DEFINING THE CHARACTERISTICS AND ROLES OF FUNCTIONAL GENOMIC SEQUENCES USING COMPUTATIONAL APPROACHES

By

John P. Lloyd

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology – Doctor of Philosophy

2017

ABSTRACT

DEFINING THE CHARACTERISTICS AND ROLES OF FUNCTIONAL GENOMIC SEQUENCES USING COMPUTATIONAL APPROACHES

By

John P. Lloyd

Advances in biotechnology have provided a wealth of sequencing data that is transforming our view of a genome. Eukaryotic genomes, initially thought to contain discrete genes in a sea of non-functional DNA, have been found to exhibit pervasive biochemical activity, particularly transcription. However, whether this biochemical activity is functional (i.e. under evolutionary selection) or the result of noisy activity of cellular machinery represents a fundamental debate of the post-genome era. The research described in this dissertation focuses on two open questions confronting genome biology: 1) Where are the functional elements within a genome? 2) What roles are functional elements performing?

For the first question, I focused on transcribed regions in unannotated, intergenic regions of genomes, which represent functionally ambiguous sequences. To determine which and how many intergenic transcribed regions (ITRs) represent functional sequences, machine learning-based function prediction models were established using *Arabidopsis thaliana* as a model. The prediction models were able to successfully distinguish between benchmark functional (phenotype genes) and non-functional sequences (pseudogenes) using evolutionary, biochemical, and sequence-based structural features. When applied to ITRs, ~40% of ITRs were predicted as functional, suggesting ITRs primarily represent transcriptional noise. I further investigated the evolutionary histories of ITRs in four grass (Poaceae) species. ITRs were found to be primarily species-specific and exhibit recent duplicates, with rare examples of ancient duplicate retention. In addition, ITR duplicates and orthologs were usually not expressed. Function prediction

models were also generated in *Oryza sativa* (rice) that predicted ~60% of rice ITRs as non-functional. The results of function prediction models and evaluating evolutionary histories both suggest ITRs are primarily non-functional sequences. However, I also provide a list of potentially-functional ITRs that should be considered high priority targets for future experimental studies.

For the second question, I established a machine learning framework to predict mutant phenotypes, which provide potent evidence for the role of a gene. Phenotype predictions were focused on essential genes (those with lethal mutant phenotypes) in *A. thaliana*, as these genes represent a historically well-studied group. Combining 57 expression, duplication, evolutionary, and gene network characteristics through machine learning methods accurately distinguished between genes with lethal and non-lethal mutant phenotypes. Additionally, essential gene prediction models could be applied across species; essential gene prediction models generated in *A. thaliana* could identify essential genes in rice and *Saccharomyces cerevisiae*. Thus, machine-learning represents a promising avenue of prioritization of candidate genes for large-scale phenotyping efforts. Overall, the research described in this dissertation highlight computational approaches as highly effective in defining functional sequences and classifying the likely roles of genes.

To Ani, for her unwavering love and support.

ACKNOWLEDGEMENTS

As I look back upon my five years in East Lansing, I reflect upon the many people I have to thank during this wonderful journey. First and foremost, I must thank my advisor, Shinhan, who convinced me to join his lab after just a short interview and has since proven time and again to be a thoughtful scientist and excellent mentor. Shinhan has always pushed us to generate top-notch science and not settle to churn out low-impact studies. I believe that I and all members of his lab are stronger scientists for it. I also wanted to mention my Master's advisor, David Meinke, for putting me on the path that I am now. My life would have ended up very different if he had not brought me into his lab as an undergraduate and Master's student. I also want to thank my committee members, Robin Buell, Robert Last, Doug Schemske, and George Mias, for donating their scarce time and helpful advice over the years.

In addition to being a great mentor, Shinhan always managed to bring together a great group of people with which to work and relax. To the many friends and colleagues from the Shiu lab over the years, it has been a pleasure and I will always fondly recall our time together. In particular, I would like to mention Gaurav Moghe, Sahra Uygun, Nicholas Panchy, Beth and Josh Moore, and Christina Azodi, who were fixtures in the lab during my stay there and come to mind first when I think of my Ph.D. studies. Last and certainly not least, I must thank my lovely wife, Anita. She was always there with support, advice, and help in times of need. Thank you, Ani, you have helped me grow and become a better person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AraNet	Probabilistic functional gene network of *Arabidopsis thaliana*
Araport	Arabidopsis Information Portal
AUC-ROC	Area under the curve - receiver operating characteristic
BLAST	Basic local alignment search tool
bp	Base pairs
BS-seq	Bisulfite sequencing
CAGE	Cap analysis gene expression
ChIP-seq	Chromatin immunoprecipitation sequencing
CNB	Conserved nucleotide block
DAP	Days after pollination
DHS	Deoxyribonuclease I hypersensitive site
DiProDB	Dinucleotide Property Database
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease I
EMBOSS	European Molecular Biology Open Software Suite
ENCODE	Encyclopedia of DNA Elements
eRNA	Enhancer RNA
EST	Expressed sequence tag
ETR	Exon transcribed region
FDR	False discovery rate
FET	Fisher's Exact Test

