A Model-Based Approach for Identifying Functional Intergenic Transcribed Regions and Noncoding RNAs

John P. Lloyd,[†,1] Zing Tsung-Yeh Tsai,[2] Rosalie P. Sowers,[3] Nicholas L. Panchy,[‡,4] and Shin-Han Shiu*[,1,4,5]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI

[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[3]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA

[4]Genetics Program, Michigan State University, East Lansing, MI

[5]Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI

[†]Present address: Departments of Human Genetics and Internal Medicine, University of Michigan, Ann Arbor, MI

[‡]National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN

***Corresponding author**: E-mail: shius@msu.edu.

Associate editor: Jun Gojobori

Abstract

With advances in transcript profiling, the presence of transcriptional activities in intergenic regions has been well established. However, whether intergenic expression reflects transcriptional noise or activity of novel genes remains unclear. We identified intergenic transcribed regions (ITRs) in 15 diverse flowering plant species and found that the amount of intergenic expression correlates with genome size, a pattern that could be expected if intergenic expression is largely nonfunctional. To further assess the functionality of ITRs, we first built machine learning models using *Arabidopsis thaliana* as a model that accurately distinguish functional sequences (benchmark protein-coding and RNA genes) and likely nonfunctional ones (pseudogenes and unexpressed intergenic regions) by integrating 93 biochemical, evolutionary, and sequence-structure features. Next, by applying the models genome-wide, we found that 4,427 ITRs (38%) and 796 annotated ncRNAs (44%) had features significantly similar to benchmark protein-coding or RNA genes and thus were likely parts of functional genes. Approximately 60% of ITRs and ncRNAs were more similar to nonfunctional sequences and were likely transcriptional noise. The predictive framework established here provides not only a comprehensive look at how functional, genic sequences are distinct from likely nonfunctional ones, but also a new way to differentiate novel genes from genomic regions with noisy transcriptional activities.

Key words: **intergenic transcription, ncRNAs, definition of function, molecular evolution, machine learning, data integration.**

Article

Introduction

Advances in sequencing technology have helped to identify pervasive transcription in intergenic regions with no annotated genes. These intergenic transcripts have been found in metazoa and fungi, including human (ENCODE Project Consortium 2012), *Drosophila melanogaster* (Brown et al. 2014), *Caenorhabditis elegans* (Boeck et al. 2016), and *Saccharomyces cerevisiae* (Nagalakshmi et al. 2008). In plants, 7,000–15,000 intergenic transcripts have also been reported in *Arabidopsis thaliana* (Yamada et al. 2003; Stolc et al. 2005; Moghe et al. 2013; Krishnakumar et al. 2015) and *Oryza sativa* (Nobuta et al. 2007). The presence of intergenic transcripts indicates that there may be additional genes (genomic regions generating functional RNA and/or protein products) that have escaped gene finding efforts thus far, including those that function as RNA genes (Simon and Meyers 2011; Guil and Esteller 2012; Fei et al. 2013; Palazzo and Lee 2015; Tan et al. 2015). Meanwhile, it is also possible that some of these intergenic transcripts are products of un-regulated noise (Struhl 2007; Palazzo and Gregory 2014). Given the functional significance of most intergenic transcripts remains unclear, the identification of functional intergenic transcribed regions (ITRs) represents a fundamental task that is critical to our understanding of genome evolution.

Loss-of-function study represents the gold standard by which the functional significance of genomic regions, including ITRs, can be confirmed (Ponting and Belgard 2010; Niu and Jiang 2013). In *Mus musculus* (mouse), at least 25 ITRs with loss-of-function mutant phenotypes have been identified (Sauvageau et al. 2013; Lai et al. 2015). In human, 162 long intergenic noncoding RNAs harbor phenotype-associated SNPs (Ning et al. 2013). In addition to intergenic expression, most model organisms feature an abundance of annotated noncoding RNA (ncRNA) sequences (Zhao et al. 2016), which are mostly identified through the presence of transcriptional evidence occurring outside of annotated protein-coding genes. Thus, the only difference between ITRs and most ncRNA sequences is whether or not they have been annotated. Similar to the ITR examples above, a small number of ncRNAs have been confirmed as functional through loss-of-function experiments including *Xist* in mouse (Penny et al.

1996; Marahrens et al. 1997), *Malat1* in human (Bernard et al. 2010), *bereft* in *D. melanogaster* (Hardiman et al. 2002), and *At4* in *A. thaliana* (Shin et al. 2006).

However, the number of ITRs and ncRNAs with well-established functions is dwarfed by those without functional evidence. Whereas some ITRs and ncRNAs can be novel genes, intergenic transcription may also be the byproduct of noisy transcription that can occur due to nonspecific landing of RNA Polymerase II (RNA Pol II) or spurious regulatory signals that drive expression in random genomic regions (Struhl 2007). In the ENCODE project (ENCODE Project Consortium 2012), ~80% of the human genome was defined as biochemically functional as reproducible biochemical activities, for example, transcription, could be detected. This has drawn considerable critique because the existence of a biochemical activity is not an indication of selection (Eddy 2013; Graur et al. 2013; Niu and Jiang 2013). Similarly, arguments based on genetic load indicate that no more than 25% of the human genome may be functional and experiencing sequence constraint (Comings 1972; Graur 2017). Instead, it is advocated that a genomic region with discernible activity is only functional if it is under selection (Amundson and Lauder 1994; Graur et al. 2013; Doolittle et al. 2014). Under this "selected effect" functionality definition, ITRs and most annotated ncRNA genes remain functionally ambiguous.

Because of the debate on the definitions of function postENCODE, Kellis et al. (2014) suggested that evolutionary, biochemical, and genetic evidences provide complementary information to define functional genomic regions. Consistent with this, integration of biochemical and conservation evidence was successful in identifying regions in the human genome that are under selection (Gulko et al. 2014) and classification of human disease genes and pseudogenes (Tsai et al. 2017). In this study, we adopt a similar framework to investigate if intergenic transcription reflects the activity of genes, including RNA genes (e.g., microRNAs), in plants. We first identified ITRs in 15 flowering plant species with 17-fold genome size differences and evaluated the relationship between the prevalence of intergenic expression and genome size. Next, we established machine learning models using *A. thaliana* data to predict likely functional ITRs and ncRNAs based on 93 evolutionary, biochemical, and sequence-structure features. Finally, we applied the models to ITRs and annotated ncRNAs to determine whether these functionally ambiguous sequences are more similar to benchmark functional or likely nonfunctional sequences.

Results and Discussion

Genome Size versus Prevalence of Intergenic Transcripts Indicates ITRs May Generally Be Nonfunctional

Transcription of an unannotated, intergenic region could be due to nonfunctional transcriptional noise or the activity of a novel gene. If noisy transcription occurs due to random landing of RNA Pol II or spurious regulatory signals, a naïve expectation is that, as genome size increases, the total nucleotides covered by ITRs would increase accordingly. This is because the additional genomic regions may provide additional space for nonspecific landing of RNA polymerase II and/or for spurious regulatory signals. By contrast, we find that there is no significant correlation between genome size and the number of annotated genes among 50 plant species ($r^2 = 0.01$; $P = 0.56$; Michael and Jackson 2013). Thus, we expect that the extent of expression for genic sequences will not be significantly correlated with genome size.

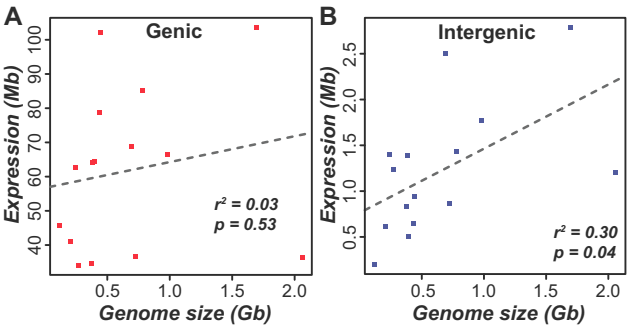

FIG. 1. Relationship between genome size and number of nucleotides covered by RNA-seq reads (expression) in 15 flowering plant species. (*A*) Annotated genic regions. (*B*) Intergenic regions. Transcribed regions were considered as intergenic if they did not overlap with any gene annotation and had no significant translated sequence similarity to plant protein sequences. Identical numbers of RNA-sequencing reads (30 million) and the same mapping procedures were used in all species. Mb: megabase. Gb: gigabase. Dotted lines: Linear model fits. r^2: Square of Pearson's correlation coefficient.

To gauge if ITRs generally behave more like what we expect of noisy or genic transcription, we first identified genic and intergenic transcribed regions using leaf transcriptome data from 15 flowering plants with 17-fold differences in genome size (supplementary table 1, Supplementary Material online). As expected, the coverage of expression originating from annotated genic regions had no significant correlation with genome size ($r^2 = 0.03$; $P = 0.53$; fig. 1*A*). In contrast, the length of genomic sequences covered by intergenic expression and genome sizes were significantly and positively correlated ($r^2 = 0.30$; $P = 0.04$; fig. 1*B*), consistent with the interpretation that a significant proportion of intergenic expression represents transcriptional noise. This relationship is maintained when three species (*Glycine max*, *Zea mays*, and *Panicum virgatum*) that experienced whole genome duplication events <15 million years ago (Ma) are removed ($r^2 = 0.28$, $P = 0.08$) or confounding effects of phylogenetic nonindependence on the relationships between genome size and expression coverage are taken into consideration ($r^2 = 0.46$, $P < 0.008$; see Materials and Methods). However, the correlation between genome size and intergenic expression explained ~30% of the variation (fig. 1*B*), suggesting that other factors also affect ITR content, including the possibility that some ITRs are truly functional, novel genes. To further evaluate the functionality of intergenic transcripts, we next identified the biochemical and evolutionary features of functional genic regions and tested whether intergenic transcripts in *A. thaliana* were more similar to functional or nonfunctional sequences.

