Defining functional intergenic transcribed regions based on heterogeneous features of phenotype genes and pseudogenes

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

[1] Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA
[2] Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan
[3] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA
[4] Genetics Program, Michigan State University, East Lansing, MI 48824, USA
[5] Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI 48824, USA
* To whom correspondence should be addressed.

Correspondence to:
Shin-Han Shiu
Michigan State University
E-mail: shius@msu.edu
Telephone number: 517-353-7196

[a] Present address: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

Running title: Functionality of intergenic transcription

Keywords: Intergenic transcription, ncRNAs, definition of function, machine learning, molecular evolution, chromatin state

ABSTRACT

With advances in transcript profiling, the presence of transcriptional activities in intergenic regions has been well established in multiple model systems. However, whether intergenic expression reflects transcriptional noise or the 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 non-functional. To further assess the functionality of ITRs, we first built machine learning classifiers using *Arabidopsis thaliana* as a model that can accurately distinguish functional sequences (phenotype genes) and non-functional ones (pseudogenes and random unexpressed intergenic regions) by integrating 93 biochemical, evolutionary, and sequence-structure features. Next, by applying the models to ITRs, we found that 2,453 (21%) had features significantly similar to phenotype genes and thus were likely parts of functional genes, while an additional 17% resembled benchmark RNA genes. However, ~60% of ITRs were more similar to nonfunctional sequences and should be considered transcriptional noise unless falsified with experiments. The predictive framework establish here provides not only a comprehensive look at how functional, genic sequences are distinct from likely non-functional ones, but also a new way to differentiate novel genes from genomic regions with noisy transcriptional activities.

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 *Homo sapiens* (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 and ~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), respectively. The presence of intergenic transcripts indicates that there may be additional genes in genomes that have escaped gene finding efforts thus far. Considering that knowledge of the complete suite of functional elements present in a genome is an important goal for large-scale functional genomics efforts and the quest to connect genotype to phenotype, identifying functional intergenic transcribed regions (ITRs) represents a fundamental task that is critical to our understanding of the gene space in a genome.

Loss-of-function phenotyping analysis represents the gold standard by which the functional significance of genomic regions, including ITRs, can be confirmed (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), indicating that they are *bona fide* genes. In addition, loss-of-function mutants have been used to confirm ITR functionality in mouse embryonic stem cell proliferation (Ivanova et al. 2006; Guttman et al. 2009) and male reproductive development (Heinen et al. 2009), as well as brain and eye development in *Danio rario* (Ulitsky et al. 2011). In human, 162 long intergenic non-coding RNAs (lincRNAs) harbor phenotype-associated SNPs, suggesting that these expressed intergenic regions may be functional (Ning et al. 2013). In addition to intergenic expression, most model organisms feature an abundance of annotated non-coding RNA (ncRNA) sequences (Zhao et al. 2016), which are mostly identified through the presence of expression occurring outside of annotated 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 experimental characterization, 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, despite

the presence of a few notable examples, the number of ITRs and ncRNAs with well-established functions is dwarfed by those with no known function.

While some ITRs and ncRNAs are likely novel genes, intergenic transcription can also be the byproduct of noisy expression 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). Thus, whether an intergenic transcript is functional cannot be depend on solely the fact that it is expressed. In addition to the biochemical activity, the genomic region with the activity must be under selection. This line of logic has revived the classical idea on how function can be defined based on “causal role” or “selected effect” functionality (Doolittle et al. 2014). A “causal role” definition requires a definable activity to consider a genomic region as functional (Cummins 1975; Amundson and Lauder 1994), which is adopted by the ENCODE Consortium (2012) to classify ~80% of the human genome as having biochemical functions. This finding has been used as evidence disproving the presence of junk DNA that are not under natural selection (see Eddy 2013). This has drawn considerable critique because biochemical activity itself is not an indication of selection (Graur et al., 2013; Niu and Jiang, 2013). Instead, if we are interested in if a genomic region with discernible activity is under selection, selected effect functionality is advocated to be a more suitable definition for function (Amundson and Lauder 1994; Graur et al., 2013; Doolittle et al. 2014). Under the selected effect functionality definition, ITRs and most annotated ncRNA genes remain functionally ambiguous.

If an ITR is functional, it would represent a genic sequence that is not identified with conventional gene finding programs. Gene finding programs incorporate sequence characteristics, transcriptional evidence, and conservation information to define genic regions that are expected to be functional. Thus, genes that lack the features typically associated with genic regions will remain unidentified. Due to the debate on the definitions of function post ENCODE, Kellis et al. (2014) has suggested that evolutionary, biochemical, and genetic evidences provide complementary information to define functional genomic regions. Integrating chromatin accessibility, transcriptome, and conservation evidence was shown to be successful in identifying regions in the human genome that are under selection (Gulko et al. 2014). Moreover, a comprehensive integration of biochemical, evolutionary, and genetic evidence resulted in highly-accurate identification of human disease genes and pseudogenes (Tsai et al. 2017). However, it is not known if such predictions are possible outside of animal systems or if the

features that define functional genomic regions in animals are applicable in other biological kingdoms. In plant species, despite the fact that many biochemical signatures are known to be associated with genic regions, these signatures have not been incorporated to assist in identifying the functional genomic regions.

To investigate the prevalence of intergenic transcription across species with a wide range of genome sizes, we identified ITRs in 15 flowering plant species with 17-fold genome size differences. To assess the functionality of plant intergenic transcripts, we first determined whether 93 evolutionary, biochemical, and sequence-structure features could distinguish functional sequences (phenotype genes) and non-functional ones (pseudogenes and random unexpressed intergenic regions) using *A. thaliana* as a model. Next, we jointly considered all features to establish functional gene prediction models using machine learning methods. Finally, we applied the models to ITRs and putative ncRNAs to determine whether these functionally ambiguous sequences are more similar to known functional or likely non-functional sequences.

RESULTS & DISCUSSION

Relationship between genome size and intergenic expression indicates that intergenic transcripts may generally be non-functional

Transcription of unannotated, intergenic regions can be due to either activities of novel genes or non-functional transcriptional noise. 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 amount of intergenic expression would increase accordingly. By contrast, we expect that the extent of genic sequence expression will not be significantly correlated with genome sizes because larger plant genomes do not necessarily have more genes (r^2=0.01; p=0.56). Thus, to gauge if intergenic transcribed regions (ITRs) generally behave more like what we expect of noisy or genic transcription, we assessed the correlation between genome size and the amount of intergenic expression occurring within a species.

We first identified genic and intergenic transcribed regions using leaf transcriptome data from 15 flowering plant species with 17-fold differences in genome size (Supplemental Table 1). Identical numbers of RNA-sequencing (RNA-seq) reads (30 million) and the same mapping procedures were used in all species to facilitate cross-species comparisons (see Methods). Transcribed regions were considered as ITRs if they did not overlap with any protein-coding or

RNA gene annotation and had no significant translated sequence similarity to plant protein sequences (see Methods). As expected, the amount of expression originating from annotated genic regions has no significant correlation with genomes size (r^2=0.03; p=0.53; **Fig. 1A**). In contrast, the amount of intergenic expression occurring is significantly and positively correlated (r^2=0.30; p=0.04; **Fig. 1B**). Because more intergenic expression is occurring in species with more genome space, this is consistent with the interpretation that a significant proportion of intergenic expression represents transcriptional noise. However, the correlation between genome size and intergenic expression explained ~30% of the variation, 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 non-functional sequences.

Expression, conservation, and epigenetic features are significantly distinct between benchmark functional and non-functional genomic sequences

To determine whether intergenic transcripts resemble functional sequences, we first asked what features may allow benchmark functional and non-functional genomic regions to be distinguished. For benchmark functional sequences, we used genes with visible loss-of-function phenotypes when mutated (referred to as phenotype genes, n=1,876; see Methods). These phenotype genes were considered functional based on the selected effect functionality criterion (Neander 1991) because their mutations have significant growth and/or developmental impact and likely contributes to reduced fitness. For benchmark non-functional genomic regions, we utilized pseudogene sequences (n=761; see Methods). These pseudogenes exhibit sequence similarity to known genes, but harbor disabling mutations including frame shifts and/or in-frame stop codons, that result in the production of presumably non-functional protein products. Considering that only 2% of pseudogenes are maintained over 90 million years of divergence between human and mouse (Svensson et al. 2006), it is expected that the majority of pseudogenes are 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 characteristics, transcription factor (TF) binding, and transcription activity. Feature values (Supplemental Table 2) were calculated for a randomly-selected 500 base pair (bp) window inside a phenotype gene or pseudogene. We used Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) as a metric to measure how well a feature distinguishes between phenotype genes and pseudogenes. AUC-ROC values range between 0.5 (random guessing) and 1 (perfect separation of functional and non-functional sequences), with AUC-ROC values of 0.7, 0.8, and 0.9 considered fair, good, and excellent performance, respectively. Among the seven feature categories, transcription activity features were highly informative (median AUC-ROC=0.88; **Fig. 2A**). 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. 2B-E**). By contrast, chromatin accessibility and sequence-structure features were largely uninformative (median AUC-ROC=0.51 and 0.55, respectively; **Fig. 2F-G**). The poor performance of chromatin accessibility features is likely because the DNase I hypersensitivity (HS) datasets are sparse, as only 2-6% of phenotype gene and pseudogene sequences overlap a DNase I HS site. Further, median nucleosome occupancy nucleosome occupancy of phenotype genes (median normalized nucleosome occupancy = 1.22) is only slightly higher than that of pseudogenes (median = 1.31; Mann Whitney U test, $p < 2e\text{-}4$). For sequence-structure features based on dinucleotide structures (see Methods), we found that poor performance was likely due to phenotype genes and pseudogenes sharing similar dinucleotide sequence compositions (Supplemental Fig. 1).

Error rates for functional region predictions are high when only single features are considered

Within each feature category, there is often a wide range of performance between features (**Fig. 2**, Supplemental Table 3). There are often clear biological or technical explanations for features that perform poorly. For the transcription activity category, 17 features have an AUC-ROC performance >0.8, including the best-performing feature, expression breadth (AUC-ROC=0.95; **Fig. 2A**). However, five transcription activity-related features perform poorly, including the presence of expression (transcript) evidence (AUC-ROC=0.58; **Fig. 2A**). This is because 80% of pseudogenes are considered expressed in ≥1 of 51 RNA-seq datasets, demonstrating that presence of transcripts should not be used by itself as evidence of functionality. For the sequence

conservation category, maximum and average phastCons conservation scores were highly distinct between phenotype genes and pseudogenes (AUC-ROC=0.83 and 0.82, respectively; **Fig. 2B**). On the other hand, identity to best matching nucleotide sequences found in the *Brassicaceae* and algal species were not informative (AUC-ROC=0.55 and 0.51, respectively; **Fig. 2B**). This is because 99.8% and 95% of phenotype genes and pseudogenes, respectively, had a potentially homologous sequence within the *Brassicaceae* family compared to only 3% and 1%, respectively, in algal species. Thus *Brassicaceae* genomes were too similar and algal genomes were too dissimilar to *A. thaliana* to provide meaningful information. H3 mark features also display high variability. The most informative H3 mark features are based on the number and coverage of activation-related marks (AUC-ROC=0.87 and 0.85, respectively; **Fig. 2E**), consistent with the notion that histone marks are often jointly associated with active genomic sequences to potentially provide a robust regulatory signal (Schreiber and Bernstein 2002; Wang et al. 2008). By comparison, the coverage and intensity of H3 lysine 27 trimethylation (H3K27me3) and H3 threonine 3 phosphorylation (H3T3ph) are largely indistinct between phenotype genes and pseudogenes (AUC-ROC range: 0.55-0.59; **Fig. 2E**).

