Identification of the expressome by machine learning on omics data

Ryan C. Sartor, Jaclyn Noshay, Nathan M. Springer, and Steven P. Briggs

*Division of Biology, University of California San Diego, La Jolla, CA 92093; and Department of Plant Biology, University of Minnesota, St. Paul, MN 55108

Contributed by Steven P. Briggs, July 11, 2019 (sent for review August 14, 2018; reviewed by James A. Birchler and Virginia Walbot)

Accurate annotation of plant genomes remains complex due to the presence of many pseudogenes arising from whole-genome duplication-generated redundancy or the capture and movement of gene fragments by transposable elements. Machine learning on genome-wide epigenetic markers, informed by transcriptomic and proteomic training data, could be used to improve annotations through classification of all putative protein-coding genes as either constitutively silent or able to be expressed. Expressed genes were subclassified as able to express both mRNAs and proteins or only RNAs, and CG gene body methylation was associated only with the former subclass. More than 60,000 protein-coding genes have been annotated in the reference genome of maize inbred B73. About two-thirds of these genes are transcribed and are designated the filtered gene set (FGS). Classification of genes by our trained random forest algorithm was accurate and relied only on histone modifications or DNA methylation patterns within the gene body; promoter methylation was unimportant. Other inbred lines are known to transcribe significantly different sets of genes, indicating that the FGS is specific to B73. We accurately classified the sets of transcribed genes in additional inbred lines, arising from inbred-specific DNA methylation patterns. This approach highlights the potential of using chromatin information to improve annotations of functional genes.

In maize, proteins are observed only from a subset of transcribed genes: 87% of genes observed to make proteins have syntenic orthologs in sorghum, even though syntenic genes account for only 23% of transcribed genes (1). This observation explains why nearly all genes with known functions are syntenic (2), and it raises a new question: How can the cell distinguish between syntenic and species-specific genes such that both are transcribed but only the former expresses proteins?

To begin to answer this question, we present a machine-learning-based approach that provides genome-wide classifications of annotated protein-coding genes as expressible or constitutively silent based on patterns of DNA methylation or histone modifications. The classifiers are additionally able to distinguish between genes that can express proteins and genes that can only express RNAs. Our findings address a long-standing challenge in genome biology to discover the expressible gene set (EGS) which comprises all protein-coding genes with the potential to be expressed in an individual. Efforts to identify the EGS have been based on surveys of expression and on comparative genomics. While these criteria have been useful, surveys only sample some of the conditions, cell types, and genetic diversity that affect gene expression, leaving some functional genes without evidence for expression. New methods are needed to identify the EGS. It is equally important to identify the silent gene set because these genes may or may not be expressed in other individuals with different genetic backgrounds and epigenetic marks. Our EGS for protein expression was contained within the larger EGS for mRNA expression. Collectively, the EGS from all individuals constitutes the expressome: all protein-coding genes in a species with the potential to be expressed as proteins or only as RNAs.

Most researchers study the predicted genes that are derived from whole-genome annotations. These annotation approaches can be complicated by the presence of sequences with homology to protein coding genes that may not be functional genes. These false gene annotations can result from silenced paralogs following either whole-genome duplications or tandem duplications, or they may arise from capture of gene fragments by transposable elements. Here we show that the analysis of DNA methylation patterns can help identify annotated genes that are not likely to be expressed or can be expressed only as RNAs.

We found that the most significant genome features used by our random forest classifiers are well-known patterns of DNA methylation and histone modification, indicating that these patterns may play roles in establishing permissions for gene expression. Genes that expressed both mRNAs and proteins had DNA methylation patterns that were distinct from genes that only expressed RNAs. Silent gene patterns differed from both expressible classes. Our models matched or outperformed expert curation for the ability to differentiate between expressed and silent genes. Extension of our method to other inbred lines with differential DNA methylation patterns demonstrated that the EGS differs between inbreds by thousands of genes that correlate with variations in the epigenome. A small but significant difference in the EGS was observed between organs. Discovery of the EGS for individuals and of the expressome for a species will contribute to understanding and fully utilizing the genetic potential of organisms. Our characterization of the EGS for the widely used maize inbred, B73, provides a first step toward discovery of the expressome for the world’s most valuable crop.