Benchmark Functional Protein-Coding and Nonfunctional Genomic Sequences Are Significantly Distinct in Multiple Features

To determine whether intergenic transcripts resemble functional sequences, we first asked what features allow benchmark functional protein-coding and nonfunctional genomic regions to be distinguished in the model plant *A. thaliana*. For benchmark functional sequences, we used protein-coding genes with visible loss-of-function phenotypes when mutated (referred to as phenotype genes, $n = 1{,}876$; see Materials and Methods). Because their mutations have significant growth and/or developmental impact and likely contribute to reduced fitness, these phenotype genes can be considered functional under the selected effect definition (Neander 1991). For benchmark nonfunctional genomic regions, we utilized pseudogene sequences ($n = 761$; see Materials and Methods). Considering that only 2% of pseudogenes are maintained over 90 million years (My) of divergence between human and mouse (Svensson et al. 2006), it is expected that the majority of pseudogenes is no longer under selection (Li et al. 1981).

We evaluated 93 gene or gene product features for their ability to distinguish between phenotype genes and pseudogenes. These features were grouped into seven categories, including chromatin accessibility, DNA methylation, histone 3 (H3) marks, sequence conservation, sequence-structure, transcription factor (TF) binding, and transcription activity (supplementary table 2, Supplementary Material online). We emphasize that no features were exclusive to protein-coding sequences. We used Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) as a metric to measure how well a feature distinguished between phenotype genes and pseudogenes, which ranges between 0.5 (random guessing) and 1 (perfect separation of functional and nonfunctional sequences). Among the seven feature categories, transcription activity features were highly informative (median AUC-ROC = 0.88; fig. 2*A*). Despite the strong performance of transcription activity-related features, the presence of expression (i.e., transcript evidence) was a poor predictor of functionality (AUC-ROC = 0.58; fig. 2*A*). This is because 80% of pseudogenes were considered expressed in $\geq$1 of 51 RNA-seq data sets, demonstrating that presence of transcripts should not be used by itself as evidence of functionality. Sequence conservation, DNA methylation, TF binding, and H3 mark features were also fairly distinct between phenotype genes and pseudogenes (median AUC-ROC ~0.7 for each category; fig. 2*B*–*E*). In contrast, chromatin accessibility and sequence-structure features were largely uninformative (median AUC-ROC = 0.51 and 0.55, respectively; fig. 2*F*, *G*). We also observed high performance variability within feature categories (see Supplementary material, Supplementary Material online). Whereas many features are distinct between phenotype genes and pseudogenes, functional predictions based on single features yield high error rates (supplementary table 3; Supplementary material, Supplementary Material online), indicating a need to jointly consider multiple features for distinguishing phenotype genes and pseudogenes.

Consideration of Multiple Features Produces Accurate Predictions of Functional Genomic Regions

To consider multiple features in combination, we first conducted principle component (PC) analysis and found that phenotype genes (supplementary fig. 1*A*, Supplementary Material online) and pseudogenes (supplementary fig. 1*B*, Supplementary Material online) were distributed in largely distinct space but with substantial overlap, indicating that standard parametric approaches are not well suited to distinguishing between benchmark functional and nonfunctional sequences. We next integrated all 93 features to establish a machine learning model distinguishing phenotype gene and pseudogenes (referred to as the full model; see Materials and Methods). The full model provided more accurate predictions (AUC-ROC = 0.98; False Negative Rate [FNR] = 4%; False Positive Rate [FPR] = 10%; fig. 3*A*) than any individual feature (fig. 2; supplementary table 3, Supplementary Material online). An alternative measure of performance based on the precision (proportion of predicted functional sequences that are truly functional) and recall (proportion of truly functional sequences predicted correctly) values also indicated that the model was performing well (fig. 3*B*). When compared with the best-performing single feature (expression breadth), the full model had a similar FNR but half the FPR (10% compared with 21%). Thus, the full model is highly capable of distinguishing between phenotype genes and pseudogenes.

We next determined the relative contributions of different feature categories in predicting phenotype genes and pseudogenes and established seven prediction models each using only the subset of features from a single category (fig. 2). Although none of these category-specific models had performance as high as the full model (fig. 3*A*), the transcription activity feature category model performed almost as well as the full model (AUC-ROC = 0.97, FNR = 6%, FPR = 12%). Instead of the presence of expression evidence, the breadth and level of transcription are the causes of the strong performance of the transcription activity-only model. We also found that a model excluding transcription activity features (full [-TX], fig. 3*A* and *B*) performed almost as well as the full model and similarly to the transcription activity-feature-only model, but with an increased FPR (AUC-ROC = 0.96; FNR = 3%; FPR = 20%). These findings indicate that a diverse array of features can be considered jointly to make highly accurate predictions of the functionality of a genomic sequence. Meanwhile, our finding of the high performance of the transcription activity-only model highlights the possibility of establishing an accurate model for functional prediction in species with only a modest amount of transcriptome data.

The Functional Likelihood (FL) Measure Can Be Used to Classify Functional and Nonfunctional Sequences

To provide a measure of the potential functionality of any sequence in the *A. thaliana* genome, including ITRs and ncRNAs, we utilized the confidence score from the full model

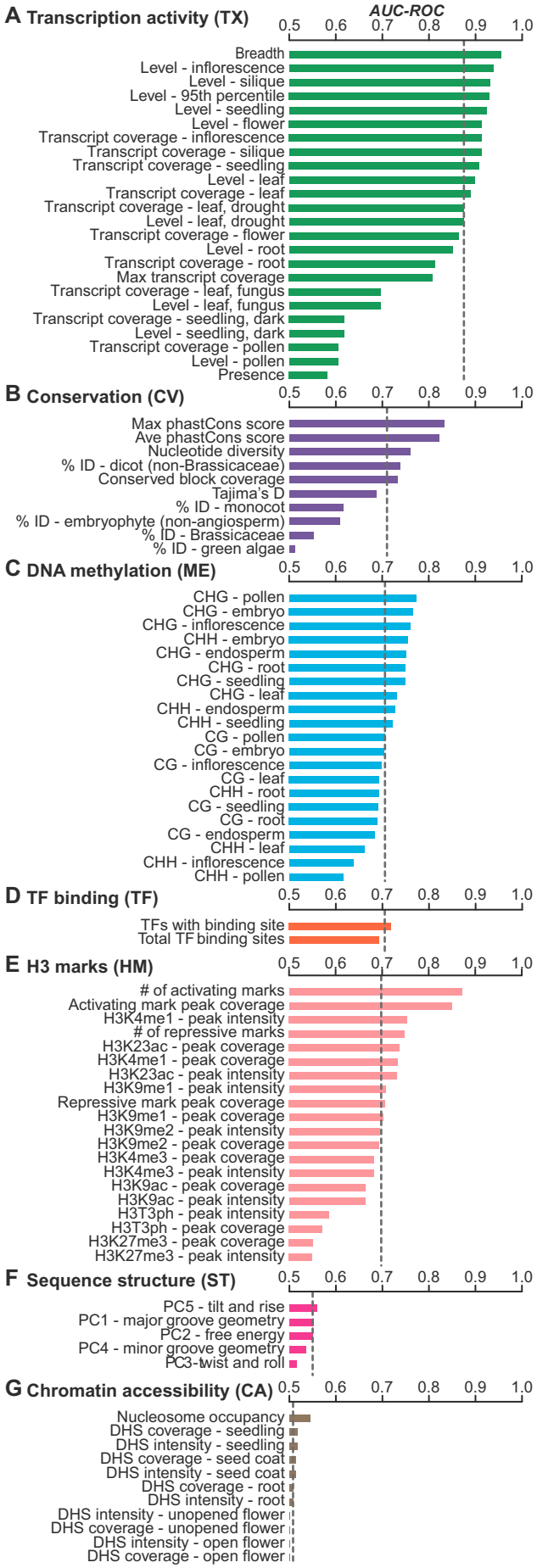

Fig. 2. Predictions of functional (phenotype gene) and nonfunctional (pseudogene) sequences based on each individual feature. Prediction performance is measured using Area Under the Curve-Receiver Operating Characteristic (AUC-ROC). AUC-ROC values range between 0.5 (random) and 1 (perfect separation), with AUC-ROC values of 0.7, 0.8, and 0.9 considered fair, good, and excellent performance, respectively. Features include those in the categories of (*A*) transcription activity, (*B*) sequence conservation, (*C*) DNA methylation, (*D*) transcription factor (TF) binding, (*E*) histone 3 (H3) marks, (*F*) sequence structure, and (*G*) chromatin accessibility. Dotted lines: Median AUC-ROC of features in a category.

as a "FL" value (Tsai et al. 2017). The FL score ranges between 0 and 1, with high values indicating that a sequence is more similar to phenotype genes (functional) and low values indicating a sequence more closely resembles pseudogenes (nonfunctional). FL values for all genomic regions examined in this study are available in supplementary table 4, Supplementary Material online. As expected, phenotype genes had high FL values (median = 0.97; fig. 4*A*) and pseudogenes had low values (median = 0.01; fig. 4*B*). To call sequences as functional or not, we defined a threshold FL value (0.35) by maximizing the *F*-measure (see Materials and Methods). Using this threshold, 96% of phenotype genes (fig. 4*A*) and 90% of pseudogenes (fig. 4*B*) are correctly classified as functional and nonfunctional, respectively, demonstrating that the full model is highly capable of distinguishing functional and nonfunctional sequences.

We next applied our model to predict the functionality of annotated protein-coding genes, transposable elements (TEs), and unexpressed intergenic regions. Most annotated protein-coding genes not included in the phenotype gene data set had high FL scores (median = 0.86; fig. 4*C*) and 80% were predicted as functional. The features exhibited by low-scoring protein-coding genes and high-scoring pseudogenes are discussed in Supplementary material, Supplementary Material online. Among putatively nonfunctional sequences, the FLs were low for both TEs (median = 0.03, fig. 4*D*) and unexpressed intergenic regions (median = 0.07; fig. 4*E*), and 99% of TEs and all unexpressed intergenic sequences were predicted as nonfunctional. We should emphasize that, using the criteria set out by the ENCODE Project Consortium (2012), 92% of pseudogenes, 85% of TEs, and 97% of randomly sampled, unexpressed intergenic sequences would be considered biochemically functional (≥2 signatures derived from transcription activity, H3 marks, TF binding, or DNase I hypersensitivity). We find that the FL measure provides a useful metric to distinguish between phenotype genes and pseudogenes. In addition, the FLs of annotated protein-coding genes, TEs, and unexpressed intergenic sequences agree with a priori expectations regarding the functionality of these sequences.