Despite this high variability in performance, some features and feature categories have high AUC-ROCs suggesting that these features may individually provide sufficient information for distinguishing between functional and non-functional genomic regions. To assess this possibility, we next evaluated the error rates of functional predictions based on single features. We first considered expression breadth of a sequence, the best predicting feature of functionality. Despite high AUC-ROC (0.95), the false positive rate (FPR; % of pseudogenes predicted as phenotype genes) is 21% when only expression breadth is used, while the false negative rate (FNR; % of phenotype genes predicted as pseudogenes) is 4%. Similarly, the best-performing H3 mark- and sequence conservation-related features have FPRs of 26% and 32%, respectively, while also incorrectly classifying at least 10% of phenotype genes as pseudogenes. Thus, even when considering well-performing single features, error rates remain high indicating the need to jointly consider multiple features for distinguishing phenotype genes and pseudogenes.

Consideration of multiple features in combination produces accurate predictions of functional genomic regions

To consider multiple features in combination, we first conducted principle component (PC) analysis to investigate how well phenotype genes and pseudogenes could be separated. Between the first two PCs, which jointly explain 40% of the variance in the feature dataset, phenotype genes (**Fig. 3A**) and pseudogenes (**Fig. 3B**) are distributed in largely distinct space. However, there remains substantial overlap, indicating that standard parametric approaches are not well suited to distinguishing between benchmark functional and non-functional sequences. Thus, we instead considered all 93 features in combination using random forest (see Methods), which generated a binary classifier that can be used to predict whether a sequence resembled phenotype genes or pseudogenes. This classifier is referred to as the full model. The phenotype gene and pseudogene sequences and associated conservation, biochemical, and sequence-structure features were separated into distinct training and testing sets such that the full model was generated and validated using independent data subsets (cross-validation). The resulting full prediction model provided much more accurate predictions (AUC-ROC=0.98; FNR=4%; FPR=10%; **Fig. 3C**) compared to any individual feature (**Fig. 2**). An additional measure of performance based on the precision (proportion of predicted functional sequences that are truly functional) and recall (proportion of functional sequences predicted as functional) values among predictions generated by the full model (**Fig. 3D**) also indicate that the model is performing well. When compared to the best-performing single feature (expression breadth), the full model has a similar FNR but only half the FPR (10% compared to 21%). Thus, the full model is more capable of distinguishing between phenotype genes and pseudogenes.

We next determined what the relative contributions of different feature categories were in predicting phenotype genes and pseudogenes and whether models based on a subset of features would perform similarly as the full model. Seven prediction models were established, 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, the models based on transcription activity, sequence conservation, and H3 mark features scored highly (AUC-ROC=0.97, 0.92, and 0.91, respectively; **Fig. 3C**). Particularly, the transcription activity feature category model performed almost as well as the full model (FNR=6%, FPR=12%). We should emphasize that, instead of the presence of expression evidence, other transcription activity-related features are significantly distinct between functional and non-functional regions that produce useful predictions.

Considering that investigating the functionality of ITRs is a primary goal of this study and that ITRs are defined based on the presence of expression evidence, we also built a model did not consider any transcription activity features (full w/o TX, **Fig. 3C-D**). We found that the model excluding transcription activity features performed almost as well as the full model and similarly to the transcription activity-feature-only model although with an increased FPR (AUC-ROC=0.96; FNR=3%; FPR=20%). This indicates that predictions of functional regions are not reliant solely on transcription data, but instead a diverse array of features can be considered 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 distinguishing functional genic and non-functional genomic sequences in plant species with only a modest amount of transcriptome data.

Functional likelihood allows the prediction of functional and non-functional genomic regions

To provide a measure of the potential functionality of any sequence, including ITRs and ncRNAs, in the *A. thaliana* genome, we utilized the confidence score from the full model as a "functional likelihood" value (Tsai et al. 2017; see Methods). The functional likelihood 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 resemble pseudogenes (non-functional). Functional likelihood values for all genomic regions examined in this study are available in Supplemental Table 4. As expected, phenotype genes have high functional likelihood values (median=0.97; **Fig. 4A**) and pseudogenes have low values (median=0.01; **Fig. 4B**). To call sequences as functional or not, we defined a threshold functional likelihood value of 0.35 (see Methods). Using this threshold, 96% of phenotype genes (**Fig. 4A**) and 90% of pseudogenes (**Fig. 4B**) are correctly classified as functional and non-functional, respectively, demonstrating that the full model is highly capable of distinguishing functional and non-functional sequences.

We next applied our model to predict the functionality of annotated protein-coding genes, transposable elements, and random unexpressed intergenic regions. Most annotated protein-coding genes not included in the phenotype gene dataset have high functional likelihood scores (median=0.86; **Fig. 4C**) and 80% are predicted as functional. Of the 20% of protein-coding

genes that were predicted as non-functional, we expect that at least 4% represent false negatives based on the FNR of the full model. The actual FNR among protein-coding genes may be higher, however, as phenotype genes represent a highly active and well conserved subset of all genes. However, a subset of the low-scoring protein-coding genes may also represent gene sequences undergoing functional decay and *en route* to pseudogene status. To assess this possibility, we examined 1,940 *A. thaliana* "decaying" genes that may be experiencing pseudogenization due to promoter disablement (Yang et al. 2011) and found that while they represent only 7% of all *A. thaliana* annotated protein-coding genes, they make up 45% of protein-coding genes predicted as non-functional (Fisher's Exact Test (FET), $p < 1E\text{-}11$). In addition to protein-coding genes, we evaluated the functional likelihoods of transposable elements (TEs) and randomly-selected, unexpressed intergenic regions that are most likely non-functional. As expected, the functional likelihoods were low for both TEs (median=0.03, **Fig. 4D**) and unexpressed intergenic regions (median=0.07; **Fig. 4E**), and 99% of TEs and all unexpressed intergenic sequences were predicted as non-functional, further demonstrating the utility of the function prediction model. Overall, the functional likelihood measure provides a useful metric to distinguish between phenotype genes and pseudogenes. In addition, the functional likelihoods of annotated protein-coding genes, TEs, and unexpressed intergenic sequences agree with *a priori* expectations regarding the functionality of these sequences.

Exclusion of features from multiple tissues increases prediction performance for narrowly-expressed sequences

Although the full model performs exceedingly well, there remain false predictions. There are 76 phenotype genes (4%) predicted as non-functional (referred to as low-FL phenotype genes). We assessed why these phenotype genes were not correctly identified by first asking what category of features were particularly distinct between low-FL and the remaining phenotype genes. We found that the major category that led to the misclassification of phenotype genes was transcription activity, as only 7% of low-scoring phenotype genes were predicted as functional in the transcription activity-only model, compared to 98% of high FL phenotype genes (**Fig. 5A**). By contrast, >65% of low-FL phenotype genes were predicted as functional when sequence conservation, H3 mark, or DNA methylation features were used. This could suggest that the full model is less effective in predicting functional sequences that are weakly or narrowly expressed.

While sequence conservation features are distinct between functional and non-functional sequences when considered in combination, a significantly higher proportion of low-FL phenotype genes were specific to the *Brassicaceae* family, with only 33% present in dicotyledonous species outside of the *Brassicaceae*, compared to 78% of high-scoring phenotype genes (FET, $p < 4e\text{-}12$), thus our model likely has reduced power in detecting lineage-specific genes.

Given the association between transcription activity features and functional predictions, we next investigated how functional predictions performed for conditionally-functional and narrowly-expressed sequences. We found that genes with conditional phenotypes (see Methods) had no significant differences in functional likelihoods (median=0.96) as those with phenotypes under standard growth conditions (median=0.97; U test, p=0.38), indicating that our model can capture conditionally functional sequences. Next, we evaluated functional likelihood distributions among sequences with different breadths of gene expression. For this comparison, we focused on non-stress, single-tissue expression datasets (Supplemental Table 5), which was distinct from the expression breadth feature in the prediction model that considered all datasets. While phenotype genes are better predicted than pseudogenes among sequences with the same number of tissues with expression evidence (U tests, all $p < 1.7E\text{-}06$; Supplemental Fig. 2A), 65% of the 62 phenotype genes expressed in ≤3 tissues are predicted as non-functional. Further, there is a significant correlation between the number of tissues with expression evidence and functional likelihood values of all sequences in our analysis (r^2=0.77; $p < 2E\text{-}16$). Thus, the functional prediction model is biased against narrowly-expressed phenotype genes.

We also found that 80 pseudogenes (10%) were defined as functional (high-FL pseudogenes). Consistent with misclassifications among phenotype genes, a key difference between high-FL pseudogenes and those that were correctly predicted as non-functional was that high-FL pseudogenes tend to be highly and broadly expressed (**Fig. 5A**). A significantly higher proportion of high-FL pseudogenes come from existing genome annotation as 19% of annotated pseudogenes were classified as functional, compared to 4% of pseudogenes identified through a computational pipeline (Zou et al 2009) (FET, $p < 1.5E\text{-}10$). We found that high-FL pseudogenes might be more recently pseudogenized and thus have not yet lost many genic signatures, as the mean number of disabling mutations (premature stop or frameshift) per kb in high-scoring pseudogenes (1.9) was significantly lower than that of low-scoring pseudogenes (4.0; U test, $p <$

0.02). Lastly, we cannot rule out the possibility that a small subset of high-scoring pseudogenes represent truly functional sequences, rather than false positives (e.g. Karreth et al. 2015; Poliseno et al. 2010). Overall, the misclassification of both narrowly-expressed phenotype genes and broadly-expressed pseudogenes highlights the need for an updated prediction model that is less influenced by expression breadth.

To tailor functional predictions to narrowly-expressed sequences, we generated a "tissue-agnostic" model that attempts to minimize the contribution of biochemical activities occurring in many tissues by excluding expression breadth and features that were available across multiple tissues (see Methods). The tissue-agnostic model performed similarly to the full model (AUC-ROC=0.97; FNR=4%; FPR=15%; Supplemental Fig. 3; Supplemental Table 4). Importantly, the proportion of phenotype genes expressed in ≤3 tissues predicted as functional increased by 23% (35% in the full model to 58% in the tissue-agnostic model, Supplemental Fig. 2B), indicating that the tissue-agnostic model is more suitable for predicting the functionality of narrowly-expressed sequences than the full model, although there is an increase in FPR (from 10% to 15%). We next sought to evaluate the functional likelihood of ITR and annotated ncRNA sequences utilizing both the full model and the tissue-agnostic model, in case that these sequences are narrowly-expressed.

Intergenic transcribed regions and annotated ncRNAs are mostly predicted as non-functional

ITRs and ncRNAs represent functionally ambiguous sequences, as they are usually identified by the presence of expression evidence and few have been functionally characterized. Nevertheless, a subset of ITRs likely represent novel genes and may also represent unannotated exon extensions of known genes (Johnson et al. 2005). To evaluate the functionality of ITRs and ncRNAs, we next applied both the full and tissue-agnostic models to these sequences. Additionally, we investigated whether likely-functional ITRs and ncRNAs are close to annotated genes, and if so, if they may be extensions of the gene neighbors. We assessed functional likelihood values for 895 ITRs from three sources: Araport 11 annotation, Moghe et al. (2013), and an additional set identified in this study from 206 RNA-seq datasets. We also analyzed the functional likelihood of TAIR ncRNAs (n=136), and Araport long ncRNAs (referred to as Araport ncRNAs, n=252) TAIR and Araport ncRNAs are collectively referred to as annotated

ncRNAs. The functional likelihoods based on the full model were low (median=0.09) for both ITRs (**Fig. 4F**) and Araport ncRNAs (**Fig. 4G**), and only 15% and 9% of these sequences are predicted as functional, respectively. By contrast, TAIR ncRNAs have higher functional likelihood values (median=0.53; **Fig. 4H**) and 68% are predicted as functional. We next asked what features were distinct among TAIR ncRNAs compared to ITRs and Araport ncRNAs that led to a greater proportion of these sequences predicted as functional and found that transcription activity features of TAIR ncRNAs are more similar to phenotype genes when compared to ITRs and Araport ncRNAs (**Fig. 5B**). By contrast, only 40% of TAIR ncRNAs are predicted as functional if sequence conservation features are considered, potentially because RNA genes experience less selective constraint at the primary sequence level compared to protein-coding genes (Pang et al. 2006). When looking at the performance of single-category predictions, we also find that a greater proportion of ITRs and Araport ncRNAs are predicted as functional when considering only DNA methylation or H3 mark features (**Fig. 5B**). However, these two category-specific models are also marked by increased false positive rates and predict a substantial proportion of unexpressed intergenic sequences as functional (**Fig. 5B**). Notably, 88% of unexpressed intergenic sequences are predicted as functional based on the DNA methylation-only model. Thus, while single-category models are useful for determining features that are similar or dissimilar across sequences types, they may not be useful as a basis for predicting sequences as functional or non-functional.

