-range interactions between proximal and distal regulatory regions in maize. *Nat. Commun.* **10**, 2633 (2019).
20. Z. Lu, W. A. Ricci, R. J. Schmitz, X. Zhang, Identification of cis-regulatory elements by chromatin structure. *Curr. Opin. Plant Biol.* **42**, 90–94 (2018).
21. S. J. Burgess *et al*., Genome-wide transcription factor binding in leaves from C_3 and C_4 grasses. *Plant Cell* **31**, 2297–2314 (2019).
22. S. R. Eichten, M. W. Vaughn, P. J. Hermanson, N. M. Springer, Variation in DNA methylation patterns is more common among maize inbreds than among tissues. *Plant Genome* **6**, 1–10 (2013).
23. R. J. Schmitz *et al*., Patterns of population epigenomic diversity. *Nature* **495**, 193–198 (2013).
24. T. Kawakatsu *et al*., Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nat. Plants* **2**, 16058 (2016).
25. D. R. Ganguly, P. A. Crisp, S. R. Eichten, B. J. Pogson, Maintenance of pre-existing DNA methylation states through recurring excess-light stress. *Plant Cell Environ.* **41**, 1657–1672 (2018).
26. P. A. Crisp *et al*., Rapid recovery gene downregulation during excess-light stress and recovery in Arabidopsis. *Plant Cell* **29**, 1836–1863 (2017).
27. D. R. Ganguly, P. A. Crisp, S. R. Eichten, B. J. Pogson, The Arabidopsis DNA methylome is stable under transgenerational drought stress. *Plant Physiol.* **175**, 1893–1912 (2017).
28. S. R. Eichten, N. M. Springer, Minimal evidence for consistent changes in maize DNA methylation patterns following environmental stress. *Front. Plant Sci.* **6**, 308 (2015).
29. M. Chen *et al*., Seed genome hypomethylated regions are enriched in transcription factor genes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E8315–E8322 (2018).
30. T.-F. Hsieh *et al*., Genome-wide demethylation of Arabidopsis endosperm. *Science* **324**, 1451–1454 (2009).
31. A. Zemach *et al*., Local DNA hypomethylation activates genes in rice endosperm. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18729–18734 (2010).
32. C. A. Ibarra *et al*., Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* **337**, 1360–1364 (2012).
33. J. P. Calarco *et al*., Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* **151**, 194–205 (2012).
34. R. K. Slotkin *et al*., Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**, 461–472 (2009).
35. J. Walker *et al*., Sexual-lineage-specific DNA methylation regulates meiosis in Arabidopsis. *Nat. Genet.* **50**, 130–137 (2018).
36. K. Park *et al*., DNA demethylation is initiated in the central cells of Arabidopsis and rice. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 15138–15143 (2016).
37. Y. C. An *et al*., Dynamic changes of genome-wide DNA methylation during soybean seed development. *Sci. Rep.* **7**, 12263 (2017).
38. J.-Y. Lin *et al*., Similarity between soybean and *Arabidopsis* seed methylomes and loss of non-CG methylation does not affect seed development. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9730–E9739 (2017).
39. R. Narsai *et al*., Extensive transcriptomic and epigenomic remodelling occurs during Arabidopsis thaliana germination. *Genome Biol.* **18**, 172 (2017).
40. T. Kawakatsu, J. R. Nery, R. Castanon, J. R. Ecker, Dynamic DNA methylation reconfiguration during seed development and germination. *Genome Biol.* **18**, 171 (2017).
41. D. Bouyer *et al*., DNA methylation dynamics during early plant life. *Genome Biol.* **18**, 179 (2017).
42. L. Ji *et al*., Genome-wide reinforcement of DNA methylation occurs during somatic embryogenesis in soybean. *Plant Cell* **31**, 2315–2331 (2019).
43. N. M. Springer, D. Lisch, Q. Li, Creating order from chaos: Epigenome dynamics in plants with complex genomes. *Plant Cell* **28**, 314–325 (2016).
44. N. M. Springer, R. J. Schmitz, Exploiting induced and natural epigenetic variation for crop improvement. *Nat. Rev. Genet.* **18**, 563–575 (2017).