Most ITRs and Annotated ncRNAs Do Not Resemble Benchmark Phenotype Genes

We next applied the full model to 895 unannotated ITRs, 136 ncRNAs annotated by The Arabidopsis Information Resource

(TAIR), and 252 long ncRNAs annotated by the Araport database that do not overlap with any other annotated genome features. Note that the primary difference between ITRs and ncRNAs is whether they have been identified by annotation projects. Consistent with previous studies (Moghe et al. 2013), ITRs and ncRNAs in our data set were more narrowly and weakly expressed and less conserved compared with phenotype genes (supplementary fig. 2*A* and *B*, Supplementary Material online). In particular, ITRs had H3 mark and DNA methylation patterns that were generally more similar to pseudogenes (supplementary fig. 2*C*–*F*, Supplementary Material online) compared with phenotype genes. When the full prediction model was applied to these sequences, the median FLs were low (0.09) for both ITRs (fig. 4*F*) and Araport ncRNAs (fig. 4*G*), and only 15% and 9% of these sequences were predicted as functional, respectively. By contrast, TAIR ncRNAs had a significantly higher median FL value (0.53; *U* tests, both $P < 5e\text{-}31$; fig. 4*H*) and 68% were predicted as functional, which is best explained by differences in features from the transcription activity category (fig. 5). We also note that ITRs and annotated ncRNAs that were close to genes were more frequently predicted as functional (supplementary fig. 3, Supplementary Material online), suggesting a subset may represent unannotated exons of known genes or that the regions proximal to genes are disproportionately covered by independent functional sequences. Alternatively, the accessible and active chromatin states of nearby genes may represent a confounding factor for prediction models. Given the challenge in ascertaining the origin and likely functionality of ITRs/ncRNAs proximal to genes, we instead conservatively estimate that 50 ITRs (9%) and 60 annotated ncRNAs (23%) that are >456 bp from nearby genes (95th percentile of annotated intron lengths) and predicted as functional may represent parts of novel genes.

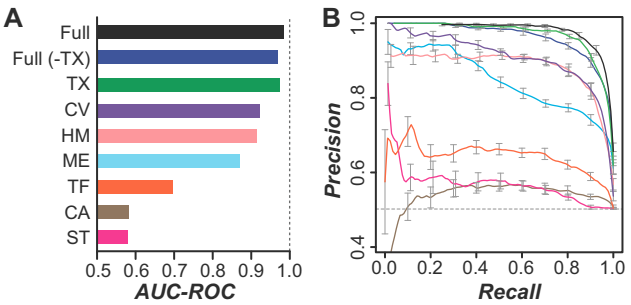

Fig. 3. Predictions of functional and nonfunctional sequences based on multiple features. (*A*) AUC-ROC values of function prediction models built when considering all features (Full), all except transcription activity (TX)-related features (Full [-TX]), and all features from each category. The category abbreviations follow those in figure 2. (*B*) Precision-recall curves of the models with matching colors from (*A*). The models were built using feature values calculated from 500 bp sequence windows.

Given the association between transcription activity features and functional predictions (figs. 2*A* and 3*A*), we considered that the low proportion of ITRs and annotated ncRNAs predicted as functional may be due to the full model being biased against conditionally functional or narrowly expressed sequences. We defined conditionally functional genes as those that exhibited a mutant phenotype under nonstandard (i.e., stress) conditions, but exhibited no phenotype under standard growth conditions. Genes with conditional phenotypes had no significant differences in FLs (median = 0.96) as those with phenotypes under standard growth conditions

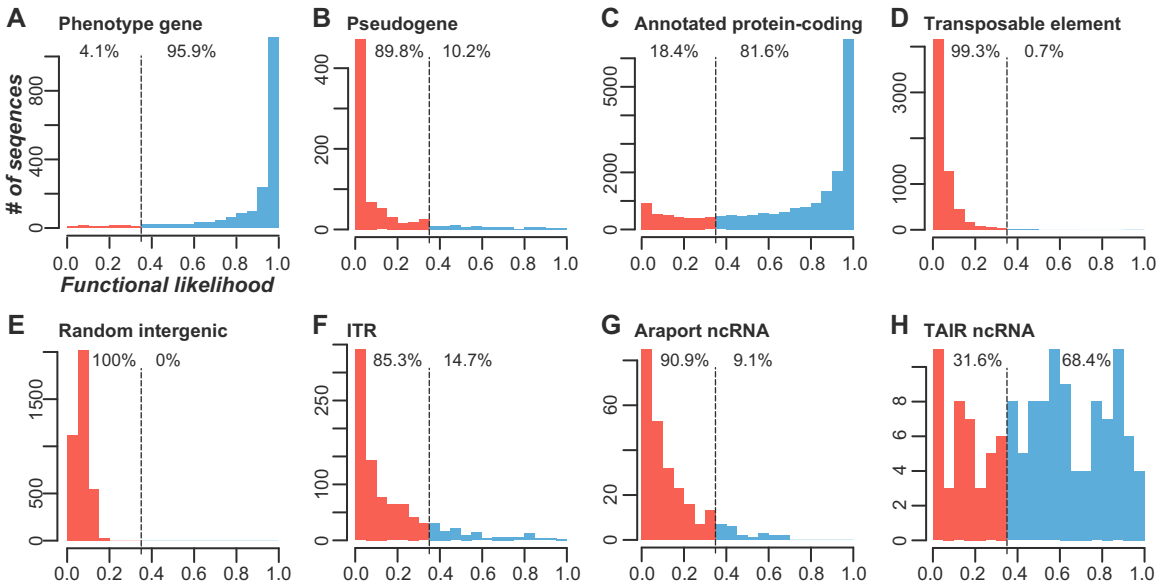

Fig. 4. Functional likelihood distributions of various sequence classes based on the full model. (*A*) Phenotype genes. (*B*) Pseudogenes. (*C*) Annotated protein-coding genes. (*D*) Transposable elements. (*E*) Random unexpressed intergenic sequences. (*F*) Intergenic transcribed regions (ITR). (*G*) Araport11 ncRNAs. (*H*) TAIR10 ncRNAs. The full model was established using 500 bp sequence windows. Higher and lower functional likelihood values indicate greater similarity to phenotype genes and pseudogenes, respectively. Vertical dashed lines indicate the threshold for calling a sequence as functional or nonfunctional. The percentages to the left and right of the dashed line indicate the percent of sequences predicted as functional or nonfunctional, respectively.

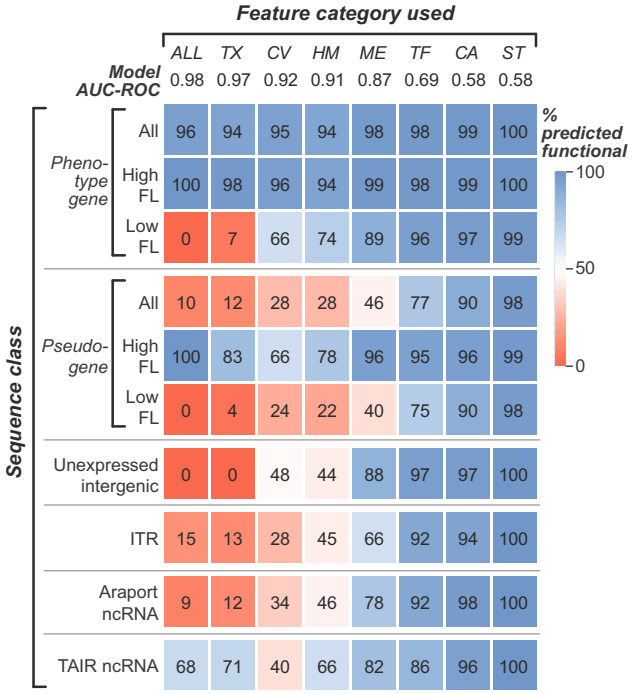

Fig. 5. Proportion of phenotype genes, pseudogenes, ITRs, and ncRNAs predicted as functional in the full and single-category models. Percentages of sequence classes that are predicted as functional in models based on all features and the single category models, each using all features from a category (abbreviated according to fig. 2). The models are sorted from left to right based on performance (AUC-ROC). The colors of and numbers within the blocks indicate the proportion sequences predicted as functional by a given model. Phenotype gene and pseudogene sequences are shown in three subgroups: All sequences (All), and those predicted as functional (high functional likelihood [FL]) and nonfunctional (low FL) in the full model. ITR: intergenic transcribed regions. A greater proportion of ITRs and Araport ncRNAs are predicted as functional when considering only DNA methylation or H3 mark features compared with the full (fig. 3) or tissue-agnostic (supplementary fig. 5, Supplementary Material online) models. However, these two category-specific models also had higher false positive rates (unexpressed intergenic sequences and pseudogenes). Thus, these single feature-category models do not provide additional support for the functionality of most Araport ncRNAs and ITRs.

(median = 0.97; *U* test, $P = 0.38$, supplementary fig. 4*A*, Supplementary Material online), indicating the full model can capture conditionally functional sequences. However, the full model is biased against narrowly expressed ($\leq$3 tissues) phenotype genes as 65% of them were predicted as nonfunctional (supplementary fig. 4*B*, Supplementary Material online). Further, pseudogenes that were more highly and broadly expressed were disproportionately predicted as functional (fig. 5; supplementary fig. 4*B*, Supplementary Material online). To tailor functional predictions to narrowly expressed sequences, particularly ITRs and ncRNAs, we generated a "tissue-agnostic" model by excluding expression breadth and features available across multiple tissues (see Materials and Methods). In addition, in the tissue-agnostic model tissue-specific features were replaced with summary features (e.g., maximum expression level across 51 RNA-seq data sets). This tissue-agnostic model performed similarly to the full model (AUC-ROC = 0.97; FNR = 4%; FPR = 15%; supplementary fig. 5; supplementary table 4, Supplementary Material online), although there was a 5% increase in FPR (from 10% to 15%). Importantly, the proportion of phenotype genes expressed in $\leq$3 tissues predicted as functional increased by 23% (35% in the full model to 58% in the tissue-agnostic model; supplementary fig. 4*C*, Supplementary Material online), indicating that the tissue-agnostic model is more suitable for predicting the functionality of narrowly expressed sequences than the full model.