Significance

Our new method uses only epigenomic patterns to classify the expression potential of annotated genes and identifies pseudogenes that are difficult to classify based solely on sequence. Genes were divided into those with protein expression, those with mRNA expression, and those that are silent. A large fraction of annotated genes are constitutively silent in one lineage but can be transcribed in others. We refer to the species-wide set of transcribed genes as the expressome and show that it is much larger than the expressible gene set in any individual. Additionally, we find that DNA methylation patterns within the gene body can differentiate between genes that express proteins and genes that express only RNAs.

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

Reviews: J.A.B., University of Missouri; and V.W., Stanford University.

Conflict of interest statement: N.M.S. and V.W. are coauthors on a 2016 Review article. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1 To whom correspondence may be addressed. Email: sbriggs@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1813645116/-/DCSupplemental.
**Results and Discussion**

**Genic DNA Methylation Can Classify Expression Potential.** Syntenic genes are hypomethylated relative to their nonsyntenic counterparts, suggesting that epigenomic features may enable robust gene classifications (3, 4). To further explore the relationship between DNA methylation and gene expression, we used the random forest algorithm (5) to build classifiers for all genes of the maize inbred line B73. Two classifiers were built based solely on genic DNA methylation features. Both used a combination of proteome and transcriptome data for training from 23 different tissues or times of development (1). For the expressible protein classifier (EPC), the silent class consisted of annotated genes with no observed mRNAs or proteins (NR_NP). The expressible class consisted of genes with high levels of mRNAs \[\text{fragments per kilobase per million reads (FPKM)} > 1\] defined by (1) and observed proteins (HR_OP). The training classes of the second classifier (expressible mRNA classifier, ERC) were defined using all genes with no detectable mRNAs (NR) vs. all genes with high mRNA levels (HR), and it did not use protein data (Fig. 1).

Several DNA methylation features were tested. The importance of features was determined using the mean decrease in accuracy upon random permutation of each individual variable (5). Three methylation sequence contexts (CHG, CG, and CHH) were quantified separately and summarized within gene regions (Fig. 1A), including the promoter (2 kilobases upstream of the transcription start site [TSS], split into 4 bins), the TSS, 5′ UTR, 3′ UTR, introns, exons, and a summed value encompassing the gene model. The summarized random forest feature importance is shown for each methylation context and genomic region (SI Appendix, Fig. S1). Based on these scores, multiple features with low importance were deleted from the models. Features retained were CHG and CG methylation in exons and introns, plus the aggregation of all retained features (labeled “Gene”).

![Genomic Regions Diagram](image)

**Fig. 1.** Overview of model features and training set definitions. (A) The various genomic regions where DNA methylation levels were quantified and used as features for classification. Features with gray labels were discarded after initial testing. Each gene was also split into 5 equivalent regions, called bins, and features were quantified separately in each bin. (B) The distribution of detected mRNA abundance is bimodal. The 2 mRNA populations can be roughly separated using an FPKM of 1. Here the nondetected mRNA (No mRNA) is represented as a separate population and given an artificial value of −12. Each population can be further refined into observed vs. nonobserved protein (No Protein) to yield 6 different groups of genes indicated by the different colors. LR_OP refers to all annotated genes that were observed to express low levels of mRNAs and detectable levels of proteins. (C) Three separate random forest models were built. Colored blocks correspond to the gene sets (from B) used for each training class. Blocks on the left indicate the positive (true) training instances vs. blocks on the right that indicate the negative (false) training instances. Numbers in parentheses indicate the number of genes in each training class.
Fig. 2. Results for random forest models. (A) Receiver operating characteristic (ROC) curves showing classification accuracy of the EPC, ERC, and PFI models. (B and C) Binned scatterplot showing prediction accuracy for quantitative abundance models considering only genes with observed expression for mRNA abundance (B) and protein abundance (C). (D–F) Signed feature importance measures for 3 different models. The values reflect the random forest “mean decrease in accuracy” measure of feature importance. The sign is based on the relationship of the feature values to the training class assignments. Positive values indicate a positive correlation between the feature and either protein observation (EPC and PFI) or high mRNA (ERC).

**Gene Body (4, 6, 7),** we divided each gene into 5 equal-proportion bins; methylation features were summarized separately for each bin (Fig. L4 and SI Appendix, Figs. S2 and S3). Classification accuracies were determined using random forest out-of-bag cross-validation on the training set genes (Fig. 2A and SI Appendix, Fig. S4A). Both classifiers had high accuracy with areas under the curve (AUCs) of 0.94 or higher for receiver operating characteristic (ROC) curves and precision vs. recall (PR) curves. The EPC achieved a near-perfect AUC of 0.99.