As ITRs and annotated ncRNAs are generally narrowly-expressed, it is likely that we are underestimating the proportion that is functional. We next applied the tissue-agnostic model to ITRs and annotated ncRNAs, as this model is less biased against narrowly-expressed sequences (Supplemental Fig. 2B). Compared to the full model, twice as many ITRs (30% compared to 15% in the full model; FET, $p < 4E\text{-}15$) and Araport ncRNAs (19% compared to 9%; FET, $p < 0.003$) are predicted as functional. A similar proportion of TAIR ncRNAs are predicted as functional (67% compared to 68%; FET, $p=0.80$), which is likely a result of TAIR ncRNAs being more broadly expressed than ITRs and Araport ncRNAs (Supplemental Fig. 4A). Considering both the full and tissue-agnostic models, we predict a total of 268 ITRs (32%), 57 Araport ncRNAs (23%), and 105 TAIR ncRNAs (77%) as functional.

Intergenic transcripts can represent evidence for unannotated extensions or alternative splicing variants of known genes (Johnson et al. 2005). Thus, we next evaluated whether ITRs

and annotated ncRNAs that are predicted as functional are close to annotated genes and if these sequences share features with neighboring genes. We found that ITRs and annotated ncRNAs closer to annotated genes tend to be predicted as functional (Supplemental Fig. 5A). Using the 95th percentile of intron lengths for all genes as a threshold to call ITRs and annotated ncRNAs as proximal or distant to neighboring genes, 57% of functional ITRs and annotated ncRNAs are considered proximal, compared to 35% for non-functional ITRs and annotated ncRNAs (FET, $p < 2E-09$), suggesting that a subset these likely-functional sequences may be unannotated exons of known genes. If ITRs and annotated ncRNAs represent unannotated extensions of known genes, they may share features with their gene neighbors. However, functional ITRs/ncRNAs have features that bear little similarity to neighboring genes, regardless of if they are proximal or distant to neighboring genes (Supplemental Fig. 5B-C). In contrast, genes are generally more similar to their neighbors, regardless of proximity, than ITRs or annotated ncRNAs are to their nearest neighbor (Supplemental Fig. 5B-C). This is also true compared to random gene pairs (Supplemental Fig. 5D). Thus, despite their proximity to annotated genes, we expect that few ITRs or annotated ncRNAs represent unannotated exon extensions of known genes. For proximal functional ITRs/annotated ncRNAs, we cannot rule out the possibility that they represent false-positive functional predictions due to the accessible and active chromatin states of nearby genes that serve as a confounding factor. For the 116 functional ITRs and annotated ncRNAs that are distal, they may represent fragments of novel genes.

Overall, we find that ITRs and annotated ncRNAs are generally predicted as non-functional. Furthermore, tissue-specific or conditional functionality does not fully explain these non-functional predictions and few predicted-functional ITRs and ncRNAs are likely unannotated extensions of neighboring genes. In addition to the ITRs and ncRNAs investigated thus far, there are 12,344 ITR and ncRNA sequences that are shorter than 500 bp and were unable to be investigated by the full model. We next evaluated methods to assess the functionality of these shorter sequences.

Short RNA genes have mixed predictions based on a binary classification model

The functional predictions performed thus far require 500 bp of sequence. However, there are an additional 10,938 ITRs and 1,406 annotated ncRNAs (12,344 in total) that are shorter than 500 bp. To evaluate the functionality of short ITRs and ncRNAs, we generated a new binary

classification model using features calculated from a randomly-selected 100 bp sequence within a gene or pseudogene body (for features, see Supplemental Table 6). ITRs and annotated ncRNAs tend to be more narrowly expressed than phenotype genes (U tests, all $p < 6e\text{-}15$; Supplemental Fig. 4B) and the tissue-agnostic model was shown to improve false negative rates among low-FL phenotype genes. Therefore, we generated this model while excluding expression breadth and tissue-specific features (referred to as 100 bp tissue-agnostic model). The 100bp tissue-agnostic model performed similarly to the full 500 bp model in distinguishing between phenotype genes and pseudogenes (AUC-ROC=0.97; FNR=13%; FPR=5%; Supplemental Fig. 6). Most importantly, focusing on entries <500 bp in length, this 100 bp model led to the prediction of an additional 366 ITRs (11%), 109 Araport ncRNAs (8%), and 10 TAIR ncRNAs (44%) as functional (Supplemental Fig. 6F-H).

In addition to allowing the evaluation of 12,344 short ITRs and annotated ncRNAs, the 100 bp tissue-agnostic model can be applied to annotated short RNA genes. Thus, we next sought to evaluate functional likelihood scores for Pol II-transcribed RNA genes that have been annotated in TAIR10, including the primary transcripts of microRNAs (miRNAs; n=151), small nucleolar RNAs (snoRNAs; n=15), and small nuclear RNAs (snRNAs; n=6). We found that 15% of miRNAs (Supplemental Fig. 6I), 73% of snoRNAs (Supplemental Fig. 5J), and 50% of snRNAs (Supplemental Fig. 6K) were predicted as functional. Because most TAIR10 annotated RNA genes are computationally predicted and have not been experimentally validated, it is possible that some may represent false positive gene annotations, particularly among miRNA entries. Meanwhile, we cannot rule of the possibility that the 100bp tissue-agnostic model performs sub-optimally for RNA genes. To further assess these possibilities, we identified six RNA genes (four miRNAs, one lncRNA, and one *trans*-acting small interfering RNA) with loss-of-function mutant phenotypes (referred to as RNA phenotype genes; Supplemental Table 7). Of these six genes, we correctly identify three as functional (Supplemental Fig. 6L). Although this is significantly higher than the proportion of pseudogenes (FET, $p < 0.004$) and miRNAs ($p = 0.05$) predicted as functional, this finding suggests that the 100 bp tissue-agnostic model has a substantial false negative rate for detecting functional RNA genes. One immediate question is whether the suboptimal prediction is because RNA genes belong to a class of their own. To further evaluate functional predictions of RNA gene sequences, TAIR ncRNAs, Araport

ncRNAs, and ITRs, we next built multi-class functional prediction models for distinguishing RNA genes from other types of functional and non-functional sequences.

Intergenic transcribed regions and annotated ncRNAs do not resemble benchmark RNA genes

To build a model that considers genomic sequences that are likely functional at the RNA level as a distinct class, we generated a four-class function prediction model aimed at distinguishing four classes of sequences: benchmark RNA genes, phenotype protein-coding genes (same as phenotype genes from previous sections), pseudogenes, and randomly-selected, unexpressed intergenic regions. Here, unexpressed intergenic sequences were included to provide another set of likely non-functional sequences distinct from pseudogenes. The benchmark RNA gene training set was composed of six RNA phenotype genes discussed in the previous section and 40 high-confidence primary miRNA sequences from miRBase (Kozomara and Griffiths-Jones 2014). The model provides four scores, one for each sequence class (for scores, see Supplemental Table 4), and the maximum score was used to classify sequences. We excluded expression breadth and tissue-specific features when generating the four-class model.

Based on predictions from the four-class model, the RNA gene training set was well-classified, with 87% predicted as either RNA gene-like (65%) or phenotype protein-coding gene-like (22%; **Fig. 6A**). Notably, all six RNA phenotype genes were predicted as functional (four and two predicted as RNA genes and phenotype protein-coding genes, respectively). To assess whether sequences predicted as RNA gene-like had evidence of translation, we identified genomic regions with translation evidence based on two shotgun proteomics datasets (Baerenfaller et al. 2008; Castellana et al. 2008). We find that phenotype protein-coding genes and other protein-coding genes predicted as benchmark RNA gene-like are less likely to have evidence of translation compared to those predicted as phenotype protein-coding gene-like (FET, both p<6e-5, Supplemental Fig. 7). Taken together with the predictions of benchmark RNA genes, these results suggest that the benchmark RNA gene prediction score allows sequences that function at the RNA level to be distinguished from other sequence types. For the remaining three classes in the four-class model, 95% of phenotype genes were predicted as either phenotype protein-coding gene-like or benchmark RNA gene-like (**Fig. 6B**), while 70% of pseudogenes (**Fig. 6C**) and 100% of unexpressed intergenic regions (**Fig. 6D**) resembled either pseudogenes

or unexpressed intergenic sequences. Importantly, among phenotype genes expressed in ≤3 tissues, 80% were correctly predicted as phenotype protein-coding or benchmark RNA gene-like in the four-class model, an increase of 22% over the 500 bp tissue-agnostic model.

Since the four-class model was generally able to distinguish benchmark RNA genes from other sequence classes, regardless of breadth of expression, we next evaluated whether ITRs and annotated ncRNAs resemble benchmark RNA genes. We find that 20%, 19%, and 15% of ITRs, Araport ncRNAs, and TAIR ncRNAs, respectively, are predicted as RNA genes (**Fig. 6E-G**). We also considered that ITRs and annotated ncRNAs that were predicted as phenotype protein-coding gene-like may also be functioning at the RNA level. Consistent with this notion, fewer than 5% of phenotype protein-coding gene-like ITRs and annotated ncRNAs have evidence of translation, compared to 37% of phenotype genes and 27% of protein-coding genes (Supplemental Fig. 7). This suggests that the majority of ITRs and annotated ncRNAs predicted as benchmark RNA gene-like or phenotype protein-coding gene-like are likely functional RNA genes.

To provide an overall estimate the proportion of likely-functional ITRs and annotated ncRNAs, we considered the outcome of all four models presented in this study (full 500 bp, 500 bp and 100 bp tissue-agnostic, and four-class models) in combination. We classify 2,453 ITRs (21%) and 506 annotated ncRNAs (28%) as functional, as they resemble phenotype protein-coding genes in at least one of the four models. An additional 1,984 ITRs (17%) and 290 ncRNAs (16%) resemble benchmark RNA genes and therefore could be functional at the RNA level. Ultimately, we find that the majority of ITRs (62%) and annotated ncRNAs (56%) are predicted as non-functional, suggesting that these sequences do not primarily represent novel protein-coding or RNA genes. Moreover, at least a third of ITRs (**Fig. 6E**) and Araport ncRNAs (**Fig. 6F**) are most similar to unexpressed intergenic regions. Given that these sequences have not been functionally characterized, it is possible that many represent regions of noisy transcription and, in the cases of annotated ncRNAs, false positive gene annotations.

CONCLUSION

We identify a collection of evolutionary, biochemical, and sequence-structure signatures that represent defining features of functional genic regions in a plant genome. Considering these features jointly via machine learning methods produces highly accurate predictions that can

distinguish between functional and non-functional genomic regions with low false positive and false negative rates. Expression evidence is particularly distinct between phenotype genes and pseudogenes. However, it is the level and breadth of expression that is important for predictions as most pseudogenes have evidence of expression. In addition, predictions performed without expression evidence also performed well, indicating that functional regions are not defined solely by expression features. We also identified ITRs occurring across 15 diverse land plant species with a wide range of genome sizes and find that the amount of intergenic expression occurring in a species increases with genome size while the amount of genic expression does not. Considering that noisy expression should be expected to increase with additional genome space, this hints that much of the intergenic transcription occurring in a species may be non-functional.