Next, we applied the tissue-agnostic model to ITRs and TAIR/Araport ncRNAs. Compared with the full model, around twice as many ITRs (30%) and Araport ncRNAs (19%) but a similar number of TAIR ncRNA (67%) were predicted as functional. Considering the union of the full and tissue-agnostic model predictions, 268 ITRs (32%), 57 Araport ncRNAs (23%), and 105 TAIR ncRNAs (77%) were likely functional. Thus, the majority of ITRs and Araport ncRNAs is more similar to pseudogenes than to phenotype genes that are predominantly protein coding.

Benchmark Protein-Coding and RNA Genes Exhibit Distinct Characteristics

We demonstrated that the majority of ITR and annotated ncRNA sequences does not exhibit characteristics of benchmark phenotype genes. Note that the phenotype genes are predominantly protein coding. Although the features utilized to generate functional predictions were not exclusive to protein-coding sequences, RNA genes may exhibit a distinct feature profile from protein-coding genes. The full and tissue agnostic models described above were established with 500 bp windows and most known RNA genes are too short to be considered by these models. Thus, to evaluate functional predictions among annotated RNA genes, we generated a new tissue-agnostic model using 100 bp sequences (for features, see supplementary table 5, Supplementary Material online) that performed similarly to the full 500 bp model, except with 9% higher FNR (AUC-ROC = 0.97; FNR = 13%; FPR = 5%; supplementary fig. 6, Supplementary Material online). With this new tissue agnostic model, 50% (three out of six) of RNA genes with documented mutant phenotypes (phenotype RNA genes) were predicted as functional (supplementary fig. 6*I*, Supplementary Material online). We also applied this model to other RNA Pol II-transcribed RNA genes (without documented phenotypes) and found that 15% of microRNA (miRNA) primary transcripts (supplementary fig. 6*J*, Supplementary Material online), 73% of small nucleolar RNAs (snoRNAs; supplementary fig. 6*K*, Supplementary Material online), and 50% of small nuclear RNAs (snRNAs; supplementary fig. 6*L*, Supplementary Material online) were predicted as functional. Although the proportion of phenotype RNA genes predicted as functional (50%) is significantly higher than the proportion of pseudogenes predicted as functional (5%, FET, $P < 0.004$), this finding suggests that a model trained using protein-coding genes has a substantial FNR for detecting RNA genes. In contrast, functional human ncRNAs could be accurately predicted by a model generated with

protein-coding sequences (Tsai et al. 2017). Thus, in plants, the idiosyncrasies of RNA genes cannot be adequately captured by evaluating protein-coding sequences.

To determine whether the suboptimal predictions by the phenotype protein-coding gene-based models are because RNA genes belong to a class of their own, we next built a multi-class function prediction model (as opposed to the binary, two-class models described above) aimed at distinguishing four classes of sequences: Benchmark RNA genes ($n = 46$, supplementary table 5, Supplementary Material online), phenotype protein-coding genes (1,882), pseudogenes (3,916), and randomly selected, unexpressed intergenic regions (4,000). In the four-class model, 87% of benchmark RNA genes, including all six phenotype RNA genes, were predicted as functional sequences (65% RNA gene-like and 22% phenotype protein-coding gene-like; fig. 6A). In addition, 95% of phenotype protein-coding genes were predicted as functional (fig. 6B), including 80% of narrowly expressed genes, an increase of 22% over the 500 bp tissue-agnostic model (supplementary fig. 4C, Supplementary Material online). For benchmark nonfunctional sequences, 70% of pseudogenes (fig. 6C) and 100% of unexpressed intergenic regions (fig. 6D) were predicted as nonfunctional (either as pseudogenes or unexpressed intergenic sequences). Overall, the four-class model improves prediction accuracy of RNA genes and narrowly expressed genes. In addition, given the 30% FPR among pseudogenes, the four-class model provides a liberal estimate of sequence functionality and high confidence estimate of nonfunctionality.

Most ITRs and Annotated ncRNAs Do Not Resemble Benchmark RNA Genes

By applying the four-class model on ITRs and annotated ncRNAs, we found that 34% of ITRs, 38% of Araport ncRNAs, and of 65% TAIR ncRNAs were predicted as functional sequences (fig. 6E–G). Specifically, $\leq$20% of ITR and annotated ncRNA sequences were classified as RNA genes (fig. 6E–G). Although miRNAs dominate the benchmark RNA sequences, we should emphasize that the four-class prediction model also increased the proportion of snoRNAs and snRNAs that were predicted as functional (91%) compared with the 100 bp tissue agnostic model (67%). Thus a lack of similarity to benchmark miRNAs provides evidence that most ITRs and Araport ncRNAs are not functioning as RNA genes.

To provide an overall estimate of the proportion of likely functional and nonfunctional ITRs and annotated ncRNAs, we considered the predictions from the four-class model (fig. 6), the full model (figs. 3 and 4), and the tissue-agnostic models (supplementary figs. 5 and 6, Supplementary Material online), which cover both protein-coding and RNA gene functions. On the basis of support from $\geq$1 of the four models, we classified 4,437 ITRs (38%) and 796 annotated ncRNAs (44%) as likely functional, as they resembled either phenotype protein-coding or RNA genes. ITR and ncRNA sequences predicted as functional were most consistently associated with high expression breadth across data sets, low levels of nucleotide diversity (i.e., within-species conservation) and low levels of DNA methylation, particularly in CG contexts (fig. 7). The lower nucleotide diversity among predicted-functional ITRs and ncRNAs indicates that the prediction models are identifying sequences that may be under selection and are therefore likely to be functional.

Among benchmark sequence classes, 99% of phenotype protein-coding genes, 89% of benchmark RNA genes, and 31% of pseudogenes were predicted as functional based on support from $\geq$1 of the four models. Given the relatively high FPR among pseudogenes (31%), we should stress that the estimate of functional ITRs/ncRNAs is a liberal one. As a result, the set of ITRs and ncRNAs predicted as functional should be interpreted with caution, as they may contain a substantial proportion of false positive predictions. Assuming that pseudogenes are nonfunctional, we generated a conservative estimate of functional ITRs and ncRNAs by subtracting the pseudogene FPR from the proportion of predicted-functional sequences. Under this framework, we estimate that 7% of ITRs and 13% of annotated ncRNAs are likely functional. Most importantly, we find that the majority of ITRs (62%) and annotated ncRNAs (56%) is predicted as nonfunctional. Moreover, at least a third of ITRs (fig. 6E) and Araport ncRNAs (fig. 6F) most closely resemble unexpressed intergenic regions. On the basis of these findings, we conclude that the majority of ITRs and annotated ncRNA regions resembles nonfunctional genomic regions, and therefore are derived from noisy transcription.

Conclusion

Discerning the location of functional regions within a genome represents a key goal in genomic biology and is fundamental to molecular evolutionary studies. Despite advances in computational gene finding, it remains challenging to determine whether ITRs represent functional or noisy biochemical activity. We established robust function prediction models based on the evolutionary, biochemical, and structural characteristics of phenotype genes and pseudogenes in *A. thaliana*. The prediction models accurately define functional and nonfunctional regions and are applicable genome-wide and echo recent findings using human data to evaluate RNA gene functionality (Tsai et al. 2017). We utilized prediction models to assess the functionality of both protein-coding and annotated RNA genes. As benchmark examples of more recently identified RNA gene classes become available in *A. thaliana*, such as *cis*-acting regulatory (Guil and Esteller 2012) or competitive endogenous (Tan et al. 2015) RNAs, it will be interesting to see if sequences that encode RNA products with these roles can be predicted as functional based on a similar predictive framework. Given that function predictions were successful in both plants and metazoans, integrating the evolutionary and biochemical features of known genes for functional genomic region prediction will likely be applicable to any species. The next step will be to test whether function prediction models can be applied across species, which could ultimately allow the phenotype data and omics resources available in model systems to effectively guide the identification of functional regions in nonmodels.