**CHG and CG Methylation Are Negatively Associated with Expression.**

To determine whether silent genes were associated with high or low methylation levels, each feature was given a positive or negative sign (Fig. 2D and E and SI Appendix, Materials and Methods and Fig. S5). The structures of the EPC and ERC models were very similar based on feature importance (Fig. 2D and E and SI Appendix, Fig. S4D). Methylation of CHG and CG at the 5′ ends of genes (bin 1) was most strongly associated with silent genes; CHG methylation at the 3′ ends of genes (bin 5) was also significant. In vitro methylation of CG sites in the 5′ region of the gene is a potent inhibitor of transgene expression (8).

**Gene Classification Is Not Associated with Patterns or Levels of Expression.**

To test whether genomic DNA methylation is associated with the patterns or levels of expression, random forest models were run using the same methylation training data but replacing the binary class vector with quantitative mRNA and protein abundance for the ERC and EPC, respectively. This produced the protein expression-level predictor and the mRNA expression-level predictor; both failed to accurately predict expression levels. When examining the full set of predictions (SI Appendix, Fig. S4D and C), we observed good R² values but observed low R² values when considering only genes with detectable expression (Fig. 2B and C). Therefore, the good R² values observed on the full set of predictions may be mostly due to the ability of the model to discriminate between observed and nonobserved products of expression.

**Genome-Wide Classifications.** The ERC and EPC were used to reclassify all protein-coding genes based solely on DNA methylation patterns, including the 98,296 members of the working gene set for which we had methylation coverage. The ERC classified 41,056 genes as able to express mRNAs, but only 32,979 genes were classified by the EPC as able to express proteins; 55% of the ERC expressible genes were absent from the training set (SI Appendix, Fig. S6A). This highlights the power of classification models to learn from a high-confidence subset of genes and then provide accurate genome-wide classifications. Comparison of results from the ERC and EPC identified 2 groups of genes that are expressed as RNAs only (8,078) or as mRNAs plus proteins (32,979) (Dataset S1).

We compared our classifications to the most recent (RefGen version 4 [v4]) and the previous (RefGen v2) curated classifications. Maize RefGen v2 was the last version where a full gene set was annotated (5a working gene set [WGS]), yielding over 110,000 gene models at distinct loci. The maize filtered gene set (FGS) is a subset of high-confidence protein-coding genes from the RefGen v2 WGS (see SI Appendix, Materials and Methods for description). For the newest assembly, RefGen v4 (9), only a filtered gene set has been annotated. We cross-referenced the RefGen v4 FGS to the RefGen v2 accessions (Dataset S1) so that our classifications could be compared to both versions of the maize
genome. Only ~77% of the RefGen v4 FGS can be converted to RefGen v2 accessions, which constrained our ability to make comparisons.

Our ERC differs from the curated FGS by 17,684 genes and 18,019 genes for RefGen v2 and RefGen v4, respectively (SI Appendix, Fig. S7A). We classified the remaining genes (57,239) as silent in B73. However, epialleles of silent genes can be transcribed in other genetic backgrounds as described below.

The EPC and ERC classified as silent 33 and 23% of the RefGen v2 FGS and 16 and 12% of the RefGen v4 FGS (SI Appendix, Fig. S7A); these proportions rose to 66 and 58% for all potential genes (RefGen v2 WGS). The Maize Genetics and Genomics Database (MaizeGDB) curation project (10) annotates a biotype to each gene model. The biotype can be used to filter out all likely transposable elements (TEs) and pseudogenes to yield 63,331 probable protein-coding genes in the WGS. Of these, 60,295 have coverage in our DNA methylation data. Using the higher-confidence EPC classifier, 48% were classified as silent (SI Appendix, Fig. S8), and nearly all of the TEs (97%) and pseudogenes (94%) in the WGS were classified as silent.

