PNAS
www.pnas.org

Supplementary Information for

Identification of the expressome by machine learning on omic data

Ryan C. Sartor[1], Jaclyn Noshay[2], Nathan M. Springer[2], Steven P. Briggs[1A]
[1]University of California, San Diego, La Jolla, CA; [2]University of Minnesota, St. Paul, MN
[A]Corresponding author;
sbriggs@ucsd.edu

This PDF file includes:

Supplementary text
Figs. S01 to S13
Captions for additional supplementary files
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Supplemental Datasets S01 to S17
R Source Code Datasets S18 & S19

www.pnas.org/cgi/doi/10.1073/pnas.1813645116

Supplementary Information Text

Materials and Methods

Data Processing and Manipulation

The R Statistical Programing Language [1] along with the Rstudio integrated development environment [2] was used for all data processing and manipulation unless otherwise specified.

Figure Plotting

All figure were generated using R software. All details can be found in the supplemental source code file (Supplemental_Dataset_S18 source code: Lines 172:976)

Quantitation of Whole Genome Bisulfite Sequencing Data

(Supplemental Dataset S04 & Supplemental Dataset S05)

Raw sequencing reads for Whole Genome Bisulfite Sequencing (WGBS) of maize B73 14-day old seedlings were downloaded from the NCBI Sequence Read Archive (**SRA**) (SRA# SRR850328; [3]). Files were converted to fastq format using fastq-dump from the SRA toolkit. The Trim Galore program [4] was used with the "--paired" argument to remove paired-end adapter sequences. The Bismark program [5] was used to identify methylated cytosines against the maize B73 RefGen_v2 genome, allowing for 1 miss-match "-N 1". Individual methylated cytosines were summarized separately for the 3 methylation contexts (CG, CHG and CHH) using "bismark_methylation_extractor" with arguments "--paired-end" and "--no-overlap" which prevents duplicated methyl-C counts from read pairs. Gene model chromosomal coordinates were obtained from the B73 RefGen_v2 working gene set (known as 5a_WGS). These loci were extended 250 bp upstream in order to include sequences around the transcription start site (TTS). All three methylation contexts were then mapped to these loci using the R statistical computing language. Several genomic regions were identified for each gene model using the 5a WGS annotations. These regions included 4 promoter regions (bins 1-4), covering 2Kb upstream of the TTS in 500 bp increments. 500 base pairs around the TTS (250 bp up and down stream), the 5'UTR, All Exons, All Introns, the 3' UTR and the whole gene model (from TSS to transcription stop sight). Methylation of each of these regions was quantified separately. For each context, the fraction of methylated cytosines relative to the total detected cytosines was used as a measure of methylation level (Supplemental Dataset S04), known as the weighted methylation level [6].

Gene-wise Binning of DNA Methylation Levels

(Supplemental Dataset S05)

All gene models were divided into 5 equal sized, consecutive sections called bins (Fig 01_A). Quantitation of DNA methylation levels was carried out separately for each bin.

Handling of Missing DNA Methylation Data

Of the 110028 gene models in the maize B73 5a Working Gene Set, 11732 did not have coverage in the WGBS data. These genes were discarded from further analysis. For genes with coverage along the gene model, not all genomic features have WGBS coverage. In this case, missing data was simply substituted with a place-holder value of 0.5 which represents an un-informative number as this feature is neither hypo or hyper-methylated.

Observed Protein List and Protein Abundance Data

(Supplemental Dataset S06 & Supplemental Dataset S07)

All protein data is from [7]. Protein abundance (Supplemental Dataset S06) was taken as the average across 23 potential tissues in which the protein had detectable levels [7]. For each protein, any tissue where no protein was detected, this tissue was not included in the average. Only proteins with at least one uniquely mapped peptide were used in this study (Supplemental Dataset S07).

mRNA Abundance Data

(Supplemental Dataset S08)

mRNA abundance data used for the Express-able mRNA Classified (ERC), Express-able protein classifier (EPC) and the Protein-specific Feature Illuminator (PFI) were taken from [7] by averaging across 23 potential tissues in which the transcript was detected. For each transcript, any tissue where no transcript was detected, this tissue was not included in the average.

Filtered Gene Set Annotation

(Supplemental Dataset S09)

The list of 45,348 gene models (parent genes) was taken from the Maize GDB [8] annotation file "ZmB73_5b.60_FGS.translations.fasta" found on the Maize GDB ftp site (http://ftp.maizegdb.org/). This list is constructed and maintained by the Maize GDB project. its construction is described on the site as:

"The Filtered Gene Set (FGS) is a subset of the Working Gene Set intended to exclude transposons, pseudogenes, contaminants, and other low-confidence annotations. Maizesequence.org used essentially the same method as described in [9] but with modifications. First, the inclusion criterion of synteny (relative rice, sorghum, and Brachypodium distachyon) was given higher precedence than the exclusion criteria of pseudogene and transposon. This measure was taken to avoid exclusion of possibly legitimate genes that may have been miss-assembled or miss-annotated. Second, selection of the FGS was additionally informed by evidence of expression, taking advantage of RNA-seq data displayed on this site [10]. In addition, a total of 786 genes in the 4a filtered set [9] were rejected this time around primarily because they physically overlap with other genes in the FGS and have inferior evidence/confidence."

(Supplemental Dataset S02)

The list of 131,496 gene models in the RefGen_V4 was taken from the annotation file: "Zea_mays.AGPv4.cdna.all.fa" found on the Gramene FTP site (ftp://ftp.gramene.org/pub/gramene/). This list was reduced to 39,498 gene models by only retaining the locus name (i.e. ignoring various isoforms from the same gene model). These V4 genes were converted to V3 names using a cross reference table downloaded from Maize GDB "V3_V4_xref.txt". Only 30601 gene models could be successfully cross-referenced back to V4, limiting the comparisons that can be made (Supplemental Dataset S09).

Gene Biotype Annotation
(Supplemental Dataset S10)

Gene Biotypes were extracted for the RefGen_V2 working gene set (5a) gff annotation file "ZmB73_5a_WGS.gff" that was downloaded from the Maize GDB ftp site (http://ftp.maizegdb.org/). Biotype information is listed under the information field for "gene" features.

Annotation of the Number of Introns
(Supplemental Dataset S11)

For each gene model in the RefGen_V2 working gene set, the number of introns was determined by counting all unique introns listed in the gff annotation file "ZmB73_5a_WGS.gff" that was downloaded from the Maize GDB ftp site (http://ftp.maizegdb.org/).

Identifying Genes with Transposable Elements
(Supplemental Dataset S12)

The RefGen_V2 working gene set annotation (5a) provides a gene biotype. Anything with biotype "transposable element" was taken as a TE. Also, 7,612 additional gene models were found to contain one or more regions with a significant blast hit (E-value = 0) to one or more reference TE sequences from the Maize TE database [11].

Whole Genome Bisulfite Sequencing of Multiple Inbred Lines and Multiple B73 Tissues
(Supplemental Figshare Data & Supplemental Dataset S13)

Whole Genome Bisulfite (WGBS) of leaf tissue from several maize genotypes (Supplemental Figshare Data) [12] as well as multiple tissues from B73 (Supplemental Dataset S13) [13] have been used. Briefly, 3rd leaf tissue of five genotypes (B73, Mo17, Oh43, CML322, and Tx303) and 3 additional tissues from B73 (anther, developing ear and shoot apical meristem) were used for DNA isolation and bisulfite sequencing. Sodium bisulfite converted sequencing libraries were generated as described in [14]. 100-bp paired end reads were obtained and assessed for quality by FASTQC [15]. Read tails lower than quality 20 were removed. All reads were then mapped to B73 reference genome (version 2: ZmB73_5a) with BSMAP and methratio.py for calling methylation levels at a single cytosine. Methylation levels were summarized across 100-bp sliding windows for all three sequence contexts (CG, CHG, and CHH).

RNA-sequencing of Multiple Inbred Lines and Multiple B73 Tissues
(Supplemental Dataset S14 & Supplemental Dataset S15)

RNA sequencing reads were generated from the 3rd leaf of 5 genotypes (B73, Mo17, Oh43, CML322, and Tx303) [14] (Supplemental Dataset S14) and 3 additional tissues from B73 (anther, developing ear and shoot apical meristem) [13] (Supplemental Dataset S15). All samples consist of 3 biological replicates. The quality of the paired-end reads was assessed using FastQC [15]. Any reads containing adapters or runs with irregular base content at 5' or 3' ends were trimmed using cutadapt [16]. Reads were mapped to the Maize B73 RefGen_V2 (5a) genome using HISAT [17] with the following argument:
-k 2 --maxins 12000 --no-mixed --no-discordant --quiet --met-stderr --no-unal --threads 7
Reads mapping to exons were counted using htseq-count [18] with the following arguments:
-s no -r name -t exon -i gene_id -m union --nonunique none
Fragments per million (FPM) were calculated as (counts / (total reads per sample / 1000000)).

B73 Histone Modification Data
(Supplemental Dataset S16)

Abundance data for 3 different histone modifications (H3K36me3, H3K9Ac, and H3K4me3) was taken from [19]. Peak data for each modification was mapped to the B73 RefGen_V2 genome. The summed peak intensity for both shoot and root samples was calculated for each gene bin separately. A peak was assigned to a bin if any part of the peak overlapped the bin. The average of shoot and root intensities was used for the final values.

Construction of Classification Models (ERC and EPC)
(Supplemental_Dataset_S18 source code: Lines 64 – 119)
All models used for classification were constructed in the same way, with differences only in the class labels on which they were trained. Random Forest classification models were used [20] as implemented in the R statistical programing language [21]. For training sets, a matrix of methylation data described above was used as features (independent variables) for every model and a unique set of class labels (dependent variables) was used to train the Express-able Protein Classified (EPC) and the Express-able mRNA Classifier (ERC) separately. These class labels were defined using transcript abundance data (ERC) or a combination of transcript and protein abundance data (EPC). Therefore, two classifiers were built. For the ERC class label training vector, the "Non Expressable" class was defined as genes with no detectable mRNA abundance (No RNA: 37588 genes) and the "Express-able" Class was defined as high abundance mRNA genes with FPKM >1 (High RNA: 33696 genes). For the EPC, this class labels were further refined by also requiring no detectable Protein in the "Non Expressable" class (No RNA / No Protein: 37487 genes) and

detected protein in the “Express-able” class (High RNA / Observed Protein: 15421 genes). The random forest models were built with the class label vector as a factor type using 1000 trees and importance = T which returns the “Mean Decrease in Accuracy” measure of feature importance.
(Supplemental_Dataset_S18 source code: Lines 658-707).
A different DNA-methylation data set was used when looking across different genotypes and different B73 tissues (Fig 03, see above for details). These data were summarized independently and 100bp tiling was done for quantitation (see above). Therefore, a new EPC model was re-trained on this tiled data. The same training class label vector for the EPC was used. Prediction scores for all models were taken as the proportion of Votes for the positive class (using the out-of-bag cross validated prediction for training set genes). These scores are reported for the EPC, ERC and PFI in supplemental Dataset S01 and for all ERC-2 predictions of various maize inbred lines and B73 tissues in supplemental Dataset S03.