Among the 11,833 ITRs analyzed in this study, we predict 2,453 (21%) are likely functional as they exhibit the biochemical, evolutionary, and sequence-structure characteristics of known functional genomic regions. For annotated ncRNA regions, we classify 506 of 1,794 (28%) as likely-functional. An additional 1,984 ITRs (17%) and 290 ncRNAs (16%) resemble benchmark RNA genes and therefore could be functional at the RNA level. However, the false positive rate among RNA gene predictions could be quite high, as 15% of pseudogenes were predicted as RNA genes. More robust and reliable predictions would be possible if additional benchmark RNA genes with loss-of-function phenotype information were available. Ultimately, the ITRs and annotated ncRNAs that are predicted as functional are likely-genic regions that could be responsible for biological novelties and represent an important component of the functional gene set in *A. thaliana*. Therefore, they should be considered high priority targets in future experimental studies. However, the remaining 7,396 ITRs (63%) and 998 annotated ncRNAs (56%) are most similar to pseudogenes or unexpressed intergenic sequences, suggesting these sequences are likely non-functional and byproducts of transcriptional noise. Given that the majority of ITRs and annotated ncRNAs are predicted as non-functional, we recommend that the null hypothesis for the functionality of expressed intergenic sequences is that they represent transcriptional noise. We do not suggest that all novel intergenic transcription represents non-functional activity, but instead that ITRs should be generally regarded as non-functional until convincing experimental evidence is provided that a transcribed genomic region is functional.

METHODS

Identification of leaf intergenic transcribed regions

RNA-sequencing (RNA-seq) datasets were retrieved from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) for 15 flowering plant species (Supplemental Table 1). All datasets were generated from leaf tissue and sequenced on Illumina HiSeq 2000 or 2500 platforms. Genome sequences and gene annotation files were downloaded from Phytozome v11 (www.phytozome.net; Goodstein et al. 2011) or Oropetium Base v01 (www.sviridis.org; VanBuren et al. 2015). Genome sequences were repeat masked using RepeatMasker v4.0.5 (www.repeatmasker.org) if a repeat-masked version of a genome assembly was not available. Only one end from paired-end read datasets were utilized in downstream processing. Reads were trimmed of low scoring ends and residual adaptor sequences using Trimmomatic v0.33 (Bolger et al. 2014) and mapped to associated genome sequences using Tophat v2.0.13 (Kim et al. 2013). Reads ≥20 nucleotides in length that mapped uniquely within a genome at our mapping threshold were used in further analysis. Thirty million mapped reads were randomly selected from among all datasets for a species and assembled into transcript fragments using Cufflinks v2.2.1 (Trapnell et al. 2010). 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 overvalued. The 1st and 99th percentile of intron lengths in a given gene annotation set were used as the minimum and maximum intron lengths, respectively, for both the TopHat2 and Cufflinks steps. Intergenic transcribed regions (ITRs) were defined by transcript fragments that did not overlap existing gene annotation and did not have significant six-frame translated sequence similarity to annotated plant proteins in Phytozome v10 (BLASTX E-value < 1E-05). To determine the relationship between genome size and number of annotated genes, we calculated the correlation between assembled genome size and gene counts from the first 50 published plant genomes as described by Michael and Jackson (2013).

Arabidopsis thaliana genome annotation

Arabidopsis thaliana protein-coding gene, miRNA gene, snoRNA gene, snRNA gene, ncRNA region, pseudogene, and transposable element annotations were retrieved from The Arabidopsis Information Resource v10 (TAIR10; www.arabidopsis.org; Berardini et al. 2015). Additional miRNA gene and lncRNA region annotations were retrieved from Araport v11

(www.araport.org; Krishnakumar et al. 2015). A pseudogene-finding pipeline similar to that described by Chen et al. (Zou et al. 2009) was used to identify additional putative pseudogene fragments and count the number of disabling mutations (early stop or frameshift mutations) present in these sequences. To avoid potential confounding effects from overlapping gene annotation, protein-coding and RNA gene annotation that overlapped other gene or pseudogene annotation were excluded from further analysis, except for lncRNA annotation that overlapped with other lncRNAs, which were merged. Pseudogenes and transposable elements that overlapped genic regions were also removed. When pseudogenes from TAIR10 and the pseudogene-finding pipeline overlapped, the longer pseudogene annotation was retained.

ITRs were defined by Moghe et al. (2013; "Set 2" ITRs; coordinates provided by the authors) and Araport v11 (described as "novel transcribed regions"). Overlapping ITR annotations from Araport were merged. Additional ITRs were identified from 206 RNA-seq datasets generated using wild-type, Columbia-0 tissue on Illumina sequencing platforms (Supplemental Table 5). Datasets were identified by querying NCBI-SRA for datasets from *A. thaliana* with RNA as the source. Reads were trimmed, mapped, and assembled into transcript fragments using the steps described in the previous section, except that reads from multiple datasets were not merged and subsampled. Instead, overlapping assembled transcript fragments from across datasets were merged. ITRs were identified by transcribed fragments that did not overlap with any annotated feature from TAIR10 or Araport11 or any pseudogenes defined by the pseudogene-finding pipeline. Overlaps among ITR annotations were resolved using a priority system: Araport11 > Moghe et al. > ITRs identified in this study.

For each gene, ncRNA, pseudogene, transposable element, and intergenic transcribed sequence, a randomly-selected 100 and 500 base pair (bp) window was chosen for feature calculation (Supplemental Table 2; Supplemental Table 6; see below for feature descriptions). Sequences that were not at least 100 or 500 bp in length were excluded. This controlled for effects of sequence length and simplified gene structure considerations (e.g. exon/intron boundaries). In addition, random 100 bp (n=4,000) and 500 bp (n=3,716) regions of intergenic space (genome regions outside of gene, pseudogene, or transposable element annotation) that did not overlap with any genic or intergenic transcript fragments were also selected for feature calculation. These 100 and 500 bp windows in gene, ncRNA, pseudogene, transposable element,

and ITR annotation and unexpressed intergenic space are referred to as “feature regions” throughout the Methods section.

Single-feature prediction performance

The ability for single features to distinguish between functional and non-functional regions was tested using Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) values calculated using the scikit-learn package in Python. AUC-ROC values range between 0.5 (equivalent to random guessing) and 1 (perfect predictions) and values above 0.7, 0.8, and 0.9 are considered to be fair, good, and excellent, respectively. Thresholds to predict sequences as functional or non-functional using a single feature were defined by the feature value that produced the highest F-measure (harmonic mean of precision and recall), which gives consideration to both false positives and false negatives at a given threshold. False positive rates (FPR) were calculated as the percentage of negative cases with values above or equal to the threshold and false negative rates (FNR) were calculated was the percentage of positive cases with values below the threshold.

Phenotype data sources

Mutant phenotype data for *Arabidopsis thaliana* protein-coding genes was collected from a published dataset (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 (i.e. non-stress conditions). Genes with documented mutant phenotypes under standard conditions were considered as a distinct and non-overlapping category from other annotated protein-coding genes. We identified six RNA genes with documented loss-of-function phenotypes through literature searches: *At4* (AT5G03545; Shin et al. 2006), *MIR164A* and *MIR164D* (AT2G47585 and AT5G01747, respectively; Guo et al. 2005), *MIR168A* (AT4G19395; Li et al. 2012b), and *MIR828A* and *TAS4* (AT4G27765 and AT3G25795, respectively; Hsieh et al. 2009). An additional 23 RNA genes with documented overexpression mutant phenotypes were identified from the literature links at miRBase. Conditional phenotype genes were those belonging to the “Conditional” phenotype class as described by Lloyd and Meinke (2012). These genes had no

obvious mutant phenotype under standard growth conditions, but did exhibit a loss-of-function phenotype under stress conditions. These were compared with phenotype genes belonging to the "Morphological" phenotype class from the same study, which have visible growth defects under standard growth conditions.

Sequence conservation and structure features

Nucleotide diversity and Tajima's D were calculated among 81 *A. thaliana* accessions (Cao et al. 2011) for each feature region using custom Python scripts. The genome matrix file for the Cao et al. study was retrieved from the 1,001 genomes database (www.1001genomes.org) and analyzed with Python scripts available through GitHub (github.com/panchyni/GenomeMatrixProcessing). The genomic regions that align between *A. thaliana* and six other plant species were retrieved from Li et al. (2012a). The coverage of each feature region with these aligned blocks was calculated. In addition, phastCons conservation scores were available for each nucleotide within an aligned block. The maximum and average of phastCons scores were calculated for each feature region. Nucleotides in a feature region that did not overlap with an aligned block were assigned a phastCons score of 0. BLASTN searches were performed between feature region nucleotide sequences and Phytozome v10 genome sequences. Five plant lineages were considered: *Brassicaceae* ($n_{species}$=7), other dicotyledonous plants (n=22), monocots (n=7), other embryophyte plants (n=3), and algae (n=5). The percent identity to the most significant match by E-value (maximum E-value: 1E-05) within a lineage group for each feature region was used as the feature in functional predictions. DNA sequence-structure features consisted of the first five principal components of the 125 conformational and thermodynamic dinucleotide properties collected from DiProDB database (Friedel et al. 2009). The first five principal components (83% of variation) correspond primarily to DNA major groove geometry, free energy, twist and roll, DNA minor groove geometry, and tilt and rise, respectively (Tsai et al. 2015). Sequence-structure values corresponding to principal components were calculated in dinucleotide windows and averaged across the length of a feature region.

Transcription activity features

From the 206 *A. thaliana* RNA-seq datasets described above, we removed datasets with fewer than 20 million reads (n=134) or abnormally high RPKM distributions among resulting transcript

fragments (n=21; median of median RPKM values among retained and removed datasets=10.2 and 4065.2, respectively; Supplemental Table 5), which indicated technical issues during the read cleaning, read mapping, or transcript assembly processes. Transcript fragments assembled from the remaining 51 RNA-seq datasets were used to calculate expression breadth across datasets, 95th percentile RPKM expression levels, maximum transcription coverage in a single dataset, and presence or absence of expression evidence. Ten datasets from diverse tissues and conditions with a high number of reads were chosen to calculate max RPKM expression levels and transcription coverage as single-dataset features. The tissues and conditions included: pollen (SRR847501), light- and dark-grown seedlings (SRR1020621 and SRR974751, respectively), leaf tissue under standard, drought, and fungal-infection conditions (SRR953400, SRR921316, and SRR391052, respectively), root (SRR578947), inflorescence (SRR953399), flower (SRR505745), and silique (SRR953401). RNA-seq datasets generated from a single tissue and in standard growth conditions were used for tissue-specific expression analysis. The seven tissues were pollen, seedling, leaf, root, inflorescence, flower, and silique (Supplemental Table 5). Two additional datasets generated by sequencing RNA molecules associated with ribosomes (SRR966480 and SRR966484) were retrieved from NCBI-SRA and processed using the same steps as those used on other RNA-seq datasets.

Histone 3 mark features

Chromatin immunoprecipitation sequencing (ChIP-seq) datasets for four activation-associated (H3K4me1: SRR2001269, H3K4me3: SRR1964977, H3K9ac: SRR1964985, and H3K23ac: SRR1005405) and four repression-associated (H3K9me1: SRR1005422, H3K9me2: SRR493052, H3K27me3: SRR3087685, and H3T3ph: SRR2001289) histone 3 (H3) marks were retrieved from NCBI-SRA. Datasets were chosen due to high number of reads and presence of histone 3 or total protein controls. Reads were trimmed with Trimmomatic v0.33 (Bolger et al. 2014) and mapped to the TAIR10 genome sequence with Bowtie v2.2.5 (Langmead et al. 2009). H3 mark peaks were identified with the Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) software v1.1 (Xu et al. 2014). SICER requires an effective genome size input, which was calculated according to Koehler et al. (2011). The maximum H3 mark peak intensity and the coverage with each H3 mark peak were calculated for each feature region. The count and coverage of all activating or repressing marks in a feature region were also calculated.