The Accuracy of Random Forest Models Matches or Exceeds That of Expert Curation. We compared the abilities of the curated RefGen v2 and v4 filtered gene sets and the EPC and ERC classifiers to identify the set of expressible genes in B73. ROC and PR curves were created to evaluate each set (SI Appendix, Fig. S7 F and G). Each set represents a 2-category classification (expressible or silent). The ROC curve (SI Appendix, Fig. S7F) shows that the EPC classifier achieves the highest accuracy (solid yellow line). The RefGen v2 set achieves a similar true-positive rate but with a higher false-positive rate, meaning that the RefGen v2 FGS, being the largest set, includes nearly all observed proteins but has more false positives. Precision (SI Appendix, Fig. S7G) represents the proportion of the corresponding set that is correctly called and should be insensitive to incomplete data assuming this subset is random. Looking at precision, we see the EPC and ERC (solid lines) outperform the RefGen v2 FGS (dashed lines). The RefGen v4 FGS performed well, with higher precision than the ERC for expressible mRNAs, but the EPC is still the top performer, indicating that the addition of protein data substantially improves classifications. The observed graphs of bimodal mRNA abundance could be more or less reconstructed from the 4 classified gene sets by plotting the average mRNA abundance for each silent and expressed gene (SI Appendix, Fig. S7 B–E). This indicates that DNA methylation patterns are sufficient to explain the observed bimodal component in the distributions of gene expression.

Hypermethylation of Transposable Elements Does Not Affect the Prediction Accuracy of Protein Coding Genes. The correlation between DNA methylation and gene expression was established first in studies of the maize autonomous TEs Ac, Spm, and MuDR. Using combinations of restriction enzymes, investigators found that the methylation of TEs was associated with their ability to transpose and to cause transposition of additional members of their TE family (11). TE methylation was negatively associated with mRNA and protein expression and with cycling between active and inactive states. High levels of CHG and CG methylation repress expression of TEs and repetitive elements (4, 7, 12, 13). Subsequent work has shown that plant TEs are silenced by RNA-dependent DNA methylation in the CHH context (14). Because these elements are so abundant in the genome, many gene models in the WGS are TEs that have escaped sequence masking. Of the 110,028 RefGen v2 gene models, 29,082 have been categorized as likely TEs. In addition, we identified 7,612 gene models that have a high basic local alignment search tool (BLAST) hit to one or more reference TE sequences in the maize TE database (15) and may be protein-coding genes with TEs inserted into their gene body. To determine the extent to which these previously characterized, highly methylated elements are affecting our classifiers and conclusions, we rebuilt all of the classification models after filtering out all 36,694 TEs and TE-containing gene models. The new classifier is nearly identical to the original both in classification accuracy (SI Appendix, Fig. S9 A and B) and in feature importance (SI Appendix, Fig. S9 C–E). We have left the TEs in the final models because our goal is to examine the relationship between genic methylation and expression potential. A subset of these 36,694 TEs and TE-containing genes was observed as proteins (2,423) or as highly expressed RNAs (5,065).

Inbred-Specific Expressible Gene Sets. To determine whether inbred-specific DNA methylation is associated with an inbred-specific EGS, the ERC was remade using data from multiple maize inbreds. The third leaf from the genetically diverse inbreds Mo17, CML322, Oh43, Tx303, and B73 was used to produce DNA methylation data (16) and RNA sequencing (RNA-seq) data (17). The methylation data were processed by quantifying weighted methylation levels for consecutive 100 base pair (bp) tiles along each chromosome. A new classifier, ERC-2, was constructed using the same class definitions as the ERC (summarized expression from many tissues). The model was trained using these class definitions plus the 100 bp tile DNA methylation data from the third leaf B73 sample. Genes of the remaining 4 inbreds were classified using their DNA methylation and the ERC-2 model (Fig. 3 A, C, and E).

The ERC-2 was also used to determine whether developmentally regulated differences in genic methylation are associated with tissue-specific gene expression potentials (Fig. 3 B, D, and F). Three previously published B73 tissue data sets were examined (anther, developing ear, and shoot apical meristem [SAM]). For each sample, both DNA methylation and RNA-seq data were collected (18). We observed much greater variability in classification scores between inbreds than we did between tissues (Fig. 3 C and D; blue dots). On average, 2,160 genes have differential classifications between 2 inbreds while only 140 genes have differential classifications between 2 tissues.