Classification of Test Data
(Supplemental_Dataset_S18 source code: Lines 151 – 158)

The Random forest classifiers (ERC and EPC) were used to classify the remaining genes that were not represented in the training set. The Random Forest “predict()” function was used along with the DNA methylation data for the test genes. This same procedure was used to classify genes based on methylation data from the different maize inbred lines (Supplemental_Dataset_S18 source code: Line 710) and different B73 tissues (Supplemental_Dataset_S18 source code: Lines 762). The classifications for genes in the training set were obtained from the Out-of-Bag cross-validated predictions of each model (“$predicted” slot from the Random Forest object).

Construction of Regression Model (PFI)
(Supplemental_Dataset_S18 source code: Lines 64 - 119)

The PFI regression model was built in the same manner as the classification model described above. The only difference is the class labels used for training. For the PFI, the first class was defined as genes with mRNA expression > 1 FPKM and non-observed protein (High RNA / No Protein: 18275 genes). The other class is composed of genes with mRNA expression >1 and observed protein (High RNA / Observed Protein: 15421 genes).

Construction of Quantitative Expression Predictors
(Supplemental_Dataset_S18 source code: Lines 102 - 126)

Two predictive models were constructed in an attempt to predict quantitative mRNA and Protein expression levels using DNA methylation data. The mRNA Expression-level Predictor (REP) and Protein Expression-level Predictor (PEP) were built using random forest models in much the same way as the ERC and EPC (described above), with one difference. The dependent variable training vector was a continuous numeric variable representing either observed mRNA or observed Protein log2(abundance). Non-observed values are

represented as the integer less than the lowest observed value for the data set (since log2(0) = -infinity). This is -12 for the mRNA data and -1 for the protein data. The Out-Of-Bag cross-validated results were extracted from the "$predicted" slot in the Random Forest objects. These values were used for further analysis and plotting.

Feature Importance (Regression) Measure

(Supplemental_Dataset_S18 source code: Line 130)

The feature importance was determined by the random forest algorithm, using the mean decrease in accuracy upon random permutation of each individual variable [20]. This requires the parameter "importance = TRUE" in the call to the "randomForest()" function.

Determination of the Mathematical Sign (Direction) of Relationships Between Methylation Features and Classification

(Supplemental_Dataset_S18 source code: Lines 134 - 147)

The sign of the relationship between DNA-methylation features and classifications was determined as follows. Genes in the training set were split between the two training classes (Non Expressable and Express-able) to yield two populations (S05). For each feature, the corresponding feature values were assigned to each of the class populations and a student's t-test was carried out between the populations. The sign of the resulting t-statistic was assigned to each feature while the magnitude of each was taken from the RF variable importance (Fig02_D-F).

Receiver Operating Characteristic (ROC) Curves and Precision Vs. Recall Curves

(Supplemental_Dataset_S19 source code: PlotROCPR(): Lines 96 - 231)

The "ROCR" R package [22] was used to generate all ROC curves. See [23] For a description of ROC and PR curves. For each gene, the proportion of positive votes from the random forest model was used as a quantitative classification score and this was evaluated against the known class labels. For the ERC and EPC, the votes were taken from the out-of-bag cross-validated predictions (Supplemental_Dataset_S18 source code: Line 310). These predictions have been shown to give accurate error rates compared to an independent test set [24]. When evaluating model accuracies of independent test samples (different B73 tissues and different maize inbreds) the ERC-2 random forest model trained on B73 leaf data was used to classify these independent test sets. Again, the number of votes were used as a quantitative classification score (Supplemental_Dataset_S18 source code: Lines 710 & 762). These scores were evaluated against a response vector that was generated from separate RNA-seq data sets corresponding to each test sample. In these response vectors, "positive" genes were ones with mRNA abundance > 1 FPKM and "negative" genes were ones with undetected mRNA abundance.
(Supplemental_Dataset_S18 source code: Lines 306-315).

For the ROC and PR curves representing the predictive ability of the V2 and V4 filtered gene sets (Supplemental Figure 07), The same class labels were used as above. For the prediction "score", vectors were generated using the BuildYvec() function with MakeQuant=T. (Supplemental_Dataset_S19 source code: Lines 60-81). This function randomly generates a set of numbers around zero [-0.1,0.1] to represent all genes not in the FGS (the negative set) and randomly generates a set of numbers around 1 [0.9,1.1] to represent all genes in the FGS (the positive set).

For all curves, the "prediction()" function from the ROCR package was used to evaluate the model. For the ROC curves, the "performance()" function was used with arguments "tpr" and "fpr" to generate curves and the argument "auc" was used to calculate the area under the curves. For PR curves, the "performance()" function was used with arguments "prec" and "rec" to generate curves. The area was calculated by estimating a function for the curve with "approxfun()" and integrating across its entire length
(Supplemental_Dataset_S19 source code: Lines 247-248)

CG Gene Body Methylation Genes
(Supplemental_Dataset_S19 source code: RetrurnGBMGenes(): Lines 383 - 395)

All genes with DNA-methylation data were tested for CG Gene body methylation (gbM). A gene was said to have gbM if the average methylation of the end bins (1 and 5) was less than 0.5 and the average methylation of the center bins (2-4) was greater than 0.5.

Model Testing and Validation on Multiple Maize Genotypes and Multiple B73 Tissues
(Supplemental_Dataset_S18 source code: Lines 658:961)

A new ERC-2 model was constructed by using the same class labels from the ERC but with a different 100bp tiled methylation data feature set from B73 3rd leaf (see above). A random forest classifier was trained. Next, the WGBS data for the additional 4 genotypes (Mo17, CML322, Oh43 and Tx303) as well as from 3 additional B73 tissues (anther, developing ear, SAM) were used as test datasets and classified using the B73-trained model. These classifications were then validated using corresponding RNA-seq data for these same genotypes. To further assess genes with differential classifications between genotypes. The 4 test genotypes were compared in a pairwise fashion (6 pairs). The same was done for pairwise comparisons between the 3 different tissues (3 pairs). For each of these pairwise comparisons, a number of genes have difference in prediction score between the two samples of 0.6 or greater. These were defined as differential predictions (blue dots in Fig03_C&D). Next the relative mRNA abundance was examined for the differential predicted genes among the corresponding sample pairs (Fig03 E&F).

Categorical Enrichment

Enrichment analysis was carried out on four sets of genes defined using ERC2 predictions. The four sets are: (i) genes predicted to be expressable in all inbreds examines and have syntenic orthologs with sorghum (All_Inbreds_Syntenic), (ii) or no synteny (All_Inbreds_NonSyntenic), (iii) genes that are predicted to be expressable in a subset of the 5 inbreds (Any_Inbred) and (iv) genes that are predicted to be silenced in all inbreds (No_Inbreds). A custom R script was written to carry out the analysis separately on each gene set using the MapMan categories. MapMan categories were annotated for every gene in the RefGen_v2 working set with the Mercator annotation pipeline [25]. For the stable and variable gene lists, every MapMan bin that was represented was examined for enrichment. A hypergeometric test was performed using the R function phyper() from the stats package. The total set of genes (Number of black and white balls) used was the intersection of genes with MapMan annotations and genes with methylation data. The number of white balls was the total number of genes annotated with the MapMan category. The number of draws was the number of genes in the test set and the number of white balls drawn was the number of genes in the test set with the MapMan annotation – 1 (1 must be subtracted for right-tail calculations due to the implementation of this function). Within each module, p-values were corrected for multiple testing using the p.adjust() function with (method="fdr") which performs the Benjamini & Hochberg correction. Results are shown in supplemental Dataset S03, under the "All_Inbreds_Syntenic", "All_Inbreds_NonSyntenic", "Any_Inbred" and "No_Inbreds" tabs.

Protein Sequence Similarity Network With ERC2 Genes

A standalone version of BLAST was used to run an all vs. all blastp analysis on the protein sequence of 67,262 non-TE genes used in the ERC2 analysis of the 5 maize inbreds. A cutoff E-value of 0.01 was used to establish "edges" between genes. Next the BLAST bitscore was used as a weight for each edge and the MCL clustering algorithm [26] was used to derive clusters with default clustering parameters. Clusters are defined in Supplemental Dataset S17 and used in Fig S10_K&L.

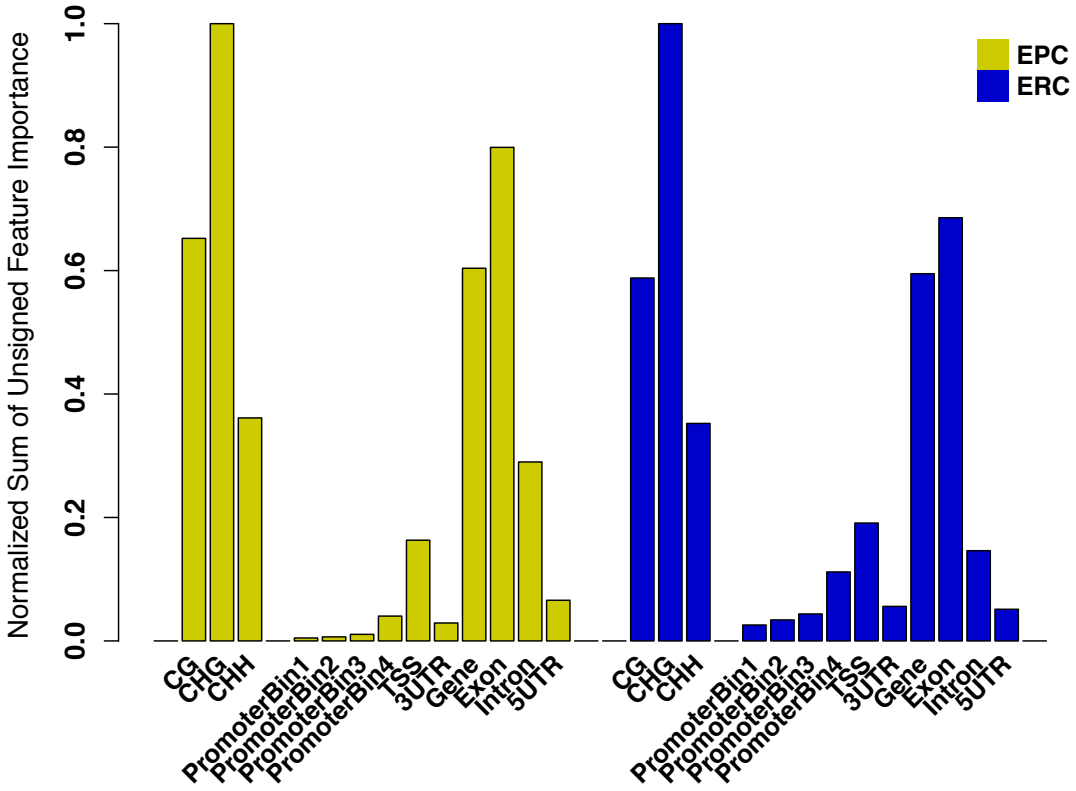

Fig. S01. Summarized measures of feature importance summed over various methylation features.