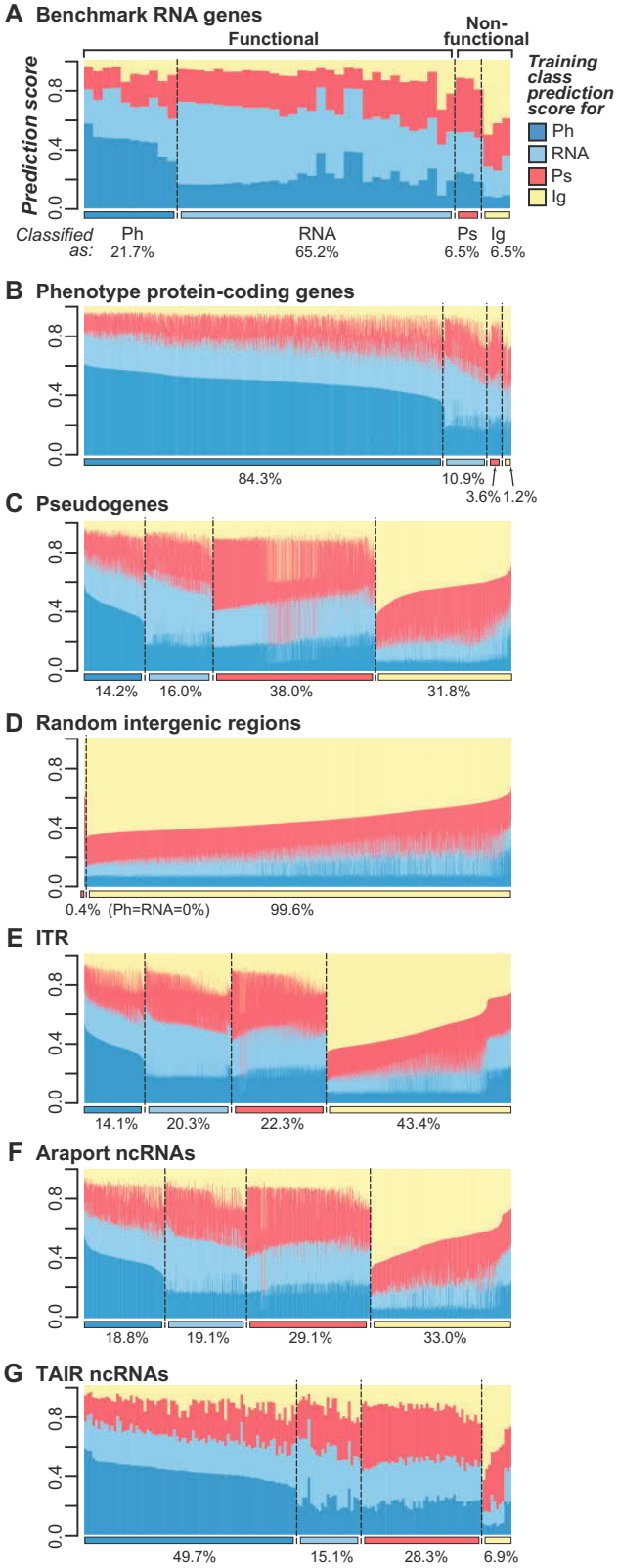

Fig. 6. Function predictions based on a four-class prediction model. (*A*) Stacked bar plots indicate the prediction scores of benchmark RNA genes for each of the four classes: Dark blue—phenotype protein-coding gene (Ph), cyan—RNA gene (RNA), red—pseudogene (Ps), yellow—random intergenic sequence (Ig). Ig were included to provide another set of likely nonfunctional sequences distinct from pseudogenes. Expression breadth and tissue-specific features were excluded and 100 bp sequences were used. A benchmark RNA gene is classified as one of the four classes according to the highest prediction score. The color bars below the chart indicate the predicted class, with the same color scheme as the prediction score. Sequences classified as Ph or RNA were considered functional, whereas those classified as Ps or Ig were considered nonfunctional. Percentages below a classification region indicate the proportion of sequences classified as that class. (*B*) Phenotype protein-coding gene prediction scores. (*C*) Pseudogene prediction scores. (*D*) Random unexpressed intergenic region prediction scores. Note that no sequence was predicted as functional. (*E*) Intergenic transcribed region (ITR), (*F*) Araport11 ncRNA regions. (*G*) TAIR10 ncRNA regions. Note that the 100 bp model used here allowed us to evaluate an additional 10,938 ITRs and 1,406 annotated ncRNAs compared with the 500 bp full and tissue-agnostic models.

Expression data were highly informative to functional predictions. We found that the prediction model based on only 24 transcription activity-related features performs nearly as well as the full model that integrates additional information including conservation, H3 mark, methylation, and TF binding data. In human, use of transcription data from cell lines also produced highly accurate predictions of functional genomic regions (Tsai et al. 2017). Importantly, our findings suggest that function prediction models can be established in any species, model or not, with a modest number of transcriptome data sets (e.g., 51 in this study and 19 in human). We emphasize that it is the breadth and level of expression that is informative for predictions, and that the presence of expression evidence by itself is an extremely poor predictor of functionality. With the effectiveness of our model noted, one major caveat of our models is that narrowly expressed phenotype genes are frequently predicted as pseudogene and broadly expressed pseudogenes tend to be called functional. To improve the function prediction model, it will be important to explore additional features unrelated to transcription, particularly those relevant to broadly expressed pseudogenes that are most likely recently pseudogenized. In addition, because few phenotype genes are narrowly expressed (5%) in the *A. thaliana* training data, more phenotyping data for narrowly expressed genes will be crucial as well.

Upon application of the function prediction models genome-wide, 4,427 ITRs and 796 annotated ncRNAs in *A. thaliana* are predicted as functional sequences. However, considering the high FPRs (e.g., 10% for the full and 31% for the four-class model), this is most likely an overestimate of the functional sequences contributed by ITRs and annotated ncRNAs. While we err on the side of calling nonfunctional sequences as functional, we reduce the error rate for calling a functional sequence as nonfunctional. Despite this conservative approach to classifying sequences as nonfunctional, the majority of ITRs and ncRNAs resembles pseudogenes and random unexpressed intergenic regions. Additionally, *A. thaliana* has a small genome and it is possible that species with larger genomes may exhibit a greater proportion of likely nonfunctional intergenic expression. Similar results were seen in human, where most ncRNAs are more similar to

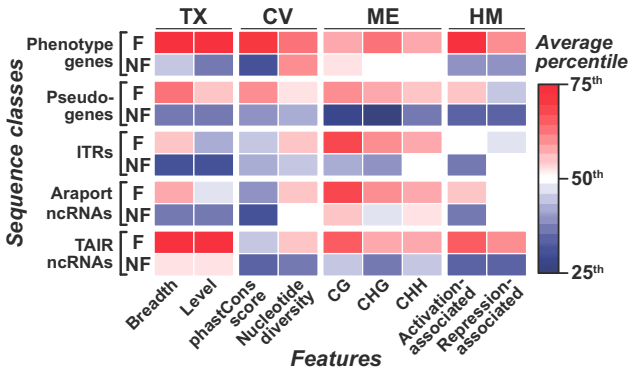

FIG. 7. Average percentile of example features among predicted functional/nonfunctional sequences relative to all sequences in our analysis. Example feature categories: Transcriptional activity (TX), sequence conservation (CV), DNA methylation (ME), and H3 marks (HM). Functional (F) and nonfunctional (NF) designations were based on support from ≥1 prediction model (full, 500 bp or 100 bp tissue-agnostic, and/or four-class model). Color indicates the average percentile of sequences for a given sequence class (phenotype gene, pseudogene, ITR, or annotated ncRNA) and prediction status (N or NF). Percentiles for each sequence were calculated with feature values sorted descending or ascending based on feature relationship with phenotype genes relative to pseudogenes. Descending: TX: breadth and level, CV: phastCons score, HM: activation-associated. Ascending: CV: nucleotide diversity, ME: CG, CHG, and CHH, HM: repression-associated.

nonfunctional sequences than they are to protein-coding and RNA genes (Tsai et al. 2017). Together with our finding of a significant relationship between the amount of intergenic expression and genome size, we conclude that a significant proportion of intergenic transcripts are nonfunctional noise. It is also important to note that there are a variety of reasons to not assume that most of a genome is functional, including ITRs (Palazzo and Gregory 2014). Thus, instead of assuming any expressed sequence must be functionally significant, we advocate that the null hypothesis should be that it is not, particularly considering that most ITRs and annotated ncRNAs have not been experimentally characterized.

Materials and Methods

Identification of Transcribed Regions in Leaf Tissue of 15 Flowering Plants

RNA-sequencing (RNA-seq) data sets were retrieved from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov/sra/; last accessed March 30, 2018) for 15 flowering plant species (supplementary table 1, Supplementary Material online). All data sets were generated from leaf tissue and sequenced on Illumina HiSeq 2000 or 2500 platforms. Genome sequences and gene annotation files were downloaded from Phytozome v.11 (www.phytozome.net; last accessed March 30, 2018; Goodstein et al. 2012) or Oropetium Base v.01 (www.sviridis.org; last accessed March 30, 2018; VanBuren et al. 2015). Genome sequences were repeat masked using RepeatMasker v4.0.5 (www.repeatmasker.org; last accessed March 30, 2018) if a repeat-masked version was not available. Only one end from paired-end read data sets were utilized in downstream processing. Reads were trimmed to be rid of low scoring ends and residual adaptor sequences using Trimmomatic v0.33 (LEADING: 3 TRAILING: 3 SLIDINGWINDOW: 4: 20 MINLEN: 20; Bolger et al. 2014) and mapped to genome sequences using TopHat v2.0.13 (default parameters except as noted below; Kim et al. 2013). Reads ≥20 nucleotides in length that mapped uniquely within a genome were used in further analysis.

For each species, thirty million mapped reads were randomly selected from among all data sets and assembled into transcript fragments using Cufflinks v2.2.1 (default parameters except as noted below; Trapnell et al. 2010), while correcting for sequence-specific biases during the sequencing process by providing an associated genome sequence with the -b flag. The expected mean fragment length for assembled transcript fragments in Cufflinks was set to 150 from the default of 200 so that expression levels in short fragments would not be overestimated. The 1st and 99th percentile of intron lengths for each species were used as the minimum and maximum intron lengths, respectively, for both the TopHat2 and Cufflinks steps. ITRs were defined by transcript fragments that did not overlap with gene annotation and did not have significant six-frame translated similarity to plant protein sequences in Phytozome v.10 (BLASTX *E*-value < 1*E*-05; Altschul et al. 1990).

Generation of Species Tree and Phylogenetically Independent Analysis

To assess potential confounding effects of phylogenetic non-independence on the relationships between genome size and expression coverage, we performed a phylogenetically independent contrasts (PIC) analysis. This analysis required the generation of a species phylogenetic tree of all 15 species (supplementary table 1, Supplementary Material online). Significant, nonself pairwise matches (maximum Expect (*E*)-value: 1×10^{-10}; ≥75% identity across ≥75% of the length of both proteins) were identified from an all-against-all BLASTP search using the annotated protein sequences from Phytozome v.12.1.5 for all 15 species. Significant matches were clustered with Markov Clustering (van Dongen 2000) using $-\log_{10}(E)$ value as the distance measure and an inflation parameter of 1.4. *E*-values equal to 0 were set to 180. We identified five clusters that contained a single sequence from each of the 15 species. For sequences of proteins in each of the five clusters, we performed multiple sequence alignments using MUSCLE (Edgar 2004). Alignments were concatenated and positions with gaps in 8 of the 15 species (>50%) were removed. RAxML (Stamatakis 2014) was used to generate 10,000 maximum likelihood trees and the final species tree was generated by midpoint rooting the highest likelihood tree using the phytools package in R (Revell 2012; supplementary fig. 7A, Supplementary Material online).