The ERC-2 performed well for all of the inbred lines (Fig. 3A) and the tested tissues (Fig. 3B), with areas under ROC curves of 0.9 or greater. The test samples were compared to each other in a pairwise fashion to give 6 comparisons among the 4 inbreds and 3 comparisons among the 3 tissues. The ERC-2 classification scores were plotted for each comparison, with the lower score plotted on the x axis and the higher score plotted on the y axis (Fig. 3 C and D). As expected, most genes receive the same classification for the 2 samples in question (98 and 99.9% for inbred and tissue comparisons, respectively), indicating that most genes possess similar methylation patterns (Fig. 3 C and D; gray dots). On average, 2,160 and 140 genes have a differential classification between 2 inbreds and tissues, respectively, resulting from differential DNA methylation (Fig. 3 C and D; blue dots), causing them to be classified as silent in one sample type and expressible in another. We designated differential classifications as a difference in score greater than 0.6. We plotted observed mRNA abundance for the differentially classified genes (Fig. 3E). As with the prediction scores, each gene was plotted with the expression of the lower-predicted sample on the x axis and the higher-predicted sample on the y axis. We used fragments per million (FPM) of 1 as a cutoff below which genes were considered not expressed. A total of 1,466 genes was expressed in at least one inbred. Of these, 1,004 (~68%) were not expressed when classified as silent, and 1,269 (~87%) had lower expression when classified as silent. This is similar to the average accuracy of the expression models (AUC = 0.92; Fig. 3A). We compared the different tissues, but only 17 genes were differentially expressed with 65% above the diagonal in Fig. 3F. These results indicate that the set of transcribed genes varies significantly.
between genotypes and much less so between tissues, consistent with previous reports on differential DNA methylation (19). Many of the silenced genes may arise from recently copied gene fragments that have been captured inside of TEs (20). We observed that as new inbreds are added to the analysis, the transcribed gene set of the species is expanded (SI Appendix, Fig. S10). Thus, maize appears to have a panexpressible gene set of significantly more genes than are transcribed in any individual inbred. The inherited epigenomic patterns that we have associated with permission for expression, along with cis and trans transcriptional regulation, give rise to the pantranscriptome where the phenomenon of inbred-specific expression has been characterized (21). The panexpressible gene set is distinct from the pangenome which arises from structural variation (9). Differential DNA methylation exhibits relatively stable transgenerational inheritance (17), and therefore, our models predict that hybrids will express a set of genes that is the sum of the sets expressed by the 2 inbred parents. This prediction has been confirmed (22). While DNA methylation is generally stable, it likely has a spontaneous rate of change greater than DNA sequence yet low enough to maintain a long-term selection response (23). Therefore, spontaneous mutations in DNA methylation will occasionally cause expressible genes to become silent and silent genes to become expressible.

Silenced Genes Have Distinct Attributes Compared to Expressible Genes. The ERC-2 results allow us to compare attributes of expressible genes to those of silenced genes to characterize the large set of silenced genes. The ERC-2 classifications predict that 32,333 genes can be transcribed across all 5 inbreds. Of these, 22,101 have syntenic orthologs with sorghum, and 10,232 are nonsyntenic. We refer to these groups as "all inbreds syntenic" and "all inbreds nonsyntenic." In addition, 18,289 genes are transcribed in some subset of the 5 inbreds. We will call this the "any inbred" group. Finally, 50,103 genes are predicted to be silenced in all inbreds, the "no inbreds" group (SI Appendix, Fig. S10 and Datasets S3 and S17). All potential TEs were discarded from all 4 groups for this analysis (SI Appendix, Fig. S10B). Categorical enrichment was carried out on each group (Dataset S3, Tabs 2–5). Interestingly, we see significant enrichment for "biotic stress" and "secondary metabolism" categories in the all inbreds nonsyntenic group, indicating more recent selection for...
Gene body methylation has been described for both plants and animals. This work was supported by NSF Grant IOS-1546899 (to S.P.B.).

ACKNOWLEDGMENTS. This work was supported by NSF Grant IOS-1546899 (to S.P.B.).


30. A. Olson et al., Expanding and vetting Sorghum bicolor gene annotations through transcriptome and methylome sequencing. Plant Genome 7, 10.3835/plantgenome2013.08.0025 (2014).


42. A. Olson et al., Expanding and vetting Sorghum bicolor gene annotations through transcriptome and methylome sequencing. Plant Genome 7, 10.3835/plantgenome2013.08.0025 (2014).