DNA methylation features

Bisulfite-sequencing (BS-seq) datasets from seven tissues (pollen: SRR516176, embryo: SRR1039895, endosperm: SRR1039896, seedling: SRR520367, leaf: SRR1264996, root: SRR1188584, and inflorescence: SRR2155684) were retrieved from NCBI-SRA. BS-seq reads were trimmed with Trimmomatic v0.33 (Bolger et al. 2014) and processed with Bismark v3 (Krueger and Andrews 2011). A cytosine was considered to be methylated if at least five reads mapped to the position and >50% of the reads indicated the position was methylated. For each feature region, the percentage of methylated cytosines in CpG, CHG, and CHH contexts were calculated if the feature region had ≥5 cytosines with ≥5 reads mapping to the position. To test the false positive rate of DNA methylation calls, we evaluated the proportion of cytosines in the chloroplast genome that are called as methylated, as the chloroplast genome is known to have few DNA methylation events (Ngernprasirtsiri et al. 1988; Zhang et al. 2006). In any nucleotide context for any BS-seq dataset, 0-1.5% (median=0) of cytosines in the chloroplast genome were defined as methylated and only 0.1-2.4% of reads suggested that a cytosine position was methylated. This indicated that the false positive rates for DNA methylation calls were low.

Chromatin accessibility and transcription factor binding features

Chromatin accessibility features consisted of DNase I hypersensitive peaks and micrococcal nuclease sequencing (MNase-seq)-derived nucleosome occupancy. DNase I peaks from five tissues (seed coat, seedling, root, unopened flowers, and opened flowers) were available from the Gene Expression Omnibus (experiment identifiers: GSE53322 and GSE53324; Sullivan et al. 2014). The max DNase I peak intensity and coverage with DNase I peaks were calculated for each feature region. MNase-seq nucleosome occupancy was produced by Liu et al. (2015). The authors provided the normalized nucleosome occupancy for each nucleotide in the TAIR10 genome sequence. The average of nucleosome occupancy values across a feature region was calculated. Transcription factor (TF) binding sites were identified from *in vitro* DNA affinity purification sequencing data of 529 TFs (O'Malley et al. 2016). The total number of TF binding sites and the number of distinct TFs bound were calculated for each feature region.

Machine learning approach

For two-class models using 500 bp sequences (positive class: phenotype genes, n=1,876; negative class: pseudogenes, n=763), the random forest (RF) implementation in the Waikato Environment for Knowledge Analysis software (WEKA; Hall et al. 2009) was utilized. We generated 100 datasets with an equal proportion of phenotype genes and pseudogenes by randomly selecting 763 phenotype genes and pairing them with all 763 pseudogene examples. For each of these 100 datasets, 10-fold stratified cross-validation was utilized during model building and testing. Therefore, training and testing of each model was performed on independent datasets. The median score from the 100 prediction models was used as the final functional prediction score ("functional likelihood"). Five hundred trees using 2, 4, 6, 9, 15, 20, and 25 randomly-selected features were built using the RF algorithm. Fifteen features provided the highest performance, as determined by AUC-ROC (calculated and visualized using the ROCR package; Sing et al. 2005). The same methods were used to test two-class RF models using 100 bp sequences (phenotype genes, n=1,882; pseudogenes, n=3,916), except that 100 datasets with equal proportions of phenotype genes and pseudogenes were generated by randomly-selecting 1,882 pseudogenes to pair with all 1,882 phenotype gene examples. In single-category predictions, fewer features were considered in parameter searches. For the H3 mark, DNA methylation, and transcription activity categories 2, 4, 7, and 10 features were tested. For the chromatin accessibility and sequence conservation categories 2, 4, and 6 features were tested. For the sequence-structure category 2, 3, and 4 features were tested. For the transcription factor binding category 1 and 2 features were tested. A tissue-agnostic model was generated by excluding the expression breadth feature and all features from tissue-specific RNA-seq, BS-seq, and DNase I hypersensitivity datasets. Tissue-specific features were replaced with the maximum FPKM and coverage from RNA-seq datasets, minimum DNA methylation proportion from any one tissue in CpG, CHG, and CHH contexts, and maximum intensity and coverage with DNaseI peaks in a single tissue.

The functional likelihood of a genomic sequence was calculated as the proportion of the 500 random forest trees that predicted a sequence as similar to a phenotype gene (Supplemental Table 4). The functional likelihood threshold to predict a sequence as functional or non-functional was defined based on the functional likelihood value that produced the maximum F-measure among all possible thresholds. F-measure is the harmonic mean of precision (proportion of predicted positive regions that are truly positive) and recall (proportion of truly positive

regions that are predicted as positive), which gives consideration to both false positives and false negatives. FPR was calculated as the percentage of pseudogenes with functional likelihood values above or equal to the functional threshold, while FNR was calculated as the percentage of phenotype genes with functional likelihood values below the threshold. Functional prediction models were also built using the Sequential Minimal Optimization - Support Vector Machine (SMO-SVM) implementation in WEKA while considering a series of complexity constant parameters: 0.01, 0.1, 0.5 (best by AUC-ROC), 1, 1.5, and 2.0. The results of SMO-SVM models were highly similar to the RF results: *PCC* between RF and SMO-SVM=0.97; AUC-ROC of SMO-SVM=0.97; FPR=12%; FNR=3%.

For the four-class model, phenotype gene, pseudogene, random unexpressed intergenic sequences, and RNA training genes were used as training classes. RNA training genes consisted of six RNA genes with documented loss-of-function phenotypes and high-confidence miRNA genes from miRBase (www.mirbase.org; Kozomara and Griffiths-Jones 2014) Random-sampling of the more populated classes in training cases was used to produce 250 datasets with equal proportions of phenotype genes, pseudogenes, intergenic sequences, and RNA training genes. Two-fold stratified cross-validation was utilized due to the low number of RNA training gene examples. The features described from the tissue-agnostic model above were also used for the four-class model. The random forest implementation in the *party* package of R with conditional inference trees method utilized was used to build the random forest classifiers. The four-class predictions provide prediction scores for each sequence type: a phenotype gene, pseudogene, unexpressed intergenic, and RNA gene score (Supplemental Table 4). The scores indicate the proportion of random forest trees that predict a given sequence as a phenotype gene, pseudogene, unexpressed intergenic, or RNA gene sequence. The median prediction score from across 100 equal-proportion runs was used as the final prediction scores, which were then scaled to sum to 1. The maximum prediction score was used to classify a sequence as phenotype gene, pseudogene, unexpressed intergenic, or RNA gene.

FIGURE LEGENDS

Figure 1. Relationship between genome size and extent of expression in 15 plant species. (*A*) Amount of expression from annotated gene regions plotted against the size of assembled genome

for 15 diverse flowering plant species. The dotted gray line indicates the line of best fit. (*B*) Amount of expression from intergenic regions plotted against the size of assembled genome.

Figure 2. Single feature predictions of functional and non-functional sequences. Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) prediction performances using single features in the categories of transcription activity (*A*), sequence conservation (*B*), DNA methylation (*C*), transcription factor binding (*D*), histone 3 (H3) marks (*E*), sequence structure (*F*), and chromatin accessibility (*G*). AUC-ROC ranges in value from 0.5 (equivalent to random guessing) to 1 (perfect predictions), with values greater than 0.7, 0.8, and 0.9 being considered fair, good, and excellent, respectively. Dotted gray lines indicate the median AUC-ROC within a feature category.

Figure 3. Multi-feature predictions of functional and non-functional sequences. Smoothed scatterplots of the first two principle components (PCs) of phenotype gene (*A*) and pseudogene (*B*) features. The percentages on the axes in (A) indicate the amount of total variation present in the associated PC. (*C*) Receiver operating characteristic curves of machine learning integration of all features (Full model), all non-transcription activity-related features (Full w/o TX), and when using all features from a single feature category. Single categories are transcription activity (TX), sequence conservation (CV), histone 3 marks (HM), DNA methylation (ME), transcription factor binding (TF), chromatin accessibility (CA), and sequence structure (ST). (*C*) Precision-recall curves of the models from (*B*).

Figure 4. Functional likelihood scores from the full, binary model. Functional likelihood distributions for (*A*) phenotype genes, (*B*) pseudogenes, (*C*) protein-coding genes, (*D*) transposable elements, (*E*) random unexpressed intergenic sequences, (*F*) intergenic transcribed regions, (*G*) ncRNAs from Araport11, and (*H*) ncRNAs from TAIR10 from models built using features calculated from 500 bp of sequence. Higher functional likelihood values indicate greater similarity to phenotype genes while lower values indicate similarity to pseudogenes. Vertical dashed lines display the threshold to predict a sequence as functional or non-functional. The numbers to the left and right of the dashed line show the percentage of sequences predicted as functional or non-functional, respectively.

Figure 5. Functional predictions from single-category predictions. (*A*) Percentages of phenotype gene and pseudogene sequences predicted as functional (high FL) or non-functional (low FL) in the full model (Full) that are predicted as functional in models based on a subset of features from a single feature category. Single feature categories are transcription activity (TX), sequence conservation (CV), histone 3 marks (HM), DNA methylation (ME), transcription factor binding (TF), chromatin accessibility (CA), and sequence structure (ST). The single category models are sorted from right to left on descending AUC-ROC and separated into informative (all AUC-ROC ≥ 0.87) and uninformative (all AUC-ROC ≤ 0.70) groups. (*B*) Percentages of sequence classes predicted as functional based on the same models in (A). ITR indicates intergenic transcribed regions.

Figure 6. Phenotype gene, pseudogene, unexpressed intergenic, and RNA gene score distributions from four-class predictions. Stacked bar plots indicate the phenotype protein-coding gene (dark blue), RNA gene (light blue), pseudogene (red), intergenic (yellow) score for each (*A*) RNA training set gene, (*B*) phenotype gene, (*C*) pseudogene, (*D*) random unexpressed intergenic region, (*E*) intergenic transcribed region, (*F*) ncRNA from Araport11, and (*G*) ncRNA from TAIR10. Black vertical lines indicate boundaries of classification regions, with sequences classified according to highest prediction score. Numbers within or pointing toward a classification regions within a chart indicate the percentage of sequences predicted as, in order phenotype gene, RNA gene, pseudogene, or intergenic. The color bars at the bottom of the chart indicate whether a region of the chart is considered phenotype protein-coding gene-like (dark blue), RNA gene-like (light blue), pseudogene-like (red), or intergenic-like (yellow).

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Relationship between dinucleotide frequencies in phenotype gene and pseudogene sequences. Percentages of all 16 dinucleotides in phenotype genes (X-axis) and pseudogenes (Y-axis). Gray dotted line indicates the line of best fit.

Supplemental Figure 2. Functional likelihood scores by expression breadth. Distributions of functional likelihood scores for phenotype genes (blue) and pseudogenes (red) for sequences expressed in one-to-seven tissues for (*A*) the full model and (*B*) a tissue-agnostic model

generated while excluding the expression breadth feature and merging tissue-specific features. The tissue-agnostic model performs better for among narrowly-expressed phenotype genes.

Supplemental Figure 3. Functional likelihood scores from the 500 bp tissue-agnostic model. Functional likelihood distributions for (*A*) phenotype genes, (*B*) pseudogenes, (*C*) protein-coding genes, (*D*) transposable elements, (*E*) random unexpressed intergenic sequences, (*F*) intergenic transcribed regions, (*G*) ncRNAs from Araport11, and (*H*) ncRNAs from TAIR10 from the tissue-agnostic model built while excluding the expression breadth and tissue-specific features. Features were calculated from a random 500 bp region from within the sequence body. Higher functional likelihood values indicate greater similarity to phenotype genes while lower values indicate similarity to pseudogenes. Vertical dashed lines display the threshold to predict a sequence as functional or non-functional. The numbers to the left and right of the dashed line show the percentage of sequences predicted as functional or non-functional, respectively.

Supplemental Figure 4. Expression breadth of sequence types. Expression breadth distributions for sequence types from (*A*) 500 bp feature regions and (*B*) 100 bp feature regions.

Supplemental Figure 5. ITR and annotated ncRNA distance to and feature similarity with neighboring genes. (*A*) Distance from intergenic transcribed regions (ITRs) and annotated ncRNAs that are predicted as functional (F) or non-functional (NF) to the closest neighboring gene. (*B,C,D*) Feature similarity based on Pearson's Correlation Coefficients between (*A*) proximal neighbors (within 95th percentile of intron lengths; distance=456), (*B*) distal neighbors (greater than 95th percentile of intron lengths), and (*C*) random pairs of ITRs, ncRNAs from Araport11, and ncRNAs from TAIR10 and annotated genes, as well as pairs of annotated genes. Pairs involving ITRs and annotated ncRNAs were further divided by whether the ITR or ncRNA sequence was predicted as functional (F) or non-functional (NF). Features were quantile normalized prior to calculating correlations.