The phytools package was utilized to calculate phylogenetic contrasts of genome size, coverage of genic expression, and coverage of intergenic expression based on the final species tree. Phylogenetic contrasts for coverage of genic and intergenic expression were then individually regressed against

the phylogenetic contrasts for genome sizes. The final species tree used for PIC contained eudicotyledonous (dicot) relationships incongruent with established phylogeny (APG IV 2016; supplementary fig. 7*A*, Supplementary Material online). Two strategies were utilized to determine whether these errors had significant effects on PIC results by: 1) manually adjusting the tree topology to be congruent with the established phylogeny and randomly altering the length of branches leading to misclassified dicot species (supplementary fig. 7*B*, Supplementary Material online) and 2) randomly shuffling the species labels for the misclassified dicot species (supplementary fig. 7*C*, Supplementary Material online).

Phenotype Data Sources

Mutant phenotype data for *A. thaliana* protein-coding genes were collected from a published data set (Lloyd and Meinke 2012), the Chloroplast 2010 database (Ajjawi et al. 2010; Savage et al. 2013), and the RIKEN phenome database (Kuromori et al. 2006) as described by Lloyd et al. (2015). Phenotype genes used in our analyses were those whose disruption resulted in lethal or visible defects under standard laboratory growth conditions. Genes with documented mutant phenotypes under standard conditions were considered as a distinct and nonoverlapping category from other annotated protein-coding genes. We identified six RNA genes with documented loss-of-function phenotypes through literature searches (supplementary table 6, Supplementary Material online): *At4* (AT5G03545; Shin et al. 2006), *MIR164A* and *MIR164D* (AT2G47585 and AT5G01747, respectively; Guo et al. 2005), *MIR168A* (AT4G19395; Li, Cui, et al. 2012), and *MIR828A* and *TAS4* (AT4G27765 and AT3G25795, respectively; Hsieh et al. 2009). Conditional phenotype genes were those belonging to the Conditional phenotype group as described by Lloyd and Meinke (2012). Loss-of-function mutants of these genes exhibited phenotype only under stress conditions.

Arabidopsis thaliana Genome Annotation

Arabidopsis thaliana protein-coding gene, miRNA gene, snoRNA gene, snRNA gene, ncRNA region, pseudogene, and TE annotations were retrieved from The Arabidopsis Information Resource v.10 (TAIR10; www.arabidopsis.org; last accessed March 30, 2018; Berardini et al. 2015). Additional miRNA gene and lncRNA region annotations were retrieved from Araport v.11 (www.araport.org; last accessed March 30, 2018). A primary difference between the TAIR ncRNAs and Araport lncRNAs (referred to as Araport ncRNAs in the Results and Discussion section) is the date in which they were annotated. For example, 221 ncRNAs were present in the v.7 release of TAIR, which dates back to 2007 (TAIR10 contains 394 ncRNA annotations; Swarbreck et al. 2007; Lamesch et al. 2012; Berardini et al. 2015). However, Araport lncRNAs were annotated in the past five years (Krishnakumar et al. 2015). Thus, that TAIR ncRNAs are generally more highly and broadly expressed is likely a result of the less sensitive transcript identification methods available for early TAIR releases. A pseudogene-finding pipeline (Zou et al. 2009) was used to identify additional pseudogene fragments and count the number of disabling mutations (premature stop or frameshift mutations). Genes, pseudogenes, and transposons with overlapping annotation were excluded from further analysis. Overlapping lncRNA annotations were merged for further analysis. When pseudogenes from TAIR10 and the pseudogene-finding pipeline overlapped, the longer pseudogene annotation was used.

Arabidopsis thaliana ITRs analyzed include: 1) the Set 2 ITRs in Moghe et al. (2013), 2) the novel transcribed regions from Araport v.11, and 3) additional ITRs from 206 RNA-seq data sets (supplementary table 7, Supplementary Material online). Reads were trimmed, mapped, and assembled into transcript fragments as described above, except that overlapping transcript fragments from across data sets were merged. ITRs analyzed did not overlap with any TAIR10, Araport11, or pseudogene annotation. Overlapping ITRs from different annotated subsets were kept based on a priority system: Araport11 > Set 2 ITRs from Moghe et al. (2013) > ITRs identified in this study. For each sequence entry (gene, ncRNA, pseudogene, TE, or ITR), a 100 and 500 base pair (bp) window was randomly chosen for calculating feature values and subsequent model building steps. Feature descriptions are provided in the following sections. The feature values for randomly selected 500 and 100 bp windows are provided in supplementary tables 2 and 5, Supplementary Material online, respectively. Additionally, nonexpressed intergenic sequences were randomly sampled from genome regions that did not overlap with annotated genes, pseudogenes, TEs, or regions with genic or intergenic transcript fragments (100 bp, $n = 4{,}000$; 500 bp, $n = 3{,}716$). All 100 and 500 bp windows described above are referred to as sequence windows throughout the Materials and Methods section.

Sequence Conservation and Structure Features

There were 10 sequence conservation features examined. The first two were derived from comparisons between *A. thaliana* accessions including nucleotide diversity and Tajima's D among 81 accessions (Cao et al. 2011) using a genome matrix file from the 1,001 genomes database (www.1001genomes.org; last accessed March 30, 2018). The python scripts are available through GitHub (https://github.com/ShiuLab/GenomeMatrixProcessing; last accessed March 30, 2018). The remaining eight features were derived from cross-species comparisons, three based on multiple sequence and five based on pairwise alignments. Three multiple sequence alignment-based features were established using aligned genomic regions between *A. thaliana* and six other plant species (*G. max*, *Medicago truncatula*, *Populus trichocarpa*, *Vitis vinifera*, *Sorghum bicolor*, and *O. sativa*; Li, Zheng, et al. 2012), which are referred to as conserved blocks. For each conserved block, the first feature was the proportion of a sequence window that overlapped a conserved block (referred to as coverage), and the two other features were the maximum and average phastCons scores within each sequence window. The phastCons score was determined for each nucleotide within conserved blocks (Li, Zheng, et al. 2012). Nucleotides in a sequence window that did not overlap with a conserved

block were assigned a phastCons score of 0. For each sequence window, five pairwise alignment-based cross-species conservation features were the percent identities to the most significant BLASTN match (if *E*-value $<$ 1*E*-05) in each of five taxonomic groups. The five taxonomic groups included the *Brassicaceae* family ($n_{species} = 7$), other dicotyledonous plants (22), monocotyledonous plants (7), other embryophytes (3), and green algae (5). If no sequence with significant similarity was present, percent identity was scored as zero.

For sequence-structure features, we used 125 conformational and thermodynamic dinucleotide properties collected from DiProDB database (Friedel et al. 2009). Because the number of dinucleotide properties was high and dependent, we reduced the dimensionality by utilizing principal component (PC) analysis as described previously (Tsai et al. 2015). Sequence-structure values corresponding to the first five PCs were calculated for all dinucleotides in and averaged across the length of a sequence window and used as features when building function prediction models.

Transcription Activity Features

We generated four multi-data set and 20 individual data set transcription activity features. To identify a set of RNA-seq data sets to calculate multi-data set features, we focused on the 72 of 206 RNA-seq data sets each with $\geq$20 million reads (see above; supplementary table 7, Supplementary Material online). Transcribed regions were identified with TopHat2 and Cufflinks as described in the RNA-seq analysis section except that the 72 *A. thaliana* RNA-seq data sets were used. Following transcript assembly, we excluded 21 RNA-seq data sets because they had unusually high RPKM (Reads Per Kilobase of transcript per Million mapped reads) values (median RPKM value range = 272–2,504,294) compared with the rest (2–252). The remaining 51 RNA-seq data sets were used to generate four multi-data set transcription activity features including: Expression breadth, 95th percentile expression level, maximum transcript coverage, and presence of expression evidence (for values see supplementary tables 2 and 5, Supplementary Material online). Expression breadth was the number of RNA-seq data sets that have $\geq$1 transcribed region that overlapped with a sequence window. The 95th percentile expression level was the 95th percentile of RPKM values across 51 RNA-seq data sets where RPKM values were set to 0 if there was no transcribed region for a sequence window. Maximum transcript coverage was the maximum proportion of a sequence window that overlapped with a transcribed region across 51 RNA-seq data sets. Presence of expression evidence was determined by overlap between a sequence window and any transcribed region in the 51 RNA-seq data sets.

In addition to features based on multiple data sets, 20 individual data set features were derived from 10 data sets: Seven tissue/organ-specific RNA-seq data sets including pollen (SRR847501), seedling (SRR1020621), leaf (SRR953400), root (SRR578947), inflorescence (SRR953399), flower, (SRR505745), and silique (SRR953401), and three data sets from nonstandard growth conditions, including dark-grown seedlings (SRR974751) and leaf tissue under drought (SRR921316) and fungal infection (SRR391052). For each of these 10 RNA-seq data sets, we defined two features for each sequence window: The maximum transcript coverage (as described above) and the maximum RPKM value of overlapping transcribed regions (referred to as Level in fig. 2). If no transcribed regions overlapped a sequence window, the maximum RPKM value was set as 0. For the analysis of narrowly and broadly expressed phenotype genes and pseudogenes (supplementary fig. 4*B* and *C*, Supplementary Material online), we used 28 out of 51 RNA-seq data sets generated from a single tissue and in standard growth conditions to calculate the number of tissues with evidence of expression (tissue expression breadth). In total, seven tissues were represented among the 28 selected RNA-seq data sets (see above; supplementary table 7, Supplementary Material online), and thus tissue expression breadth ranges from 0 to 7 (note that only 1–7 are shown in supplementary fig. 4*B* and *C*, Supplementary Material online due to low sample size of phenotype genes in the 0 bin). The tissue breadth value is distinct from the expression breadth feature used in model building that was generated using all 51 data sets and considered multiple RNA-seq data sets from the same tissue separately (range: 0–51).