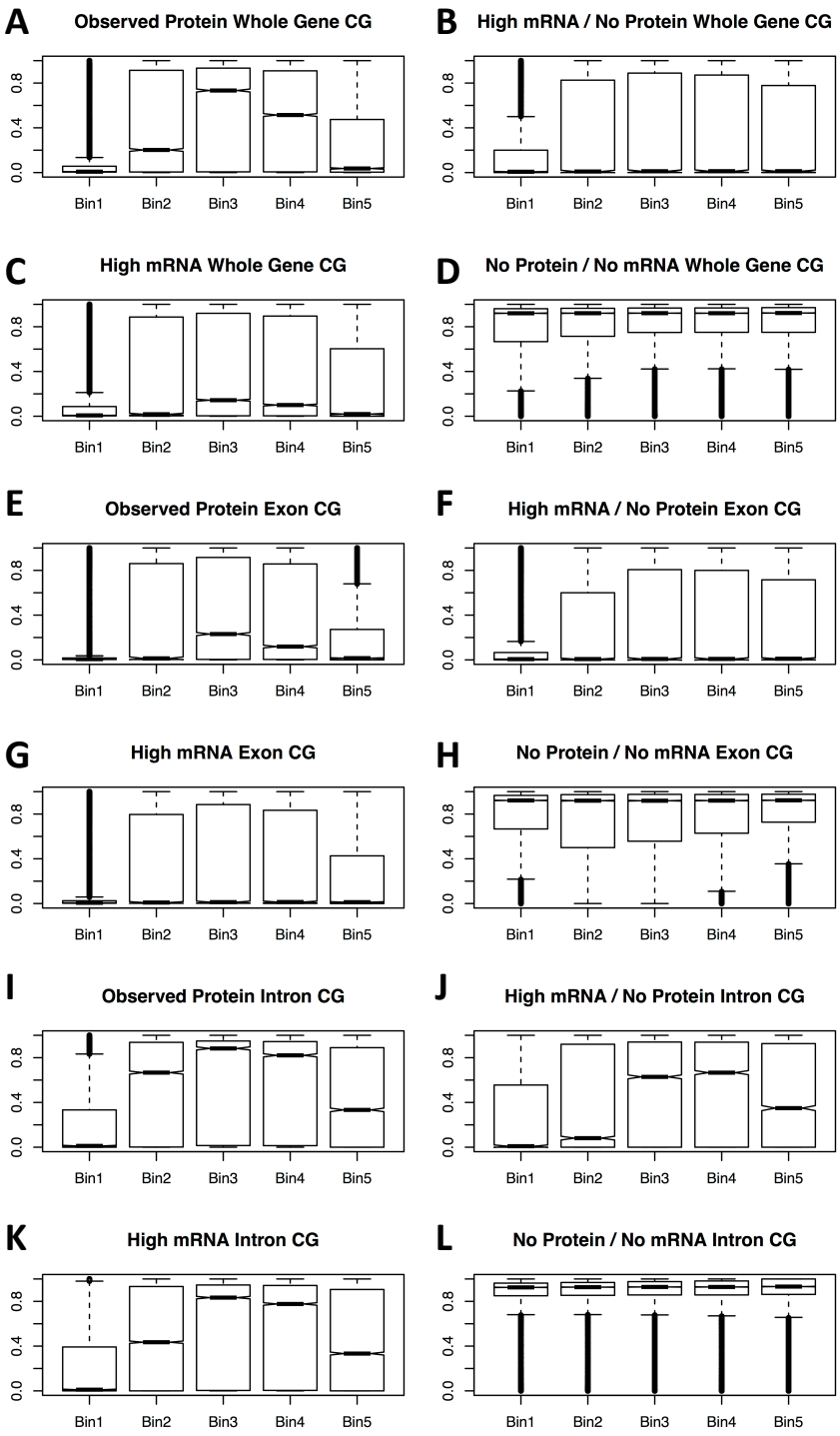

Fig. S02. Binned DNA methylation levels in the CG context. Boxes show distributions of the proportion of (methylated cytosines / all cytosines) for all 5 bins for various gene sets determined based on mRNA abundance and the presence of observed or non-observed proteins. (A-D) Summarized over the whole gene models. (E-H) Summarized over exons only. (I-L) Summarized over introns only.

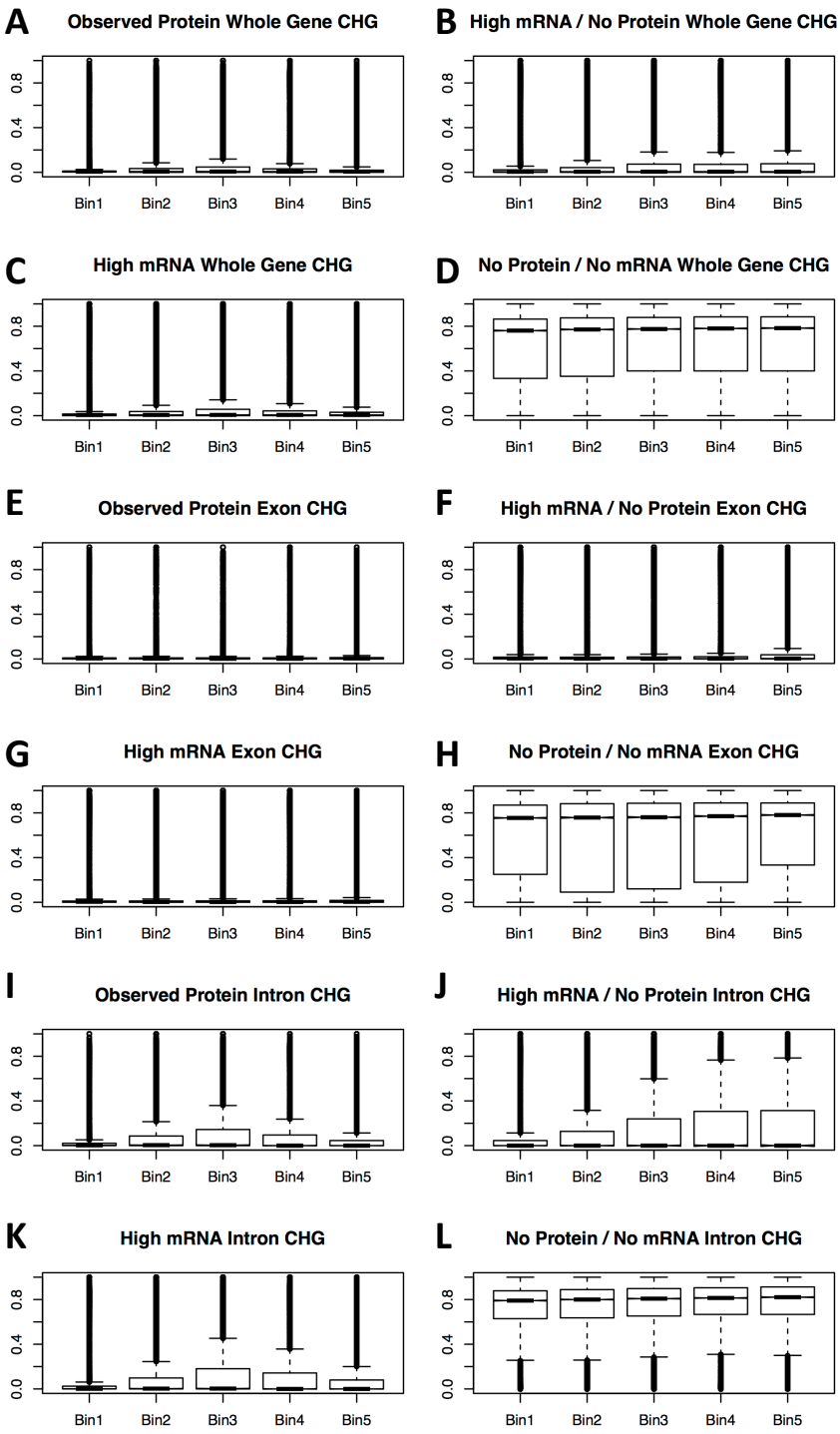

Fig. S03. Binned DNA methylation levels in the CHG context. Boxes show distributions of the proportion of (methylated cytosines / all cytosines) for all 5 bins for various gene sets determined based on mRNA abundance and the presence of observed or non-observed proteins. (A-D) Summarized over the whole gene models. (E-H) Summarized over exons only. (I-L) Summarized over introns only.