45. J. M. Noshay, P. A. Crisp, N. M. Springer, "The maize methylome" in *The Maize Genome*, J. Bennetzen, S. Flint-Garcia, C. Hirsch, R. Tuberosa, Eds. (Springer International Publishing, 2018), pp. 81–96.
46. J. M. Noshay *et al*., Monitoring the interplay between transposable element families and DNA methylation in maize. *PLoS Genet.* **15**, e1008291 (2019).
47. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4069–4074 (2011).
48. M. Galli *et al*., The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.* **9**, 4526 (2018).
49. N. Bolduc *et al*., Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes Dev.* **26**, 1685–1690 (2012).
50. A. L. Eveland *et al*., Regulatory modules controlling maize inflorescence architecture. *Genome Res.* **24**, 431–443 (2014).
51. M. Pautler *et al*., FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. *Plant Cell* **27**, 104–120 (2015).
52. K. Morohashi *et al*., A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. *Plant Cell* **24**, 2745–2764 (2012). Correction in: *Plant Cell* **24**, 3853 (2012).
53. C. Li *et al*., Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell* **27**, 532–545 (2015).
54. N. D. Heintzman *et al*., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
55. T. Wicker *et al*., The repetitive landscape of the 5100 Mbp barley genome. *Mob. DNA* **8**, 22 (2017).
56. C. E. Niederhuth *et al*., Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).
57. H. Stroud *et al*., Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife* **2**, e00354 (2013).
58. S. E. Celniker *et al*.; modENCODE Consortium, Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
59. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
60. A. K. Lane, C. E. Niederhuth, L. Ji, R. J. Schmitz, pENCODE: a plant encyclopedia of DNA elements. *Annu. Rev. Genet.* **48**, 49–70 (2014).
61. C. A. Davis *et al*., The encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
62. X. Zhang *et al*., Large DNA methylation nadirs anchor chromatin loops maintaining hematopoietic stem cell identity. *Mol. Cell* **78**, 506–521.e6 (2020).
63. W. Xie *et al*., Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
64. C. Luo, P. Hajkova, J. R. Ecker, Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).
65. M. Van Bel, F. Bucchini, K. Vandepoele, Gene space completeness in complex plant genomes. *Curr. Opin. Plant Biol.* **48**, 9–17 (2019).
66. P. A. Crisp, J. M. Noshay, S. N. Anderson, N. M. Springer, Opportunities to use DNA methylation to distil functional elements in large crop genomes. *Mol. Plant* **12**, 282–284 (2019).
67. Y. Xi, W. Li, BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
68. Z. Lu, B. T. Hofmeister, C. Vollmers, R. M. DuBois, R. J. Schmitz, Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res.* **45**, e41 (2017).
69. P. Li *et al*., The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.* **42**, 1060–1067 (2010).
70. R. M. Davidson *et al*., Utility of RNA sequencing for analysis of maize reproductive transcriptomes. *Plant Genome* **4**, 191–203 (2011).
71. W.-Y. Liu *et al*., Anatomical and transcriptional dynamics of maize embryonic leaves during seed germination. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 3979–3984 (2013).
72. C.-P. Yu *et al*., Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2477–E2486 (2015).
73. G. Li *et al*., Temporal patterns of gene expression in developing maize endosperm identified through transcriptome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7582–7587 (2014).
74. A. M. Chettoor *et al*., Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. *Genome Biol.* **15**, 414 (2014).
75. J. Chen *et al*., Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol.* **166**, 252–264 (2014).
76. S. C. Stelpflug *et al*., An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* **9**, 1–16 (2016).
77. J. W. Walley *et al*., Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).
78. P. Zhou, C. N. Hirsch, S. P. Briggs, N. M. Springer, Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. *Mol. Plant* **12**, 410–425 (2019).
79. P. Zhou *et al*., Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. *Plant Cell* **32**, 1377–1396 (2020).
80. D. Winter *et al*., An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One* **2**, e718 (2007).
81. G. M. Hoopes *et al*., An updated gene atlas for maize reveals organ-specific and stress-induced genes. *Plant J.* **97**, 1154–1167 (2019).
82. P. A. Ewels *et al*., The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).
83. J. G. Wallace *et al*., Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.* **10**, e1004845 (2014).