H3 Mark Features

Twenty H3 mark features were calculated based on eight H3 chromatin immunoprecipitation sequencing (ChIP-seq) data sets from SRA. The H3 marks examined include four associated with activation (H3K4me1: SRR2001269, H3K4me3: SRR1964977, H3K9ac: SRR1964985, and H3K23ac: SRR1005405) and four associated with repression (H3K9me1: SRR1005422, H3K9me2: SRR493052, H3K27me3: SRR3087685, and H3T3ph: SRR2001289). Reads were trimmed as described in the RNA-seq section and mapped to the TAIR10 genome with Bowtie v2.2.5 (default parameters; Langmead et al. 2009). Spatial Clustering for Identification of ChIP-Enriched Regions v.1.1 (Xu et al. 2014) was used to identify ChIP-seq peaks with a false discover rate $\leq$ 0.05 with a nonoverlapping window size of 200, a gap parameter of 600, and an effective genome size of 0.92 (Koehler et al. 2011). For each H3 mark, two features were calculated for each sequence window: The maximum intensity among overlapping peaks and peak coverage (proportion of overlap with the peak that overlaps maximally with the sequence window). In addition, four multi-mark features were generated. Two of the multi-mark features were the number of activating marks (0–4) overlapping a sequence window and the proportion of a sequence window overlapping any peak from any of the four activating marks (activating mark peak coverage). The remaining two multi-mark features were the same as the two activating multi-mark features except focused on the four repressive marks.

DNA Methylation Features

Twenty-one DNA methylation features were calculated from bisulfite-sequencing (BS-seq) data sets from seven tissues (pollen: SRR516176, embryo: SRR1039895, endosperm: SRR1039896, seedling: SRR520367, leaf: SRR1264996, root:

SRR1188584, and inflorescence: SRR2155684). BS-seq reads were trimmed as described above and processed with Bismark v.3 (default parameters; Krueger and Andrews 2011) to identify methylated and unmethylated cytosines in CG, CHH, and CHG (H = A, C, or T) contexts. Methylated cytosines were defined as those with ≥5 mapped reads and with >50% of mapped reads indicating that the position was methylated. For each BS-seq data set, the percentage of methylated cytosines in each sequence window for CG, CHG, and CHH contexts were calculated if the sequence window had ≥5 cytosines with ≥5 reads mapping to the position. To determine whether the above parameters where reasonable, we assessed the FPR of DNA methylation calls by evaluating the proportion of cytosines in the chloroplast genome that are called as methylated, as the chloroplast genome has few DNA methylation events (Ngernprasirtsiri et al. 1988; Zhang et al. 2006). On the basis of the above parameters, 0–1.5% of cytosines in CG, CHG, or CHH contexts in the chloroplast genome were considered methylated in any of the seven BS-seq data sets. This indicated that the FPRs for DNA methylation calls were low and the parameters were reasonable.

Chromatin Accessibility and TF Binding Features

Chromatin accessibility features consisted of ten DHS-related features and one micrococcal nuclease sequencing (MNase-seq)-derived feature. DHS peaks from five tissues (seed coat, seedling, root, unopened flowers, and opened flowers) were retrieved from the Gene Expression Omnibus (GSE53322 and GSE53324; Sullivan et al. 2014). For each of the five tissues, the maximum DHS peak intensity and DHS peak coverage were calculated for each sequence window. Normalized nucleosome occupancy per bp based on MNase-seq was obtained from Liu et al. (2015). The average nucleosome occupancy value was calculated across each sequence window. TF binding site features were based on in vitro DNA affinity purification sequencing data of 529 TFs (O'Malley et al. 2016). Two features were generated for each sequence window: The total number of TF binding sites and the number of distinct TFs bound.

Single-Feature Prediction Performance

The ability for each single feature to distinguish between functional and nonfunctional regions was evaluated by calculating AUC-ROC value with the Python scikit-learn package (Pedregosa et al. 2011). Thresholds to predict sequences as functional or nonfunctional using a single feature were defined by the feature value that produced the highest *F*-measure, the harmonic mean of precision (proportion of sequences predicted as functional that are truly functional) and recall (proportion of truly functional sequences predicted as functional). The *F*-measure allows consideration of both false positives and false negatives at a given threshold. FPR were calculated as the percentage of negative (nonfunctional) cases with values above or equal to the threshold and thus falsely predicted as functional. FNR were calculated as the percentage of positive (functional) cases with values below the threshold and thus falsely predicted as nonfunctional.

Binary Classification with Machine Learning

For binary classification (two-class) models that contrasted phenotype genes and pseudogenes, the random forest (RF) implementation in the Waikato Environment for Knowledge Analysis software (WEKA; Hall et al. 2009) was utilized. Three types of two-class models were established, including the full model (500 bp sequence window, figs. 3*A*, *B* and 4), tissue-agnostic models (500 bp, supplementary fig. 5, Supplementary Material online; 100 bp, supplementary fig. 6, Supplementary Material online), and single feature category models (fig. 3*A* and *B*). For each model type, we first generated 100 balanced data sets by randomly selecting equal numbers of phenotype genes (positive examples) and pseudogenes (negative examples). For each of these 100 data sets, 10-fold stratified cross-validation was utilized, where the model was trained using 90% of sequences and tested on the remaining 10%. Thus, for each model type, a sequence window had 100 prediction scores, where each score was the proportion of 500 RF trees that predicted a sequence as a phenotype gene in a balanced data set. The median of 100 prediction scores was used as the FL value (supplementary table 4, Supplementary Material online). The FL threshold to predict a sequence as functional or nonfunctional was defined based on maximum *F*-measure as described in the previous section.

We tested multiple -K parameters (2–25) in the WEKA-RF implementation, which alters the number of randomly selected features included in each RF tree (supplementary table 8, Supplementary Material online), and found that 15 randomly selected features provided the highest performance based on AUC-ROC (calculated and visualized using the ROCR package; Sing et al. 2005). Feature importance was assessed by excluding one feature at a time to determine the associated reduction in prediction performance (supplementary table 9, Supplementary Material online). All leave-one-out models performed well (AUC-ROC > 0.97), indicating that no single feature was dominating the function predictions and/or many features are correlated (supplementary fig. 8, Supplementary Material online). To demonstrate that functional predictions were not overfitted, we generated a prediction model while holding out a randomly sampled set of phenotype protein-coding genes and pseudogenes ($n = 100$ for each class) from model training and parameter optimization steps. The held-out instances were well-predicted, with a high AUC-ROC performance (0.97) and low false positive and FNRs (12% and 7%, respectively). Binary classification models were also built using all features from 500 bp sequences (equivalent to the full model) with the Sequential Minimal Optimization-Support Vector Machine (SMO-SVM) implementation in WEKA (Hall et al. 2009). The results of SMO-SVM models were highly similar to the full RF results: *PCC* between the FL values generated by RF and SMO-SVM = 0.97; AUC-ROC of SMO-SVM = 0.97; FPR = 12%; FNR = 3%. By comparison, the full RF model had AUC-ROC = 0.98, FPR = 10%, FNR = 4%.

Tissue-agnostic models were generated by excluding the expression breadth feature and 95th percentile expression level and replacing all features from RNA-seq, BS-seq, and

DHS data sets that were available in multiple tissues. For multiple-tissue RNA-seq data, the maximum expression level across 51 RNA-seq data sets (in RPKM) and maximum coverage (as described in the transcription activity section) of a sequence window in any of 51 RNA-seq data sets were used. For multi-tissue DNA methylation features, minimum proportions of methylated cytosines in any tissue in CG, CHG, and CHH contexts were used. For DHS data, the maximum peak intensity and peak coverage was used instead. In single feature category predictions, fewer total features were used and therefore lower $-K$ values (i.e., the number of random features selected when building RFs) were considered in parameter searches (supplementary table 8, Supplementary Material online).

Multi-Class Machine Learning Model

For the four-class model, benchmark RNA gene, phenotype protein-coding gene, pseudogene, and random unexpressed intergenic sequences were used as the four training classes. Benchmark RNA genes consisted of six RNA genes with documented loss-of-function phenotypes and 40 high-confidence miRNA genes from miRBase (www.mirbase.org; last accessed March 30, 2018; Kozomara and Griffiths-Jones 2014). We considered that the decreased numbers of benchmark RNA genes would not allow us to effectively distinguish between sequence classes. However, binary predictions generated using 35 phenotype gene and pseudogene instances and the 100 bp tissue-agnostic feature set resulted in an AUC-ROC performance of 0.96. We generated 250 data sets with equal proportions (larger classes randomly sampled) of training sequences. Two-fold stratified cross-validation was utilized due to the low number of benchmark RNA genes. The features included those described for the tissue-agnostic model and focused on 100 bp sequence windows. The RF implementation, *cforest*, in the *party* package of R (Strobl et al. 2008) was used to build the classifiers. The four-class predictions provide prediction scores for each sequence type: An RNA gene, phenotype protein-coding gene, pseudogene, and unexpressed intergenic score (supplementary table 4, Supplementary Material online). The prediction scores indicate the proportion of RF trees that classify a sequence as a particular class. Median prediction scores from across 100 balanced runs were used as final prediction scores. Scores from a single balanced data set models sum to 1, but not the median from 100 balanced runs. Thus, the median scores were scaled to sum to 1. For each sequence window, the maximum prediction score among the four classes was used to classify a sequence as phenotype gene, pseudogene, unexpressed intergenic region, or RNA gene.

Availability

All relevant data are within the article and supplementary material files, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors wish to thank Christina Azodi, Ming-Jung Liu, Gaurav Moghe, Bethany Moore, and Sahra Uygun for providing processed data and discussion. This work was partly supported by the National Science Foundation (grant numbers IOS-1126998, IOS-1546617, and DEB-1655386 to S.H.S.), and Research Experience for Undergraduates support to R.P.S.; and the Michigan State University Dissertation Continuation Fellowship to J.P.L.

Author Contributions

J.P.L., Z.T.-Y.T., and S.-H.S. designed the research. J.P.L., Z.T.-Y.T., R.P.S., and N.L.P. performed the research. J.P.L., Z.T-Y.T., R.P.S., N.L.P., and S.-H.S. wrote the article.

References

Ajjawi I, Lu Y, Savage LJ, Bell SM, Last RL. 2010. Large-scale reverse genetics in *Arabidopsis*: case studies from the Chloroplast 2010 Project. *Plant Physiol*. 152(2):529–540.

Altschul SF, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3):403–410.