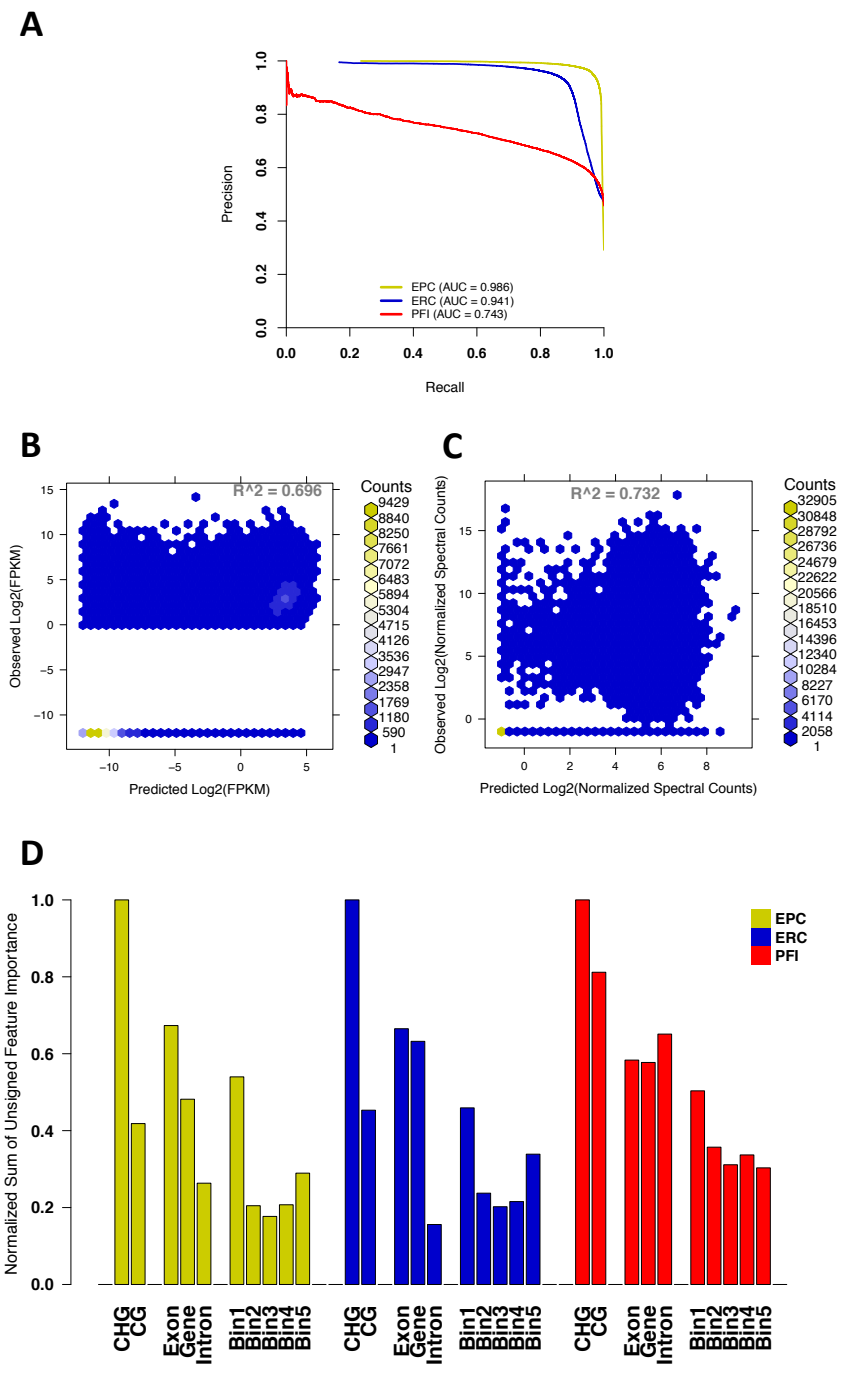

Fig. S04. Results for random forest models. (A) Precision vs. Recall (PR) curves showing classification accuracy of the EPC, ERC, and PFI models to predict expressable gene sets. (B) Prediction accuracy for quantitative mRNA abundance model measured as Log2(FPKM), looking at all genes with methylation data. (C) Prediction accuracy for quantitative protein abundance model measured as Log2(Normalized Spectral Counts), looking at all genes with methylation data. (D) The non-signed feature importance for each genomic region.

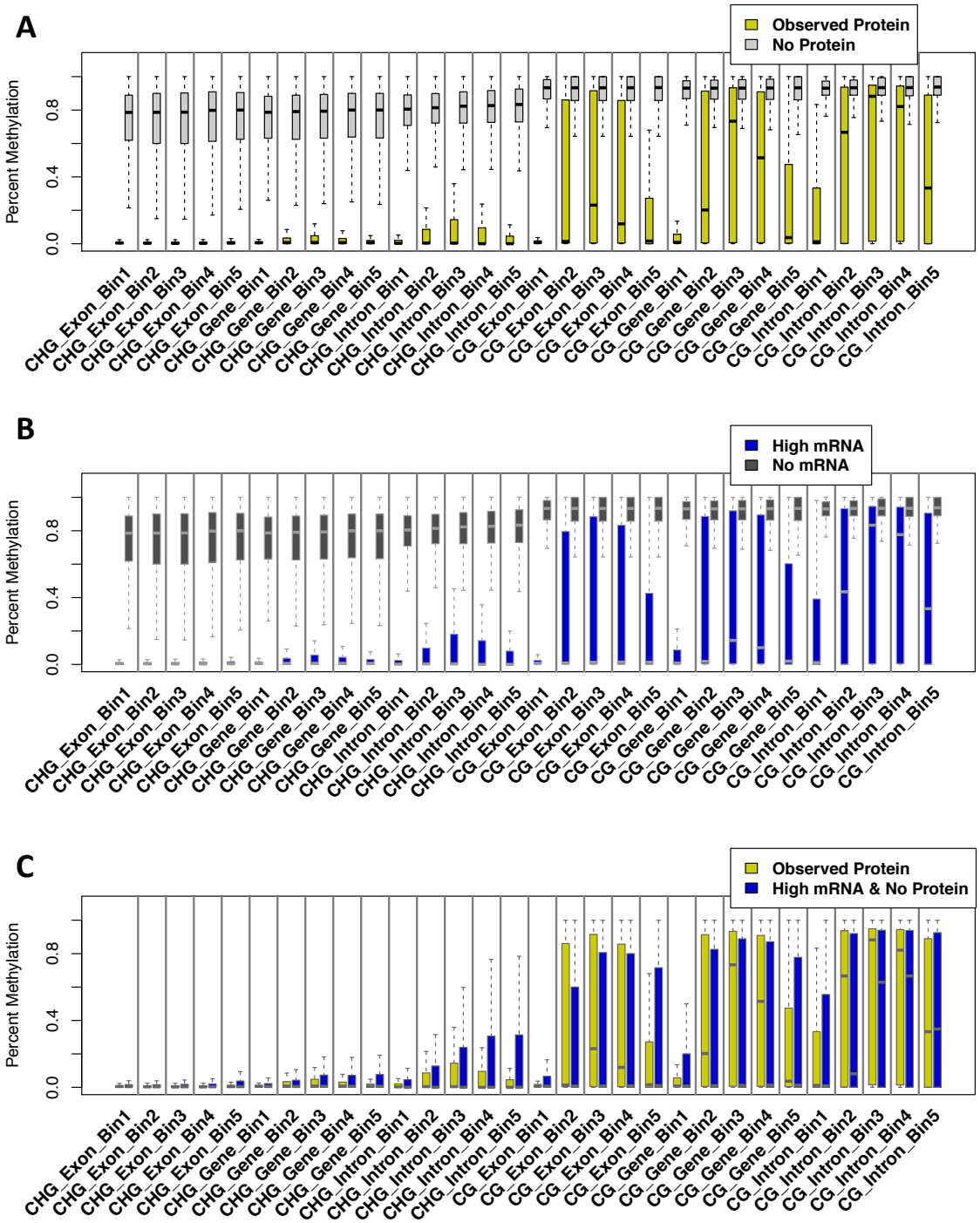

Fig. S05. Boxplots showing distributions of methylation level for all feature for the two different training classes of each random forest model. (A) Results for the EPC model. (b) Results for the EPC model and (C), results for the PFI model.

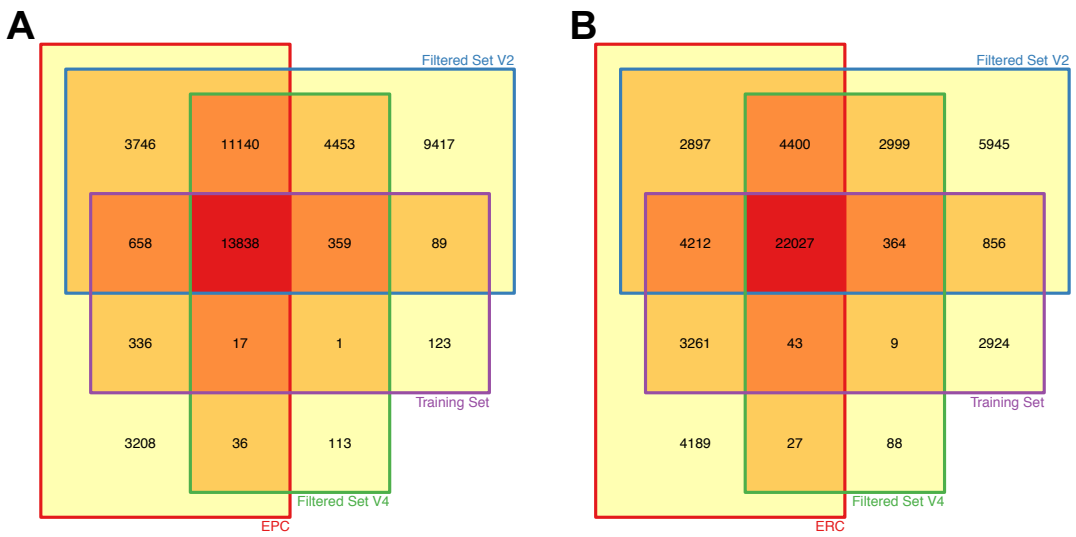

Fig. S06. Results from EPC and ERC classification. (A) Venn diagram showing the EPC results for genes classified as expressable proteins (EPC), all genes with observed proteins in the training set (Training Set), and the filtered gene sets from RefGen_V2 and RefGen_V4. (B) Venn diagram showing the ERC results for genes classified as expressable mRNA (ERC), all genes with high mRNA in the training set (Training Set) and the filtered gene sets from RefGen_V2 and RefGen_V4