Supplemental Figure 6. Functional likelihood scores from the 100 bp tissue-agnostic model. Functional likelihood distributions for (*A*) phenotype genes, (*B*) pseudogenes, (*C*) protein-coding genes, (*D*) transposable elements, (*E*) random unexpressed intergenic sequences, (*F*) intergenic transcribed regions (ITR), (*G*) ncRNAs from Araport11, (*H*) ncRNAs from TAIR10, (*I*) microRNAs, (*J*) small nucleolar RNAs, (*K*) small nuclear RNAs, and (*L*) RNA genes with

documented loss-of-function phenotypes from the tissue-agnostic model built while excluding the expression breadth and tissue-specific features. Features were calculated from a random 100 bp region from within the sequence body. Higher functional likelihood values indicate greater similarity to phenotype genes while lower values indicate similarity to pseudogenes. Vertical dashed lines display the threshold to predict a sequence as functional or non-functional. The numbers to the left and right of the dashed line show the percentage of sequences predicted as functional or non-functional, respectively.

Supplemental Figure 7. Translation evidence for sequences predicted as phenotype protein-coding gene-like and RNA gene-like. Translation evidence was based on sequence overlap in two shotgun proteomics datasets.

SUPPLEMENTAL TABLES

Supplemental Table 1. Leaf tissue RNA-sequencing datasets for 15 flowering plant species

Supplemental Table 2. Conservation, biochemical, and sequence-structure features calculated from 500 bp sequences.

Supplemental Table 3. False positive and false negative rates for single feature classifications.

Supplemental Table 4. Predictions for the full, tissue-agnostic, 100 bp, and four-class models.

Supplemental Table 5. RNA-sequencing datasets for identifying intergenic transcribed regions, calculating transcription activity features, and assessing tissue-specific predictions.

Supplemental Table 6. Conservation, biochemical, and sequence-structure features calculated from 100 bp sequences.

Supplemental Table 7. RNA genes with documented loss-of-function phenotypes.

DATA ACCESS

All data are available in the text of this article or in the supplemental material.

ACKNOWLEDGEMENTS

The authors wish to thank (in reverse alphabetical order - we do not wish our acknowledgees to feel superior or inferior to one another) Sahra Uygun, Bethany Moore, Gaurav Moghe, Ming-Jung Liu, and Christina Azodi for providing highly useful data used in this study. This work was supported by the Michigan State University Dissertation Continuation Scholarship to J.P.L, the National Science Foundation (NSF) Plant Genomics Research Experience for Undergraduates support to R.P.S., the Taiwan Ministry of Science and Technology Postdoctoral Research Abroad Program MOST-104-2917-I-564-070 to Z.T.-Y.T, and NSF grants (MCB–1119778, IOS-1126998, and IOS-1546617) to S.-H.S.

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.

DISCLOSURE DECLARATION

The authors have no conflicts of interest to disclose.

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**: 529–540.

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

Bennetzen JL. 2005. Mechanisms of Recent Genome Size Variation in Flowering Plants. *Ann Bot* **95**: 127–132.

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**: 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**: 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**: 1441–1450.

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

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.

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**: 393–399.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**: 1915–1927.

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**: 956–963.

Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, Suzuki AM, Fletcher AR, Plachetzki DC, FitzGerald PC, et al. 2014. Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Res* **24**: 1209–1223.

Cummins R. 1975. Functional Analysis. *J Philos* **72**: 741.

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

Eddy SR. 2013. The ENCODE project: missteps overshadowing a success. *Curr Biol* **23**: R259–61

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

Friedel M, Nikolajewa S, Sühnel J, Wilhelm T. 2009. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res* **37**: D37–40.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2011. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–D1186.

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**: 578–590.

Gulko B, Gronau I, Hubisz MJ, Siepel A. 2014. *Probabilities of Fitness Consequences for Point Mutations Across the Human Genome*. http://dx.doi.org/10.1101/006825.

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**: 1376–1386.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**: 223–227.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter* **11**: 10.

Hanada K, Zhang X, Borevitz JO, Li W-H, Shiu S-H. 2007. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res* **17**: 632–640.

Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, et al. 2013. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* **110**: 2395–2400.

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**: 231–247.

Heinen TJAJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a New Gene from an Intergenic Region. *Curr Biol* **19**: 1527–1531.

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**: 2120–2132.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juárez MJA, Simpson J, et al. 2013. Architecture and evolution of a minute plant genome. *Nature* **498**: 94–98.

Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR. 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**: 533–538.

Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21**: 93–102.

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**: 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**: 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**: R36.

Koehler R, Issac H, Cloonan N, Grimmond SM. 2011. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics* **27**: 272–274.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**: D68–73.

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**: D1003–9.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

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**: 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**: e0125522.

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**: R25.

Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012a. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* **24**: 4346–4359.

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**: 1182–1195.

Li W, Cui X, Meng Z, Huang X, Xie Q, Wu H, Jin H, Zhang D, Liang W. 2012b. Transcriptional regulation of Arabidopsis MIR168a and argonaute1 homeostasis in abscisic acid and abiotic stress responses. *Plant Physiol* **158**: 1279–1292.

Li WH, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* **292**: 237–239.

Lloyd J, Meinke D. 2012. A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. *Plant Physiol* **158**: 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**: 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**: 156–166.

Mattick JS. 2009. The Genetic Signatures of Noncoding RNAs. *PLoS Genet* **5**: e1000459.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**: 155–159

Michael TP, Jackson S. 2013. The First 50 Plant Genomes. *Plant Genome* **6**: 0.

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**: 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* **320**: 1344–1349.

Neander K. 1991. Functions as Selected Effects: The Conceptual Analyst's Defense. *Philos Sci* **58**: 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**: 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**: 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**: 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**: 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.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**: 1–5.

Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. 1996. Requirement for Xist in X chromosome inactivation. *Nature* **379**: 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**: 1033–1038.

Ponting CP, Belgard TG. 2010. Transcribed dark matter: meaning or myth? *Hum Mol Genet* **19**: R162–R168.

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**: e73291.

Schnable JC, Pedersen BS, Subramaniam S, Freeling M. 2011. Dose–Sensitivity, Conserved Non-Coding Sequences, and Duplicate Gene Retention Through Multiple Tetraploidies in

the Grasses. *Front Plant Sci* **2**. http://dx.doi.org/10.3389/fpls.2011.00002.

Schnable JC, Wang X, Pires JC, Freeling M. 2012. Escape from preferential retention following repeated whole genome duplications in plants. *Front Plant Sci* **3**: 94.

Schreiber SL, Bernstein BE. 2002. Signaling Network Model of Chromatin. *Cell* **111**: 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**: 712–726.

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

Slotkin RK, Keith Slotkin R, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**: 272–285.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Depamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot* **96**: 336–348.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, et al. 2013. Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution. *Science* **342**: 1367–1372.

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**: 4453–4458.

Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**: 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**: 2015–2030.

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

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**: 511–515.

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**: e1004418.

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*. http://dx.doi.org/10.1093/molbev/msx101

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**: 1537–1550.

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**: 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**: 897–903.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. 2010. Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. *PLoS Biol* **8**: e1000409.

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* **302**: 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**: 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**: 1189–1201.

Zhao J, Sun BK, Erwin JA, Song J-J, Lee JT. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**: 750–756.

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**: D203–8.

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**: 3–15.

FIGURES

Figure 1.

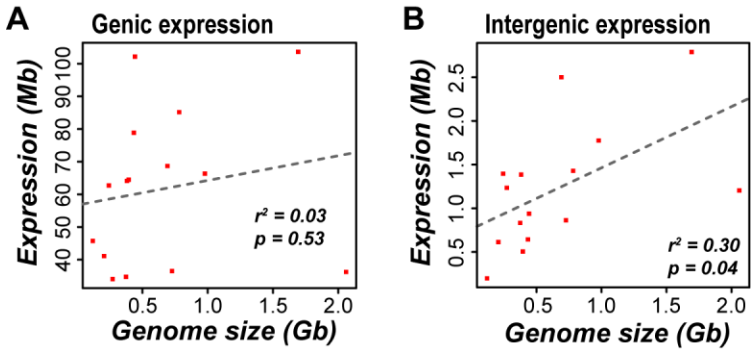


Figure 2.

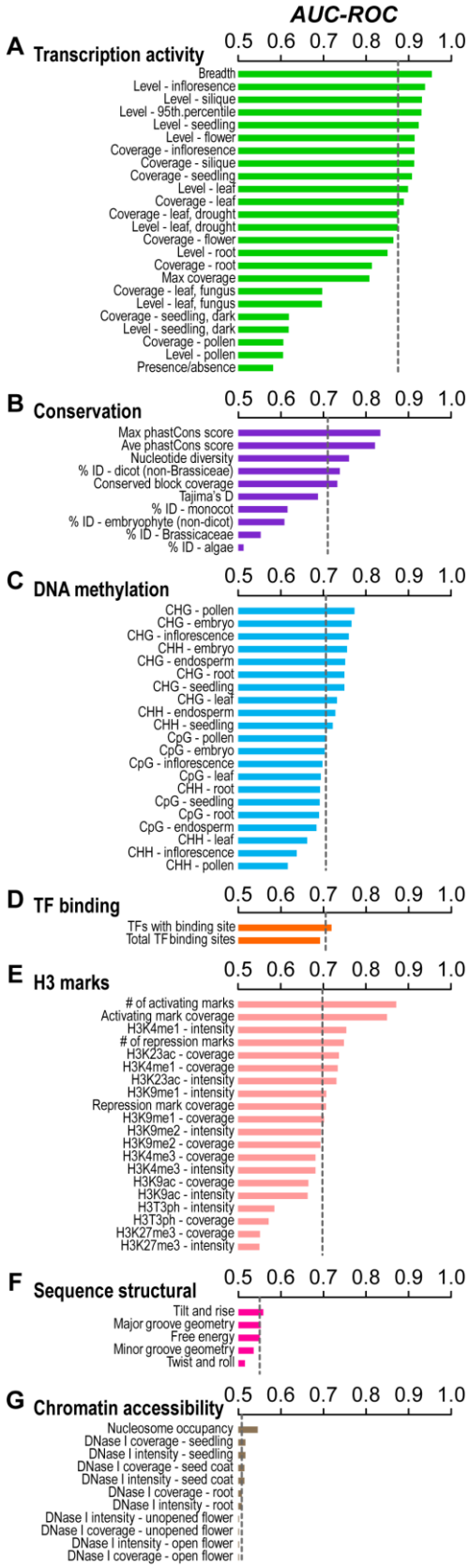

AUC-ROC
0.5 0.6 0.7 0.8 0.9 1.0
A Transcription activity
Breadth
Level - inflorescence
Level - silique
Level - 95th.percentile
Level - seedling
Level - flower
Coverage - inflorescence
Coverage - silique
Coverage - seedling
Level - leaf
Coverage - leaf
Coverage - leaf, drought
Level - leaf, drought
Coverage - flower
Level - root
Coverage - root
Max coverage
Coverage - leaf, fungus
Level - leaf, fungus
Coverage - seedling, dark
Level - seedling, dark
Coverage - pollen
Level - pollen
Presence/absence
B Conservation
Max phastCons score
Ave phastCons score
Nucleotide diversity
% ID - dicot (non-Brassiceae)
Conserved block coverage
Tajima's D
% ID - monocot
% ID - embryophyte (non-dicot)
% ID - Brassicaceae
% ID - algae
C DNA methylation
CHG - pollen
CHG - embryo
CHG - inflorescence
CHH - embryo
CHG - endosperm
CHG - root
CHG - seedling
CHG - leaf
CHH - endosperm
CHH - seedling
CpG - pollen
CpG - embryo
CpG - inflorescence
CpG - leaf
CHH - root
CpG - seedling
CpG - root
CpG - endosperm
CHH - leaf
CHH - inflorescence
CHH - pollen
D TF binding
TFs with binding site
Total TF binding sites
E H3 marks
# of activating marks
Activating mark coverage
H3K4me1 - intensity
# of repression marks
H3K23ac - coverage
H3K4me1 - coverage
H3K23ac - intensity
H3K9me1 - intensity
Repression mark coverage
H3K9me1 - coverage
H3K9me2 - intensity
H3K9me2 - coverage
H3K4me3 - coverage
H3K4me3 - intensity
H3K9ac - coverage
H3K9ac - intensity
H3T3ph - intensity
H3T3ph - coverage
H3K27me3 - coverage
H3K27me3 - intensity
F Sequence structural
Tilt and rise
Major groove geometry
Free energy
Minor groove geometry
Twist and roll
G Chromatin accessibility
Nucleosome occupancy
DNase I coverage - seedling
DNase I intensity - seedling
DNase I coverage - seed coat
DNase I intensity - seed coat
DNase I coverage - root
DNase I intensity - root
DNase I intensity - unopened flower
DNase I coverage - unopened flower
DNase I intensity - open flower
DNase I coverage - open flower

Figure 3.