Amundson R, Lauder GV. 1994. Function without purpose. *Biol Philos*. 9(4):443–469.

APG IV 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: aPG II. *Bot J Linn Soc*. 181(1):1–20.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* 53(8):474–485.

Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, Zhang MQ, Sedel F, Jourdren L, Coulpier F, et al. 2010. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J*. 29(18):3082–3093.

Boeck ME, Huynh C, Gevirtzman L, Thompson OA, Wang G, Kasper DM, Reinke V, Hillier LW, Waterston RH. 2016. The time-resolved transcriptome of *C. elegans*. *Genome Res*. 26(10):1441–1450.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature* 512(7515):393–399.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* 43(10):956–963.

Comings DE. 1972. The structure and function of chromatin. *Adv. Hum. Genet*. 3:237–431.

Doolittle WF, Brunet TDP, Linquist S, Gregory TR. 2014. Distinguishing between "function" and "effect" in genome biology. *Genome Biol Evol*. 6(5):1234–1237.

Eddy SR. 2013. The ENCODE project: missteps overshadowing a success. *Curr Biol*. 23(7):R259–R261.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.

ENCODE Project Consortium 2012. An integrated encyclopedia of {DNA} elements in the human genome. *Nature* 489(7414):57–74.

Fei Q, Xia R, Meyers BC. 2013. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell*. 25(7):2400–2415.

Friedel M, Nikolajewa S, Sühnel J, Wilhelm T. 2009. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res*. 37(Database issue):D37–D40.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a

comparative platform for green plant genomics. *Nucleic Acids Res.* 40(Database issue):D1178–D1186.
Graur D. 2017. An upper limit on the functional fraction of the human genome. *Genome Biol Evol*. 9(7):1880–1885.
Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*. 5(3):578–590.
Guil S, Esteller M. 2012. Cis-acting noncoding RNAs: friends and foes. *Nat Struct Mol Biol*. 19(11):1068–1075.
Gulko B, Gronau I, Hubisz MJ, Siepel A. 2014. Probabilities of fitness consequences for point mutations across the human genome. *Nat Genet.* 47(3):276–283.
Guo H-S, Xie Q, Fei J-F, Chua N-H. 2005. MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. *Plant Cell.* 17(5):1376–1386.
Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software. *ACM SIGKDD Explor Newsl*. 11(1):10.
Hardiman KE, Brewster R, Khan SM, Deo M, Bodmer R. 2002. The bereft gene, a potential target of the neural selector gene cut, contributes to bristle morphogenesis. *Genetics* 161(1):231–247.
Hsieh L-C, Lin S-I, Shih AC-C, Chen J-W, Lin W-Y, Tseng C-Y, Li W-H, Chiou T-J. 2009. Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. *Plant Physiol*. 151(4):2120–2132.
Karreth FA, Reschke M, Ruocco A, Ng C, Chapuy B, Léopold V, Sjoberg M, Keane TM, Verma A, Ala U, et al. 2015. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell* 161(2):319–332.
Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A*. 111(17):6131–6138.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14(4):R36.
Koehler R, Issac H, Cloonan N, Grimmond SM. 2011. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics* 27(2):272–274.
Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42(Database issue):D68–D73.
Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, Rosen BD, Cheng C-Y, Moreira W, Mock SA, et al. 2015. Araport: the Arabidopsis information portal. *Nucleic Acids Res.* 43(Database issue):D1003–D1009.
Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27(11):1571–1572.
Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34(7):1812–1819.
Kuromori T, Wada T, Kamiya A, Yuguchi M, Yokouchi T, Imura Y, Takabe H, Sakurai T, Akiyama K, Hirayama T, et al. 2006. A trial of phenome analysis using 4000 *Ds*-insertional mutants in gene-coding regions of Arabidopsis. *Plant J*. 47(4):640–651.
Lai K-MV, Gong G, Atanasio A, Rojas J, Quispe J, Posca J, White D, Huang M, Fedorova D, Grant C, et al. 2015. Diverse phenotypes and specific transcription patterns in twenty mouse lines with ablated LincRNAs. *PLoS ONE.* 10(4):e0125522.
Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40(Database issue):D1202–D1210.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10(3):R25.
Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell.* 24(11):4346–4359.
Li W, Cui X, Meng Z, Huang X, Xie Q, Wu H, Jin H, Zhang D, Liang W. 2012. Transcriptional regulation of Arabidopsis {MIR168a} and argonaute1 homeostasis in abscisic acid and abiotic stress responses. *Plant Physiol*. 158(3):1279–1292.
Li W, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* 292:237–239.
Liu M-J, Seddon AE, Tsai ZT-Y, Major IT, Floer M, Howe GA, Shiu S-H. 2015. Determinants of nucleosome positioning and their influence on plant gene expression. *Genome Res*. 25(8):1182–1195.
Lloyd J, Meinke D. 2012. A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. *Plant Physiol*. 158(3):1115–1129.
Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H. 2015. Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. *Plant Cell.* 27(8):2133–2147.
Marahrens Y, Panning B, Dausman J, Strauss W, Jaenisch R. 1997. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev*. 11(2):156–166.
Michael TP, Jackson S. 2013. The first 50 plant genomes. *Plant Genome*. 6(2):1–7.
Moghe GD, Lehti-Shiu MD, Seddon AE, Yin S, Chen Y, Juntawong P, Brandizzi F, Bailey-Serres J, Shiu S-H. 2013. Characteristics and significance of intergenic polyadenylated RNA transcription in *Arabidopsis*. *Plant Physiol*. 161(1):210–224.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (80-)* 320(5881):1344–1349.
Neander K. 1991. Functions as selected effects: the conceptual analyst's defense. *Philos Sci*. 58(2):168–184.
Ngernprasirtsiri J, Kobayashi H, Akazawa T. 1988. DNA methylation as a mechanism of transcriptional regulation in nonphotosynthetic plastids in plant cells. *Proc Natl Acad Sci U S A*. 85(13):4750–4754.
Ning S, Wang P, Ye J, Li X, Li R, Zhao Z, Huo X, Wang L, Li F, Li X. 2013. A global map for dissecting phenotypic variants in human lincRNAs. *Eur J Hum Genet*. 21(10):1128–1133.
Niu D-K, Jiang L. 2013. Can {ENCODE} tell us how much junk {DNA} we carry in our genome? *Biochem Biophys Res Commun*. 430(4):1340–1343.
Nobuta K, Venu RC, Lu C, Beló A, Vemaraju K, Kulkarni K, Wang W, Pillay M, Green PJ, Wang G-L, et al. 2007. An expression atlas of rice mRNAs and small RNAs. *Nat Biotechnol*. 25(4):473–477.
O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* 166:1598.
Palazzo AF, Gregory TR. 2014. The case for junk DNA. *PLoS Genet*. 10(5):e1004351.
Palazzo AF, Lee ES. 2015. Non-coding RNA: what is functional and what is junk? *Front Genet*. 5:1–11.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine Learning in Python. *J Mach Learn Res*. 12:2825–2830.
Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. 1996. Requirement for Xist in X chromosome inactivation. *Nature* 379(6561):131–137.
Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465(7301):1033–1038.
Ponting CP, Belgard TG. 2010. Transcribed dark matter: meaning or myth? *Hum Mol Genet.* 19(R2):R162–R168.
Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol*. 3(2):217–223.
Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, et al. 2013. Multiple knockout mouse models reveal {lincRNAs} are required for life and brain development. *Elife* 2:e01749.

Savage LJ, Imre KM, Hall DA, Last RL. 2013. Analysis of essential Arabidopsis nuclear genes encoding plastid-targeted proteins. *PLoS ONE*. 8(9):e73291.

Schreiber SL, Bernstein BE. 2002. Signaling network model of chromatin. *Cell* 111(6):771–778.

Shin H, Shin H-S, Chen R, Harrison MJ. 2006. Loss of At4 function impacts phosphate distribution between the roots and the shoots during phosphate starvation. *Plant J*. 45(5):712–726.

Simon SA, Meyers BC. 2011. Small RNA-mediated epigenetic modifications in plants. *Curr Opin Plant Biol*. 14(2):148–155.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* 21(20):3940–3941.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Bednarek S, et al. 2005. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc Natl Acad Sci U S A*. 102(12):4453–4458.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*. 9:307.

Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol*. 14(2):103–105.

Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al. 2014. Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep*. 8(6):2015–2030.

Svensson O, Arvestad L, Lagergren J. 2006. Genome-wide survey for biologically functional pseudogenes. *PLoS Comput. Biol*. 2(5):e46.

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. 2007. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*. 36(Database):D1009–D1014.

Tan JY, Sirey T, Honti F, Graham B, Piovesan A, Merkenschlager M, Webber C, Ponting CP, Marques AC. 2015. Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Res*. 25(5):655–666.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 28(5):511–515.

Tsai ZT-Y, Lloyd JP, Shiu S-H. 2017. Defining functional genic regions in the human genome through integration of biochemical, evolutionary, and genetic evidence. *Mol Biol Evol*. 34(7):1788–1798.

Tsai ZT-Y, Shiu S-H, Tsai H-K. 2015. Contribution of sequence motif, chromatin state, and DNA structure features to predictive models of transcription factor binding in yeast. *PLoS Comput Biol*. 11(8):e1004418.

van Dongen S. 2000. Graph Clustering by Flow Simulation.

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, et al. 2015. Single-molecule sequencing of the desiccation-tolerant grass Oropetium *thomaeum*. *Nature* 527(7579):508–511.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T-Y, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*. 40(7):897–903.

Xu S, Grullon S, Ge K, Peng W. 2014. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol Biol*. 1150:97–111.

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al. 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science (80-)* 302(5646):842–846.

Yang L, Takuno S, Waters ER, Gaut BS. 2011. Lowly expressed genes in *Arabidopsis thaliana* bear the signature of possible pseudogenization by promoter degradation. *Mol Biol Evol*. 28(3):1193–1203.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al. 2006. Genome-wide high-resolution mapping and functional analysis of {DNA} methylation in arabidopsis. *Cell* 126(6):1189–1201.

Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, et al. 2016. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res*. 44(D1):D203–D208.

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu S-H. 2009. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiol*. 151(1):3–15.