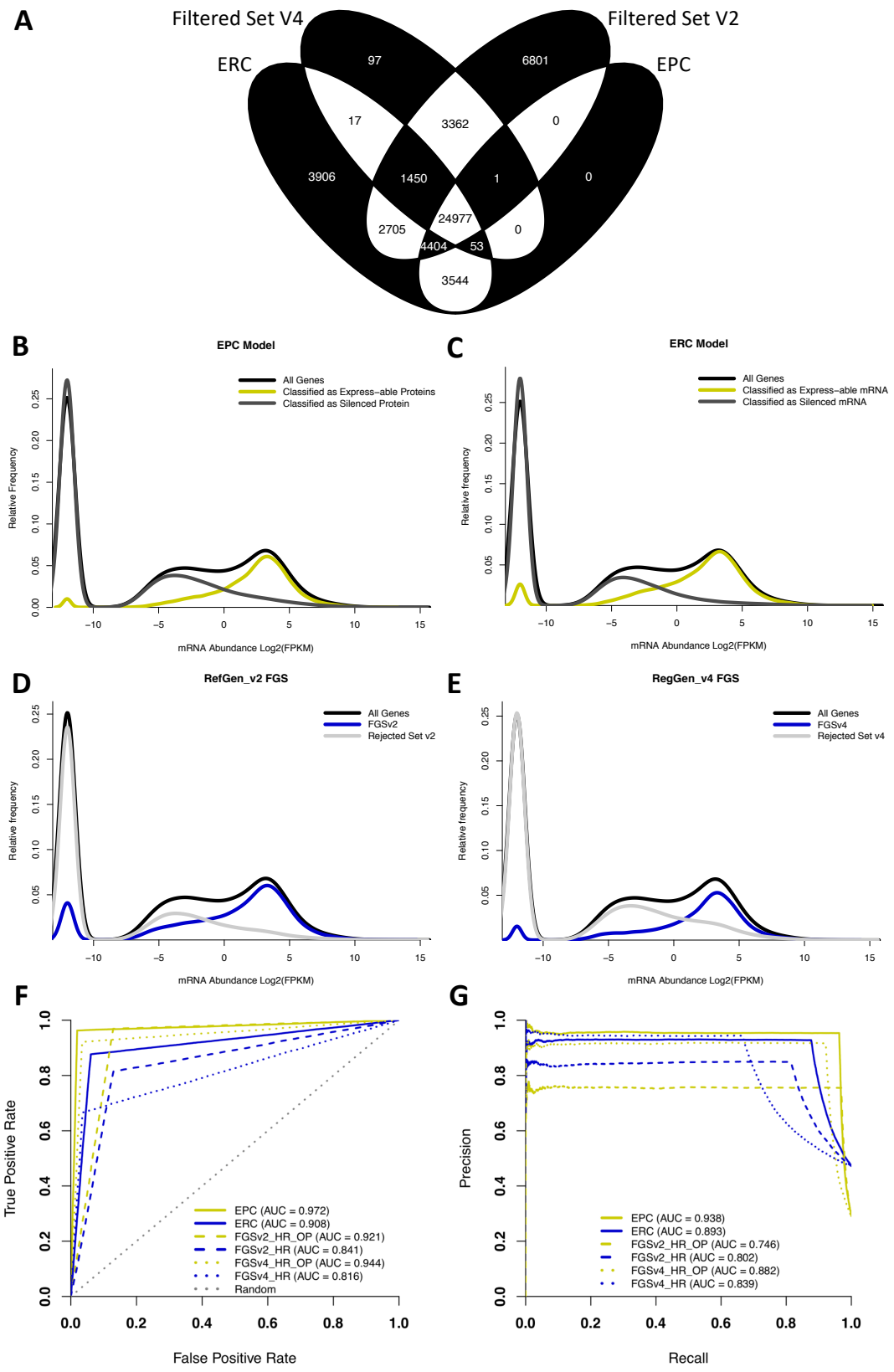
A
Filtered Set V4
Filtered Set V2
ERC
EPC
97
6801
17
3362
0
3906
1450
1
0
2705
24977
0
4404
53
3544
B
EPC Model
All Genes
Classified as Express−able Proteins
Classified as Silenced Protein
Relative Frequency
mRNA Abundance Log2(FPKM)
C
ERC Model
All Genes
Classified as Express−able mRNA
Classified as Silenced mRNA
Relative frequency
mRNA Abundance Log2(FPKM)
D
RefGen_v2 FGS
All Genes
FGSv2
Rejected Set v2
Relative frequency
mRNA Abundance Log2(FPKM)
E
RegGen_v4 FGS
All Genes
FGSv4
Rejected Set v4
Relative frequency
mRNA Abundance Log2(FPKM)
F
True Positive Rate
False Positive Rate
EPC (AUC = 0.972)
ERC (AUC = 0.908)
FGSv2_HR_OP (AUC = 0.921)
FGSv2_HR (AUC = 0.841)
FGSv4_HR_OP (AUC = 0.944)
FGSv4_HR (AUC = 0.816)
Random
G
Precision
Recall
EPC (AUC = 0.938)
ERC (AUC = 0.893)
FGSv2_HR_OP (AUC = 0.746)
FGSv2_HR (AUC = 0.802)
FGSv4_HR_OP (AUC = 0.882)
FGSv4_HR (AUC = 0.839)

Fig. S07. Classifier results and FGSs. (A) Expressable gene set annotations. Overlap between the EPC expressable class genes, the ERC expressable class genes and the pre-defined maize filtered gene sets from RefGen_V2 and RefGen_V4. mRNA abundance distributions of the various classified groups and filtered sets. (B) Results from the EPC of all genes with methylation data. (C) Results from the ERC of all genes with methylation data. (D) Results from the RefGen_v4 filtered Gene Set of all genes with methylation data. (E) Results from the RefGen_v2 filtered Gene Set of all genes with methylation data. (F) ROC curves showing the accuracy of EPC and ERC expressable gene sets as well as the v2 and v4 filtered gene sets. Plots are evaluating accuracy based on an atlas of maize expression data [7]. “HR” means High-RNA and is comparable to the ERC (blue lines). These are being evaluated based on the observed expressable mRNA. “HR_OP” means High-RNA & Observed Protein and is comparable to the EPC (yellow lines). These are being evaluated based on observed expressable mRNA and protein. (G) PR curves for the same analysis as in (F).

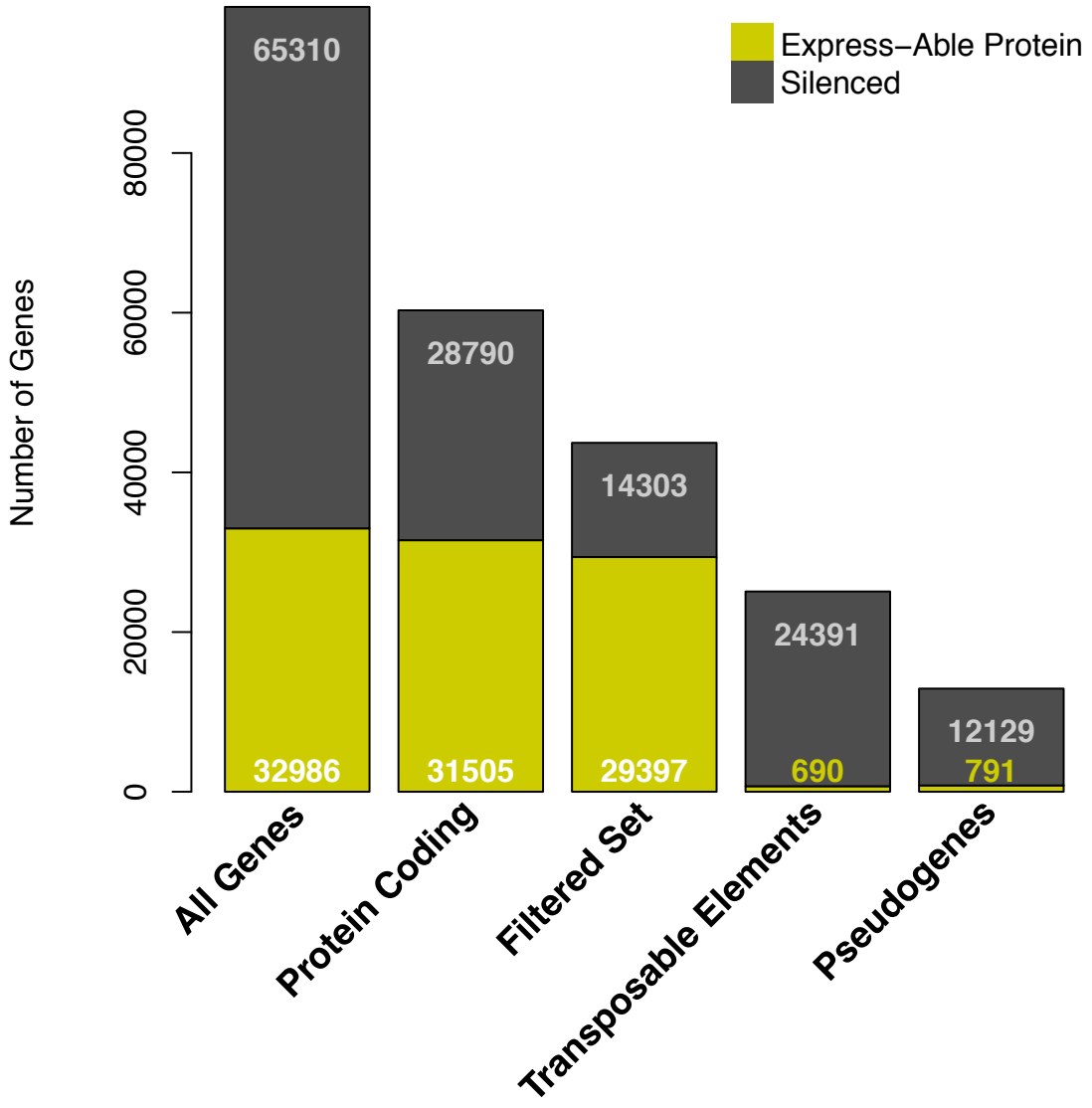

Fig. S08. Classification results from the EPC classifier. Stacked bar plot showing multiple pre-defined gene sets and the relative proportions that are classified as expressable or silenced by the EPC model. “All Genes” are defined by RefGen_V2 WGS. “Protein Coding” genes are the subset of the V2 WGS that remain after “Transposable Element” and “Pseudogene” biotypes are filtered out. “Filtered Set” is further refined using additional evidence (See Materials and Methods).