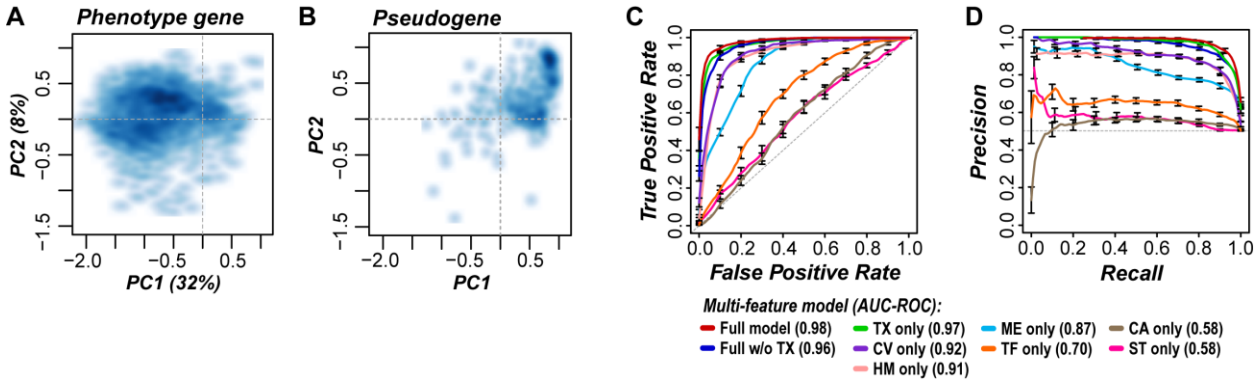

A
Phenotype gene
PC2 (8%)
PC1 (32%)
B
Pseudogene
PC2
PC1
C
True Positive Rate
False Positive Rate
D
Precision
Recall
Multi-feature model (AUC-ROC):
Full model (0.98)
Full w/o TX (0.96)
TX only (0.97)
CV only (0.92)
HM only (0.91)
ME only (0.87)
TF only (0.70)
CA only (0.58)
ST only (0.58)

Figure 4.

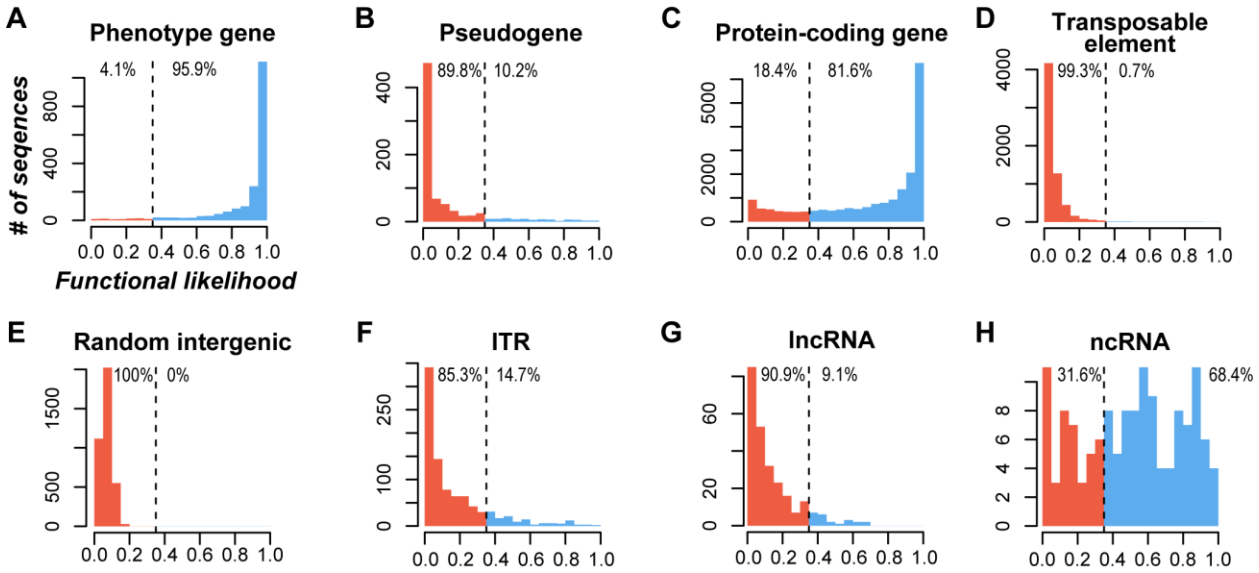

A
Phenotype gene
4.1%
95.9%
# of seqences
Functional likelihood
B
Pseudogene
89.8%
10.2%
C
Protein-coding gene
18.4%
81.6%
D
Transposable element
99.3%
0.7%
E
Random intergenic
100%
0%
F
ITR
85.3%
14.7%
G
lncRNA
90.9%
9.1%
H
ncRNA
31.6%
68.4%

Figure 5.

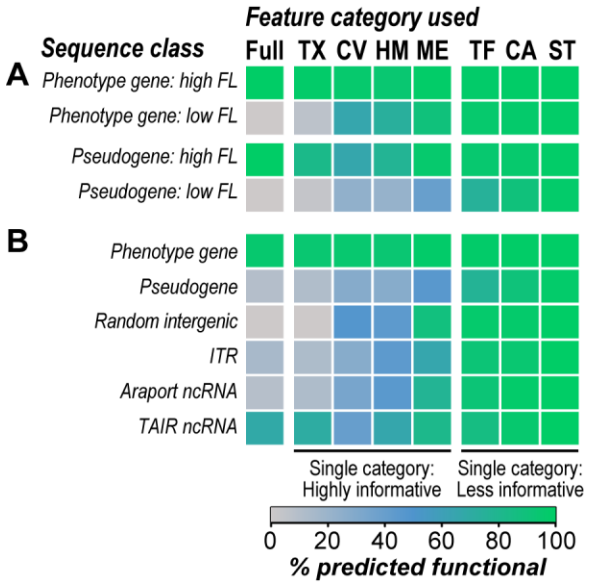
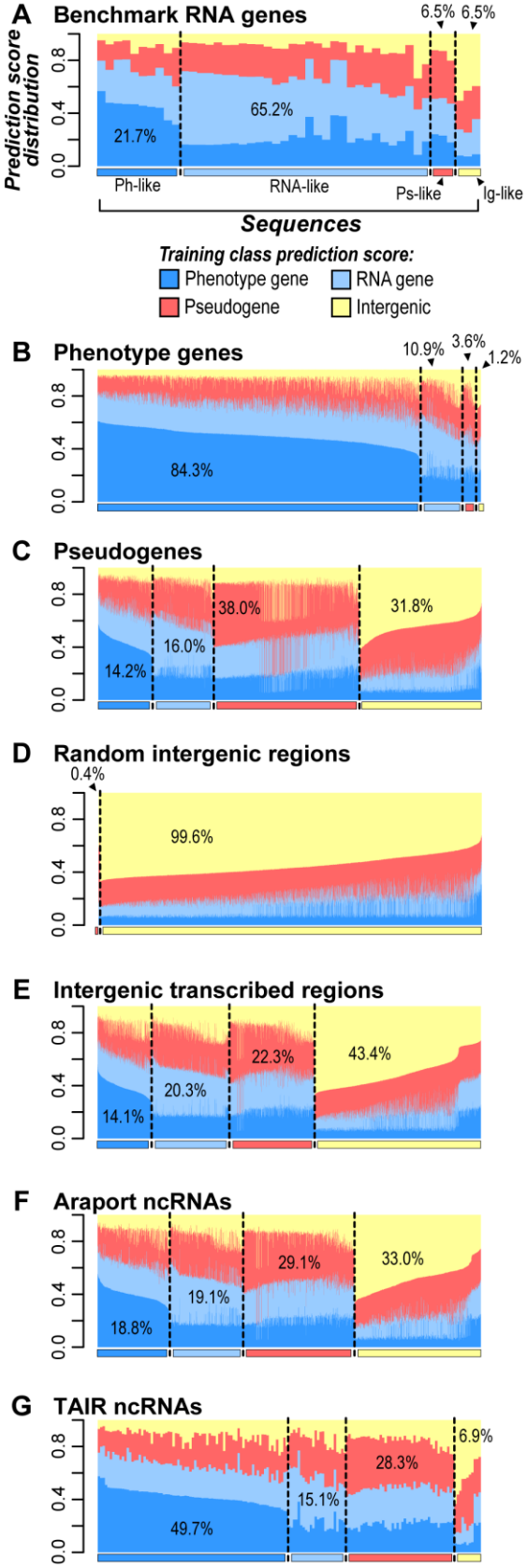
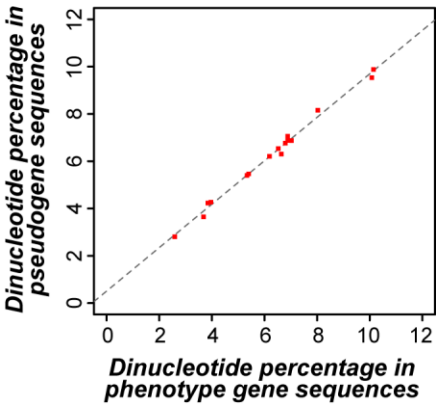

Feature category used
Sequence class
Full
TX
CV
HM
ME
TF
CA
ST
A
Phenotype gene: high FL
Phenotype gene: low FL
Pseudogene: high FL
Pseudogene: low FL
B
Phenotype gene
Pseudogene
Random intergenic
ITR
Araport ncRNA
TAIR ncRNA
Single category: Highly informative
Single category: Less informative
0
20
40
60
80
100
% predicted functional

Figure 6.