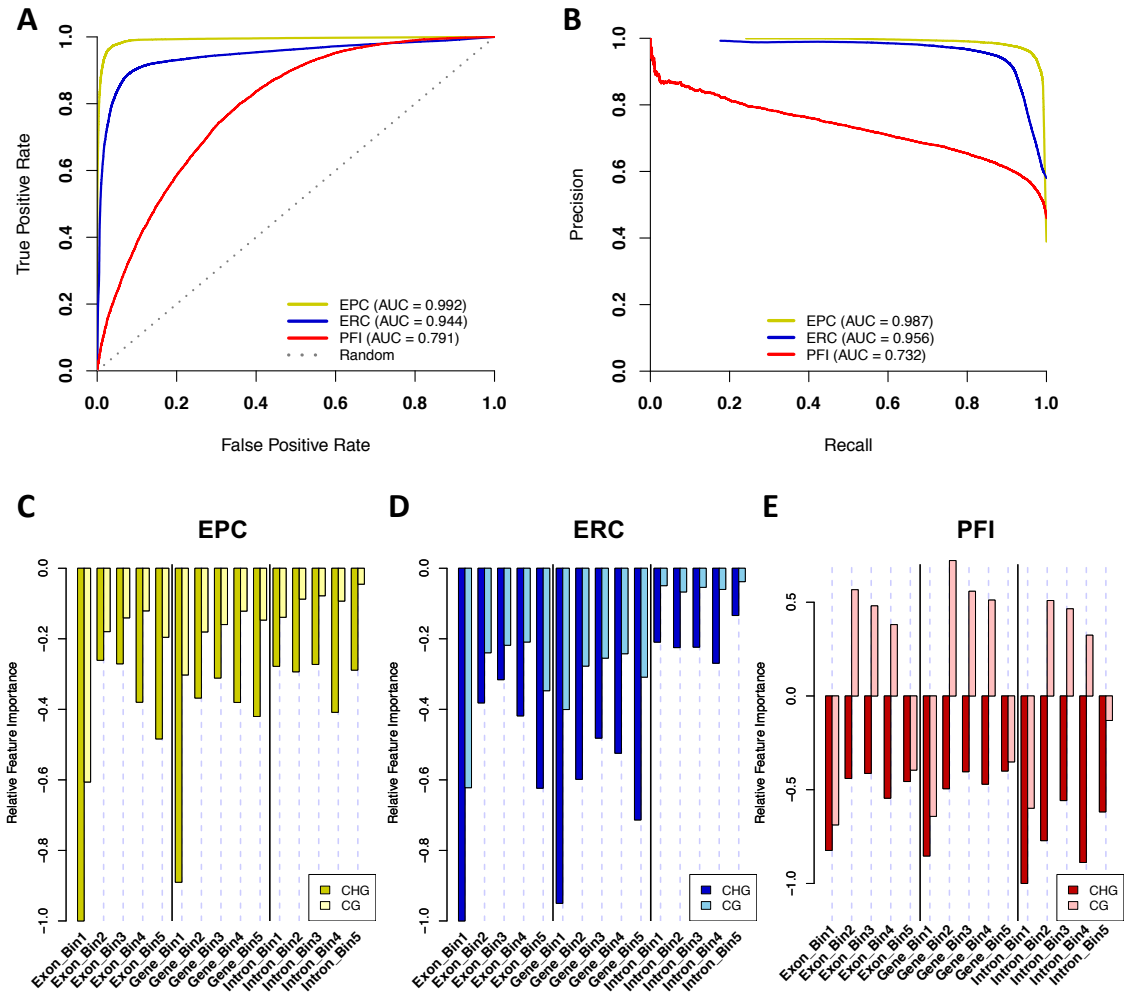

Fig. S09. Classifiers built with no transposable elements. 36,694 genes from RefGen_V2 WGS were identified as likely transposable elements (TEs). Out of the 98,296 genes with methylation data, 32,400 were likely TEs. After filtering out these TEs, classifiers were re-built. (A) Receiver Operating Characteristic (ROC) curve and (B) Precision vs. Recall (PR) curve showing predictive accuracy of the EPC, ERC and PFI. (C-E) 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).

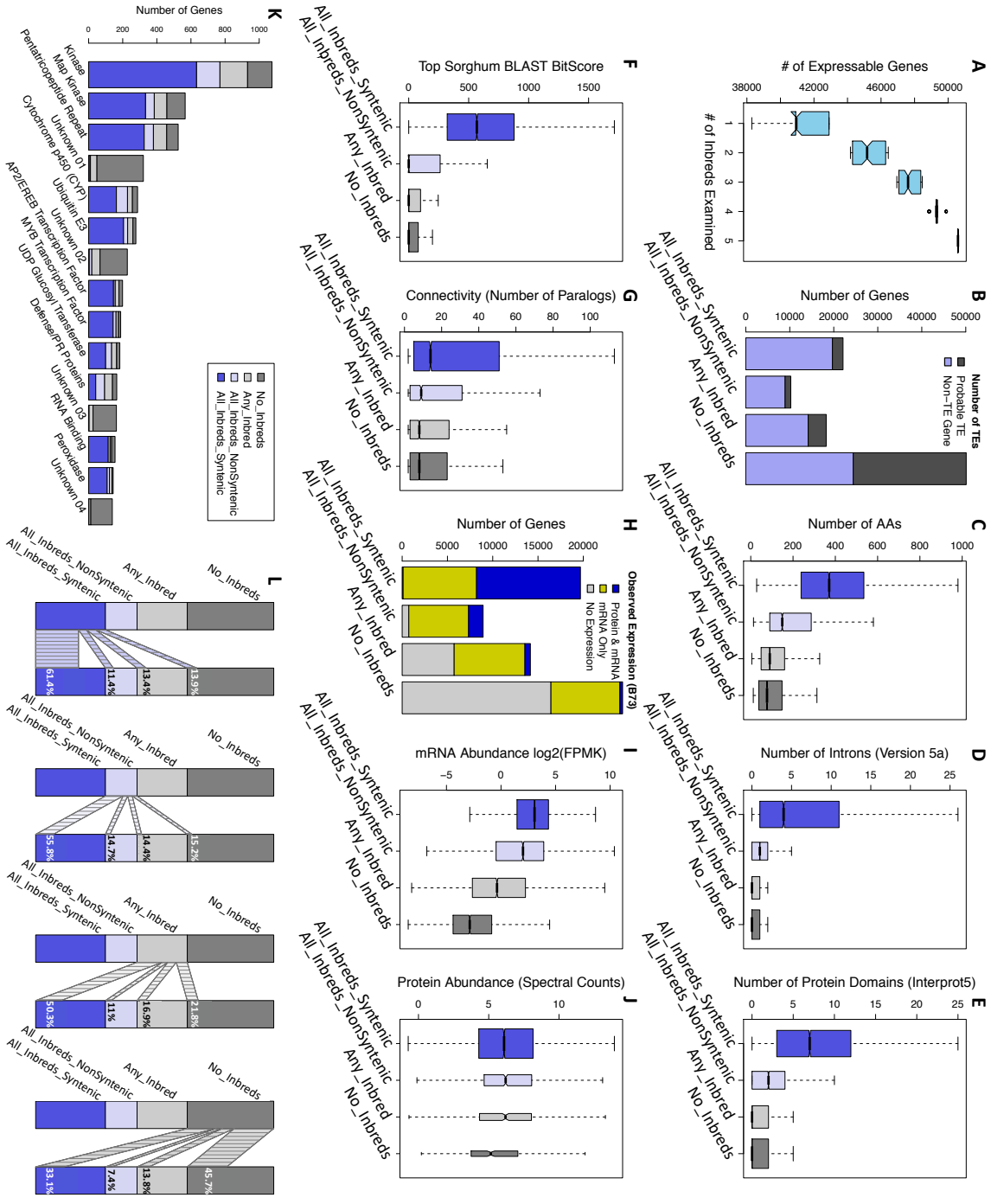
A
# of Expressable Genes
# of Inbreds Examined
B
Number of Genes
Probable TE
Non−TE Gene
Number of TEs
C
Number of AAs
D
Number of Introns (Version 5a)
E
Number of Protein Domains (Interprot5)
F
Top Sorghum BLAST BitScore
G
Connectivity (Number of Paralogs)
H
Number of Genes
Protein & mRNA
mRNA Only
No Expression
Observed Expression (B73)
I
mRNA Abundance log2(FPMK)
J
Protein Abundance (Spectral Counts)
All_Inbreds_Syntenic
All_Inbreds_NonSyntenic
Any_Inbred
No_Inbreds
K
Number of Genes
Kinase
Map Kinase
Pentatricopeptide Repeat
Unknown 01
Cytochrome P450 (CYP)
Ubiquitin E3
Unknown 02
MYB Transcription Factor
AP2/EREB Transcription Factor
UDP Glucosyl Transferase
Defense/PR Proteins
Unknown 03
RNA Binding
Peroxidase
Unknown 04
L
61.4%
11.4%
13.4%
13.9%
55.8%
14.7%
14.4%
15.2%
50.3%
11%
16.9%
21.8%
33.1%
7.4%
13.8%
45.7%

Fig. S10. Examination of various features associated with four sets of genes defined using ERC2 predictions. The four sets are: (i) genes predicted to be expressable in all inbreds examined and have syntenic orthologs with sorghum (All_Inbreds_Syntenic), (ii) or no synteny (All_Inbreds_NonSyntenic), (iii) genes that are predicted to be expressable in a subset of the 5 inbreds (Any_Inbred) and (iv) genes that are predicted to be silenced in all inbreds (No_Inbreds). (A) The cumulative number of expressable genes predicted as inbreds are added to the analysis, representing the pan-expressable gene set in maize. Distributions are derived based on the order in which inbreds are added. (B) Proportions of Each set that is likely TEs. (C-F) Distributions of various genomic features for each set. (G) All non-TE genes with ERC2 predictions were grouped into a network of clusters based on sequence similarity (all vs. all BLAST comparison). The data points making up these distributions are, for a given gene, the number of genes in the same cluster (sequence paralogs). (H) The proportions of each set that have observed mRNA, mRNA & Protein or no observations from [7]. (I) mRNA abundance distributions for observed mRNA from each set. (J) Protein abundance distributions for observed protein from each set. (K & J) The largest 15 clusters from the sequence similarity network (above) are represented (K). The total number of genes in each set is represented using various colors. Most are clustering around a known, common protein domain (x-axis labels). However, several are comprised mostly of “No_Inbred” genes and are clustering around uncharacterized protein sequences (Unknown 1-4). (L) The genes from all 4 sets are examined one at a time. The proportion of connections (each pair of genes in the same cluster is said to be connected) to the same set and the 3 other sets is displayed. For instance, we see that 61.4% of the connections of “All_Inbreds_Syntenic” genes are connected to other “All_Inbreds_Syntenic” genes.

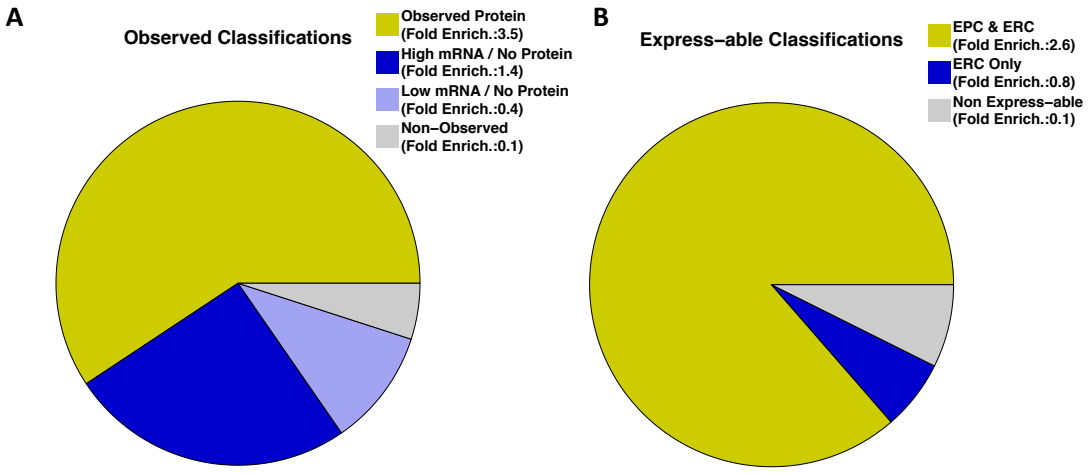

Fig. S11. CG gene body methylation. (A) Pie chart displaying the observed distribution of 9071 genes with CG gene body methylation patterns into various expression populations along with the fold enrichment ratio (Observed / Expected) for each set. (B) Pie chart displaying the distribution of Express-able Protein Classifier and Express-able RNA Classifier results for the 9071 genes with CG gene body methylation patterns. Note that no “EPC only” is shown because only 1 gene exists in this category and it is not gbM.

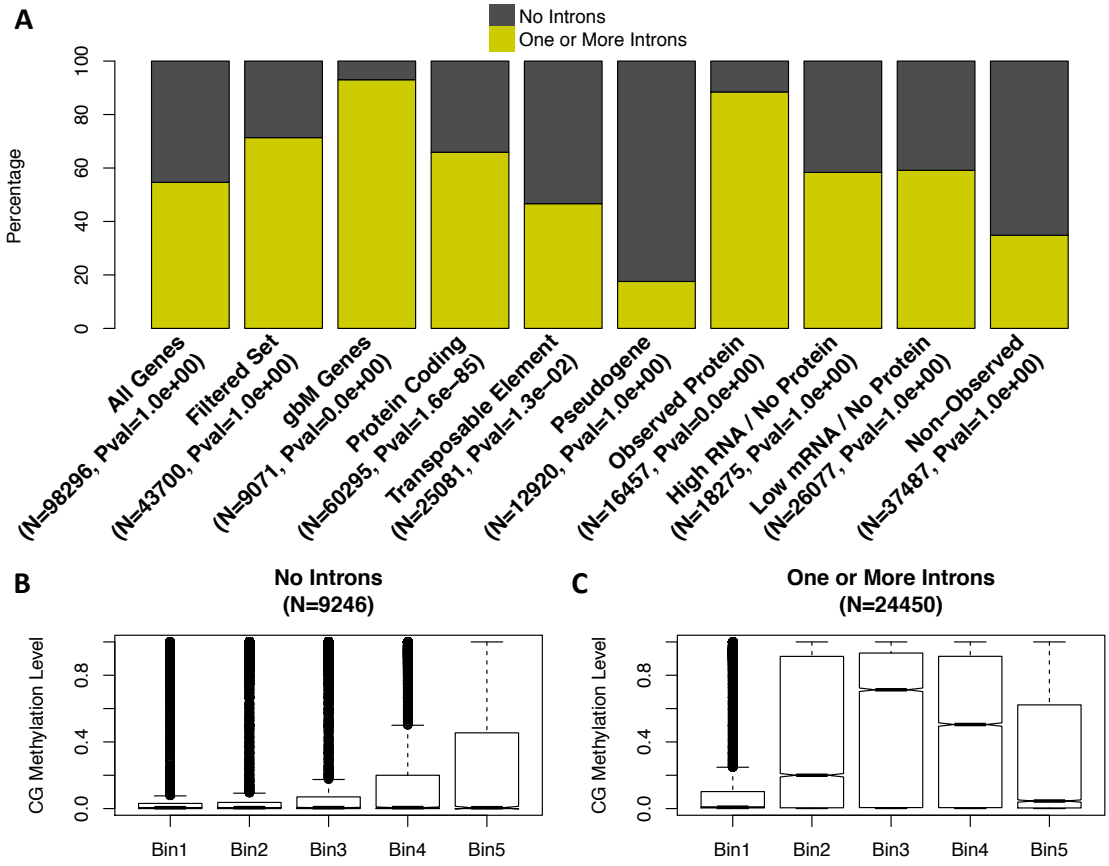

Fig. S12. Analysis of genes with one or more introns. (A) The relative proportions of intron-containing and non-intron genes in various pre-defined gene sets. “N” specifies the total number in that set. “Pval” specifies the *p*-values calculated for the significance of enrichment of intron-containing genes for each category relative to the filtered set, using a hypergeometric test and the upper tail. It should be noted that this calculation, by default, only includes genes present in the filtered set. Therefore, the ratios used in the *p*-value calculation will differ from those shown in the plot, which shows the proportions of the entire corresponding sets (i.e. relative to all genes, not subset on the filtered genes). (B) The distribution of CG methylation levels across all 5 bins for genes with high RNA and no introns. (C) The distribution of CG methylation levels across all 5 bins for genes with high RNA and one or more introns.

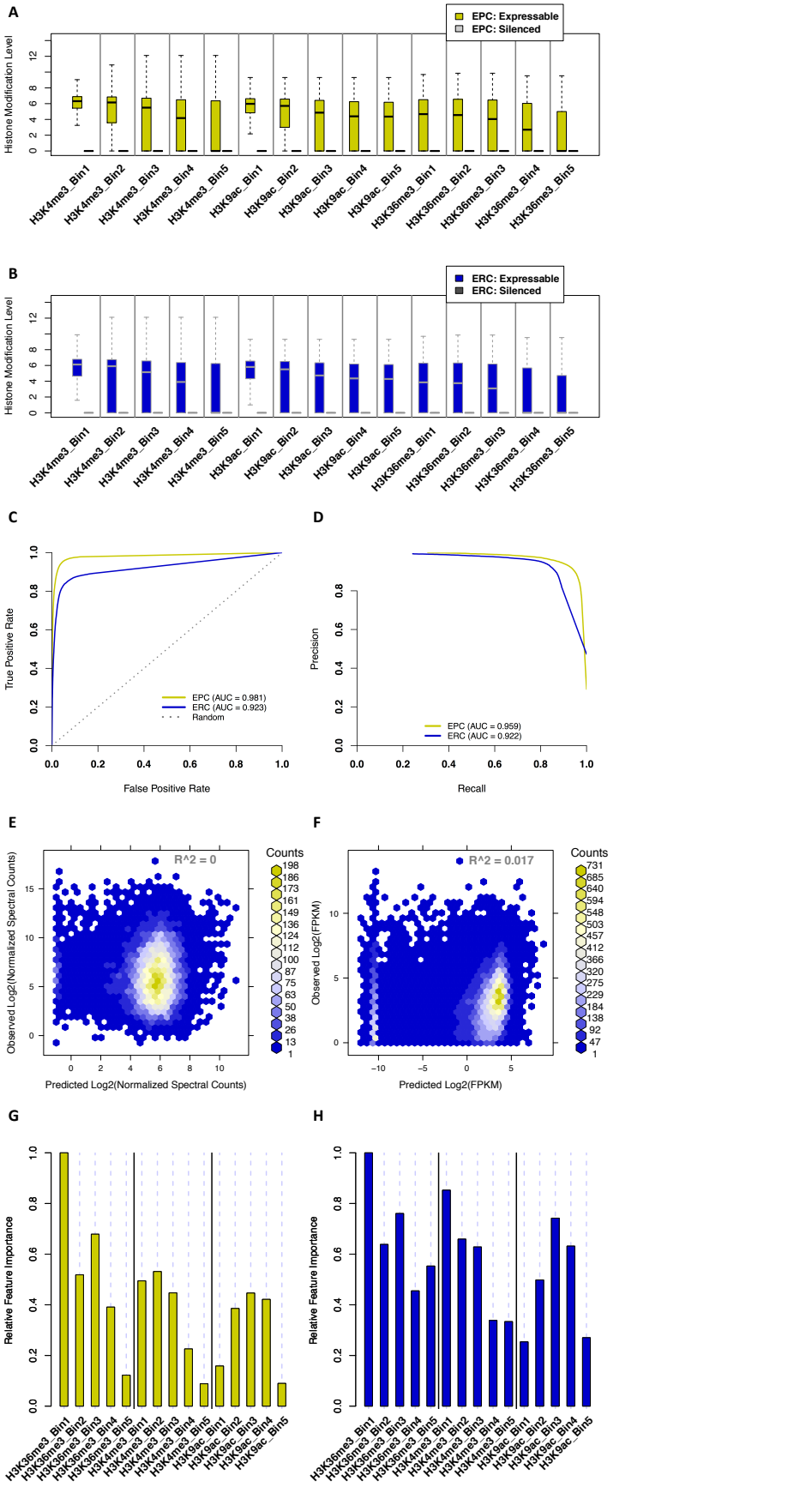
A
EPC: Expressable
EPC: Silenced
Histone Modification Level
H3K4me3_Bin1
H3K4me3_Bin2
H3K4me3_Bin3
H3K4me3_Bin4
H3K4me3_Bin5
H3K9ac_Bin1
H3K9ac_Bin2
H3K9ac_Bin3
H3K9ac_Bin4
H3K9ac_Bin5
H3K36me3_Bin1
H3K36me3_Bin2
H3K36me3_Bin3
H3K36me3_Bin4
H3K36me3_Bin5
B
ERC: Expressable
ERC: Silenced
C
True Positive Rate
False Positive Rate
EPC (AUC = 0.981)
ERC (AUC = 0.923)
Random
D
Precision
Recall
EPC (AUC = 0.959)
ERC (AUC = 0.922)
E
R^2 = 0
Observed Log2(Normalized Spectral Counts)
Predicted Log2(Normalized Spectral Counts)
Counts
F
R^2 = 0.017
Observed Log2(FPKM)
Predicted Log2(FPKM)
Counts
G
Relative Feature Importance
H
Relative Feature Importance

Fig. S13. Classification models built using data from 3 different histone modifications in maize B73 (H3K4me3, H3K36me3 and H3K9ac). (A-B) Boxplots showing the modification abundance in 5 bins across the whole gene body for genes that were classified as expressable or silenced in the EPC (A) and ERC (B). (C-D) Receiver Operating Characteristic (ROC) (C) and Precision vs. Recall (PR) (D) curves showing the accuracy of the histone modification-based classification models. (E-F) Scatter plots showing the accuracy of quantitative protein level predictions measured as Log2(Normalized Spectral Count) (E) or mRNA level predictions measured as Log2(FPKM) (F). (G-H) Signed feature importance measures for each predictive feature in the histone based EPR (G) and ERC (H).

Additional data (separate files)

Supplemental_Dataset_S01.csv Classification results from the 3 main classifiers (EPC, ERC and PFI). Each row is a gene and each column represents 1 model. The values are the proportion of votes for the positive class from the random forest classifier. For the corresponding training set genes, the cross-validated prediction is given. In general, a cutoff of 0.5 is used. Therefore, a score <= 0.5 represents a "silenced" prediction and a score >0.5 represents an "expressable" prediction. The last column lists the cross-referenced RefGen_V4 annotation when available.

Supplemental_Dataset_S02.csv A single column listing accessions for genes that are in the RefGen_V4 maize Filtered Gene Set (FGS).