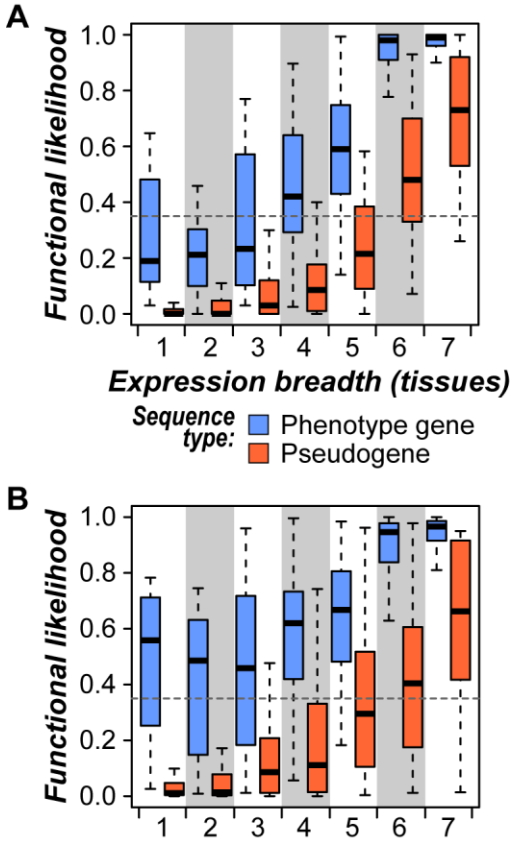

A Benchmark RNA genes
6.5% 6.5%
Prediction score distribution
0.0 0.4 0.8
21.7%
65.2%
Ph-like
RNA-like
Ps-like
Ig-like
Sequences
Training class prediction score:
Phenotype gene
RNA gene
Pseudogene
Intergenic
B Phenotype genes
10.9% 3.6% 1.2%
84.3%
C Pseudogenes
14.2%
16.0%
38.0%
31.8%
D Random intergenic regions
0.4%
99.6%
E Intergenic transcribed regions
14.1%
20.3%
22.3%
43.4%
F Araport ncRNAs
18.8%
19.1%
29.1%
33.0%
G TAIR ncRNAs
49.7%
15.1%
28.3%
6.9%

SUPPLEMENTAL FIGURES

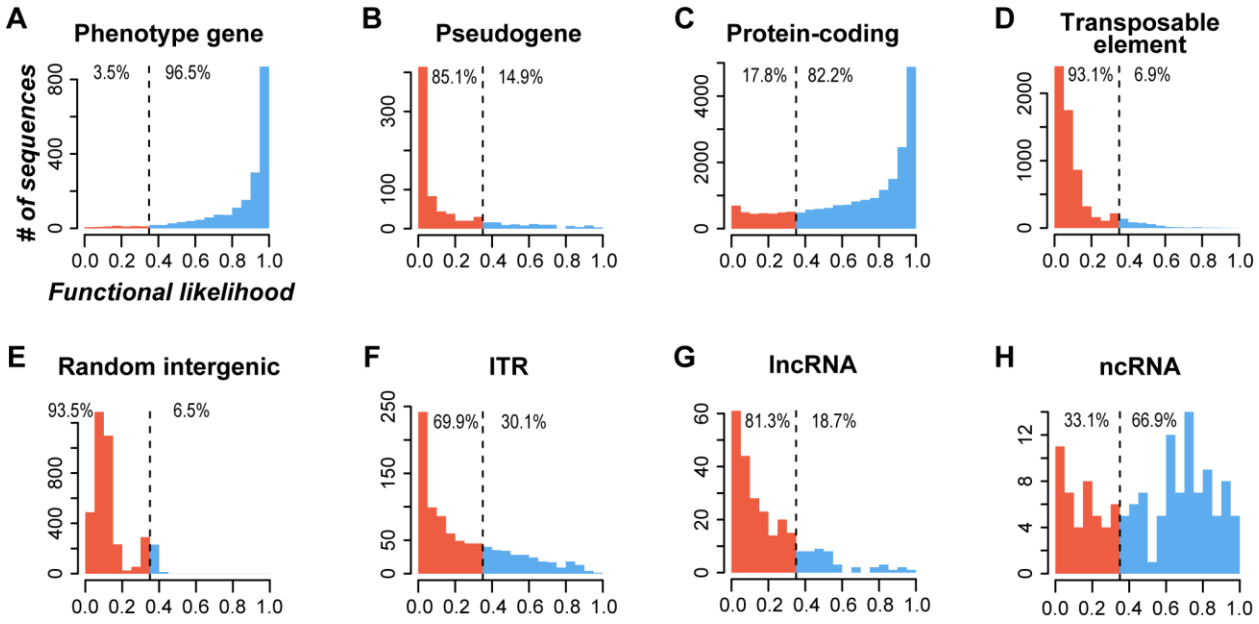
Supplemental Figure 1.



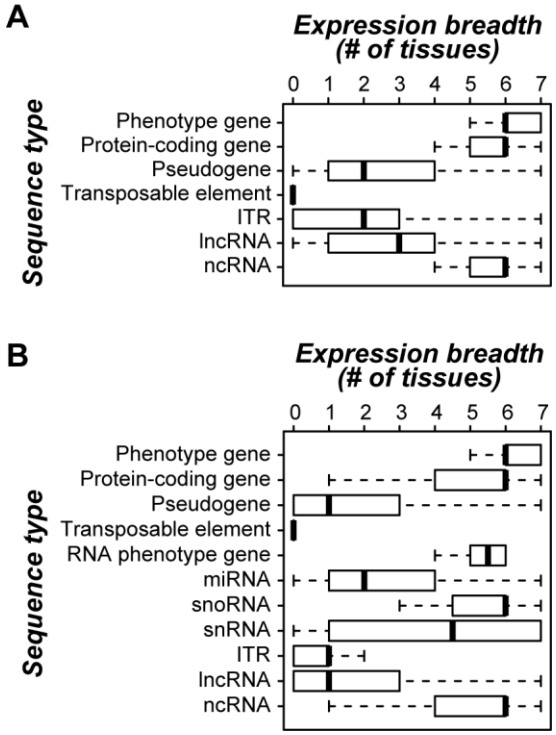
Supplemental Figure 2.



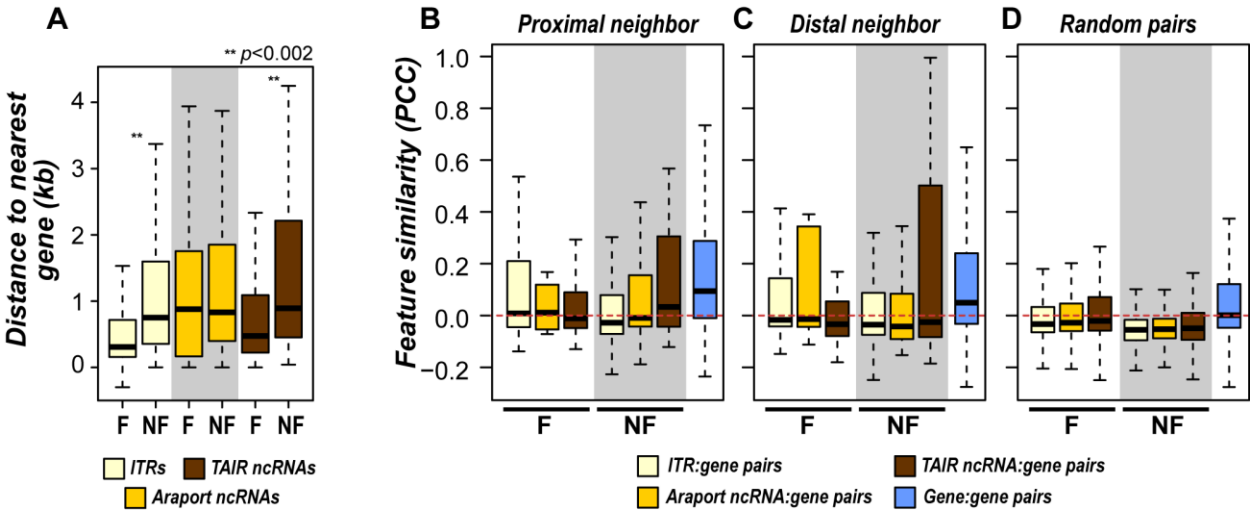
Supplemental Figure 3.



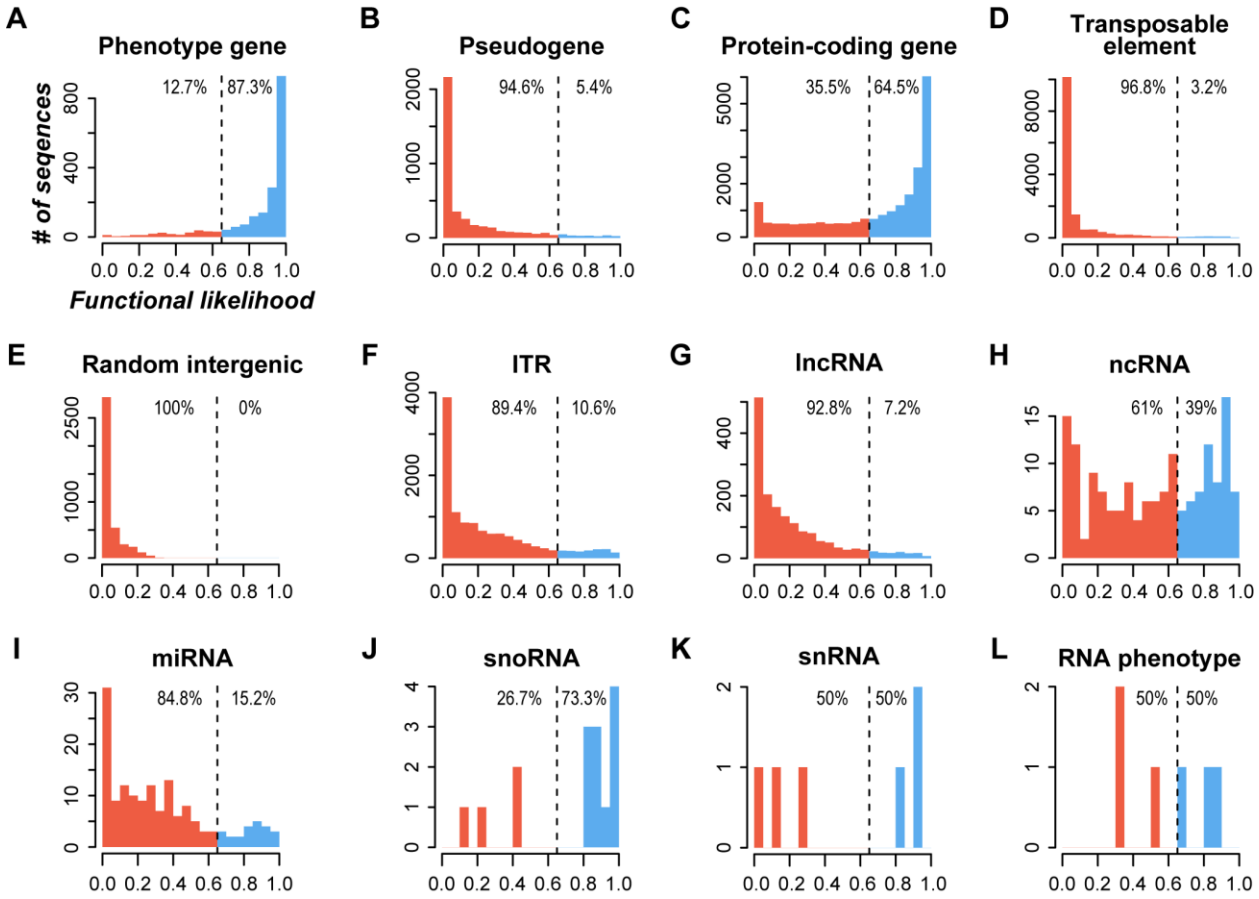
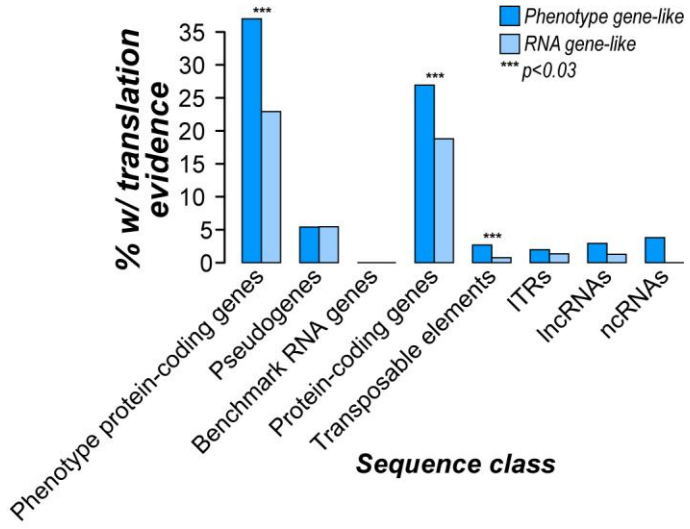
Supplemental Figure 4.



Supplemental Figure 5.



Supplemental Figure 6.



Supplemental Figure 7.