Supplemental_Dataset_S03.xlsx (Tab: ERC-2_Scores) Classification results for predictions on various maize inbred lines and various B73 tissues using the ERC-2 classifier that was trained on a single B73 3rd leaf tissue data set. Each row is a gene and each column represents a different inbred line (3rd leaf) or a different B73 tissue. The values are the proportion of votes for the positive class from the random forest classifier. For the training set genes, the cross-validated prediction is given. In general, a cutoff of 0.5 is used. Therefore, a score <= 0.5 represents a "silenced" prediction and a score >0.5 represents an "expressable" prediction. (Tabs: All_Inbreds_Syntenic, All_Inbreds_NonSyntenic, Any_Inbred and No_Inbreds) The results of categorical enrichment carried out using the MapMan annotation on the four gene sets defined using ERC2 predictions. The four sets are: (i) genes predicted to be expressable in all inbreds examined and have syntenic orthologs with sorghum (All_Inbreds_Syntenic), (ii) or no synteny (All_Inbreds_NonSyntenic), (iii) genes that are predicted to be expressable in a subset of the 5 inbreds (Any_Inbred) and (iv) genes that are predicted to be silenced in all inbreds (No_Inbreds). Columns for enrichment results are: 1) Bin code and name of the functional class, 2) P-value of enrichment, 3) FDR adjusted P-value, 4) Fold enrichment over expected, 5) Enrichment direction (either over or under-enriched), 6) Total number of genes in the background (reference) set used for the analysis (here all genes with predictions in the ERC2s were used), 7) Total number of genes with the specified annotation in the background set, 8) Total number of genes used for the analysis in the test set (may vary from actual test number because some genes do not have MapMan annotations), 9) Number of genes with specified annotation in the test set. 10) A list of genes represented in column 9.

Supplemental_Dataset_S04.csv DNA methylation data for 14-day old B73 seedling quantified on the full set of genomic features including whole gene, 5' UTR, Transcription Start Site (TSS), Exon, Intron, and 3' UTR as well as all 3 methylation contexts (CG, CHG, CHH). Each row is a gene and each column is labeled as "Methylation Context _ Genomic Feature".

Supplemental_Dataset_S05.csv DNA methylation data for 14-day old B73 seedling quantified separately for each of the 5 bins as well as whole-gene, exon and intron features and both CG and CHG methylation contexts. Each row is a gene and each column is labeled as “Methylation Context _ Genomic Feature _ Bin Number”.

Supplemental_Dataset_S06.csv A list of Log2-transformed values of protein abundance for all detected proteins. Column 1 lists gene accessions and column 2 is the abundance value.

Supplemental_Dataset_S07.csv A single column listing accessions for genes that have at least one uniquely identified peptide.

Supplemental_Dataset_S08.csv A list of Log2-transformed values for mRNA abundance of all detected mRNAs. Column 1 lists gene accessions and column 2 is the abundance value.

Supplemental_Dataset_S09.csv A single column listing accessions for genes that are in the RefGen_V2 maize Filtered Gene Set (FGS).

Supplemental_Dataset_S10.csv A list of all maize genes and their biotypes. Column 1 shows gene accessions and column 2 lists one of 3 biotypes: (“protein_coding”,”transposable_element”,”pseudogene”).

Supplemental_Dataset_S11.csv A list of all maize genes and the number of introns they contain. Column 1 shows gene accessions and column 2 is the number of introns.

Supplemental_Dataset_S12.csv A single column listing accessions for genes that are likely to be transposable elements.

Supplemental Figshare Data
https://doi.org/10.6084/m9.figshare.8316701.v1
DNA methylation data for 3rd leaf from multiple different maize inbred lines. All data is quantified via 100 bp, non-overlapping windows. This table includes methylation quantification for all 3 contexts as well as feature coverage quantified over the 5 gene bins and 3 genomic features (gene, exon and intron). Each row is a gene and each column is labeled as “Inbred Line . (Methylation Context / Coverage) . Genomic Feature . Bin Number”.

Supplemental_Dataset_S13.csv DNA methylation data from multiple maize tissues (B73). All data is quantified via 100 bp, non-overlapping windows. This table includes methylation data for all 3 contexts, quantified over the 5 gene bins and 3 genomic features (gene, exon and intron). Each row is a gene and each

column is labeled as “B73.Tissue . Methylation Context . Genomic Feature . Bin Number”.

Supplemental_Dataset_S14.csv mRNA abundance (RNA-seq, Log2(FPM)) for the 3rd leaf from the same inbred lines as Supplemental Figshare Data. 109,871 RefGen_V2 genes are included. Each row is a gene and each column represents a different sample of 5 inbred lines and 3 replicates each.

Supplemental_Dataset_S15.csv mRNA abundance (RNA-seq, Log2(FPM)) for 3 different B73 tissues. 109,871 RefGen_V2 genes are included. Each row is a gene and each column represents a different sample of 3 tissues with 3 replicates each.

Supplemental_Dataset_S16.csv Quantification of 3 histone modifications in maize B73 across gene model bins. H3K36me3, H3K9Ac, and H3K4me3 are included. Each row represents a gene model and each column is designated by (ModificationType_Bin#).

Supplemental_Dataset_S17.csv List of various features associated with every gene with ERC2 predictions. (Group column) The gene set that each gene belongs to. Four sets of genes were defined using ERC2 predictions. The four sets are: (i) genes predicted to be expressable in all inbreds examined and have syntenic orthologs with sorghum (All_Inbreds_Syntenic), (ii) or no synteny (All_Inbreds_NonSyntenic), (iii) genes that are predicted to be expressable in a subset of the 5 inbreds (Any_Inbred) and (iv) genes that are predicted to be silenced in all inbreds (No_Inbreds). (Number_Of_AAs) The number of amino acids in the coding sequence of the gene model. (Number_Of_Introns) The number of introns in the gene model. (Number_of_Protein_Domains (Interprot5)) The number of protein domains found in the gene model using Interprot 5. (Top_Sorghum_BLAST_BitScore) The protein Blast bit score for the top hit to Sorghum proteins. (Syntenic_Ortholog_With_Sorghum) Does this gene have a syntenic ortholog with sorghum. (Predicted_Protein_Expression_In_B73) Is this gene predicted to express protein using the ERC model in B73. (Observed_Protein_Expression_In_B73 [7]) Does this gene have observed protein expression from [7]. (Best_Reciprocal_Blast_Maize_Paralog) If the gene has a reciprocal best BLAST hit to another maize gene, it is listed here. (Blast_Cluster) All non-TE genes were used to create a network based on protein sequence similarity (all vs. all BLAST) and then clustered into groups. This column lists the cluster membership for each gene. (Gene_Class) This is a category containing our final curation for each gene. There are five categories: (i) Expressable – The gene is predicted to be expressable in one or more observed inbreds. (ii) Likely TE – The gene has high sequence similarity to a transposable element. (iii) Silent_Clusters-With-Expressable – The gene is predicted to be silenced in all observed inbreds but >= 25% of the genes in its cluster are expressable, providing evidence for possible expression in other maize inbreds. (iv) Silent_Clusters-With-Silent – The gene is predicted to be silenced in all

observed inbreds and > 75% of the genes in its cluster are also silent. (v) Silent_No_Cluster – The gene is predicted to be silenced in all observed inbreds and has very low sequence similarity with any other gene.

Supplemental_Dataset_S18 A source code text file in the format of the R programming language. This file contains all the source code that was used to run the model construction, analysis and figure plotting for this publication.

Supplemental_Dataset_S19 A source code text file in the format of the R programming language. This file contains auxiliary functions that are called by Supplemental_Dataset_S18 source code.

References

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2018). Vienna, Austria. URL https://www.R-project.org/.

2 Rstudio Team. RStudio: Integrated Development for R. (2015).

3. Li, Q. et al. Genetic perturbation of the maize methylome. *The Plant cell* **26**, 4602–4616 (2014). [PMID:25527708] http://doi.org/10.1105/tpc.114.133140

4. Krueger, F. Trim Galore!. (2007). Babraham bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

5. Krueger, F. et al. Bismark : a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011). http://doi.org/10.1093/bioinformatics/btr167

6. Schultz, M. D. et al. “Leveling” the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28**, 583–585 (2012). [PMID:23131467] http://doi.org/10.1016/j.tig.2012.10.012

7. Walley, J. W. et al. Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016). http://doi.org/10.1126/science.aag1125

8. Portwood, J. L. et al. MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Research* **47**, D1146–D1154 (2018). http://doi.org/doi.org/10.1093/nar/gky1046

9. Schnable, P. S. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**, 1112–1115 (2009). http://doi.org/10.1126/science.1178534

10. Li, P. et al. The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* **42**, 1060–1067 (2010). http://doi.org/10.1038/ng.703

11. Wessler, S. R. et al. Maize transposable element database http://maizetedb.org/~maize/

12. Li, Q. et al. Examining the causes and consequences of context-specific differential DNA methylation in maize. *Plant Physiology* **168**, 1262–1274 (2015). http://doi.org/10.1104/pp.15.00052

13. Li, Q. et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proceedings of the National Academy of Science* **112**, 14728–14733 (2015). http://doi.org/10.1073/pnas.1514680112

14. Eichten, S. R. et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. *The Plant cell* **25**, 2783–2797 (2013). http://doi.org/10.1105/tpc.113.114793

15. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics* (2010).

16. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011). http://doi.org/10.14806/ej.17.1.200

17. Kim, D. et al. HISAT : a fast spliced aligner with low memory requirements. *Nature Methods* **12**, (2015). http://doi.org/10.1038/nmeth.3317

18. Anders, S. et al. Genome analysis HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2018). http://doi.org/10.1093/bioinformatics/btu638

19. He, G. et al. Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biology* **14**, 1–15 (2013). http://doi.org/10.1186/gb-2013-14-6-r57

20. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001). [PMID:21816105] http://doi.org/10.1023/A:1010933404324

21. Liaw, A. et al. Classification and regression by randomForest. *R news* **2**, 18–22 (2002). [PMID:21196786]

22. Sing, T. et al. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005). http://doi.org/10.1093/bioinformatics/bti623

23. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006). http://doi.org/10.1016/j.patrec.2005.10.010

24. Breiman, L. OUT-OF-BAG ESTIMATION. *Technical Report: Department of Statistics: UC Berkeley* 1–13 (1996).

25. Lohse, M. et al. Mercator : a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell & Environment* **37**, 1250–1258 (2014). http://doi.org/10.1111/pce.12231

26. van Dongen, S.M. Graph clustering by flow simulation. *PhD thesis, University of Utrecht* **1**, (2000). http://doi.org/10.1016/j.cosrev.2007.05.001