FL	Functional likelihood
FNR	False negative rate
FPKM	Fragments per kilobase of transcript per million mapped reads
FPR	False positive rate
FungiDB	Fungal Database
GBM	Gene body methylation
GO	Gene Ontology
ITR	Intergenic transcribed region
K	Rate of substitutions
Ka	Rate of non-synonymous substitutions
Ks	Rate of synonymous substitutions
KST	Kolmogorov–Smirnov test
lncRNA	long non-coding RNA
MCScanX	Multiple collinearity scan software, version X
miRBase	microRNA Database
miRNA	microRNA
miRNA-TR	microRNA transcribed region
MNase-seq	Micrococcal nuclease sequencing
MSU	Michigan State University
MUSCLE	Multiple Sequence Comparison by Log-Expectation
MYA	Million years ago
MYO	Million years old
NCBI	National Center for Biotechnology Information

ncRNA	non-coding ribonucleic acid
OrthoMCL	Ortholog Markov clustering software
PAML	Phylogenetic Analysis by Maximum Likelihood
PC	Principal component
PCC	Pearson's correlation coefficient
PHAST	Phylogenetic Analysis with Space/Time models
Phen-TR	Phenotype exon transcribed region
PHYLIP	Phylogeny Inference Package
Pseu-TR	Pseudogene transcribed region
r^2	Square of Pearson's correlation coefficient
RAxML	Randomized Axelerated Maximum Likelihood
RF	Random forest
RNA	Ribonucleic acid
RNA Pol II	RNA Polymerase II
RNA-seq	RNA sequencing
RPKM	Reads per kilobase of transcript per million mapped reads
SMO	Sequential minimal optimization
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
SRA	Sequence read archive
SVM	Support vector machines
TAIR	The Arabidopsis Information Resource

TE	Transposable element
TF	Transcription factor
tRNA	Transfer RNA
tRNAdb	tRNA Database
U test	Mann Whitney U test
WEKA	Waikato Environment for Knowledge Analysis
WGD	Whole genome duplication

CHAPTER 1: INTRODUCTION

Prevalence of intergenic transcription

Advances in sequencing technology have led to the discovery of extensive transcriptional activities occurring in unannotated, intergenic regions of genomes. Intergenic transcription is prevalent in a variety of model systems, including *Homo sapiens* (human) (ENCODE Project Consortium, 2012), *Drosophila melanogaster* (fruit fly) (Brown et al., 2014), *Caenorhabditis elegans* (nematode) (Boeck et al., 2016), and *Saccharomyces cerevisiae* (yeast) (Nagalakshmi et al., 2008). In plant systems, intergenic transcripts have been identified in *Arabidopsis thaliana* (Yamada et al., 2003; Stolc et al., 2005; Moghe et al., 2013; Krishnakumar et al., 2015) and *Oryza sativa* (rice) (Nobuta et al., 2007). Initially, intergenic transcripts were suggested to be primarily related to nearby genes as unannotated exon extensions or run-on transcription (van Bakel et al., 2010). However, a variety of novel functions have been suggested for these sequences (Guil and Esteller, 2012; Hanada et al., 2013; Tan et al., 2015), suggesting they may also represent the activity of undetected genes.

The notion that ITRs represent independent functional elements with novel functions awaiting discovery should be tempered by the possibility that they may also represent the product of noisy activity of the cellular machinery controlling gene expression. Transcriptional noise can be produced by random landing of RNA Polymerase II or presence of spurious regulatory signals driving expression of non-functional transcripts (Struhl, 2007). As much as 90% of the transcriptional activity in the yeast genome has been estimated to be the product of noisy transcription, including many transcripts originating from intergenic regions (Struhl, 2007). In addition to transcriptional noise, ITRs may also represent contamination from genomic DNA generated while preparing RNA samples for sequencing (Moghe et al., 2013). Few ITRs have been experimentally characterized (Ivanova et al., 2006; Guttman et al., 2009; Heinen et al.,

2009; Ulitsky et al., 2011; Sauvageau et al., 2013; Lai et al., 2015) or exhibit sequence conservation within or between species (Moghe et al., 2013). Thus, the possibility that ITRs primarily represent non-functional sequences cannot be ruled out. To explore which ITRs likely result from the activities of novel genes (i.e. functional sequences), a clear conception of what defines a gene is required.

The nature of genes

Genes have been described historically as the unit of heredity (Gerstein et al., 2007) and underlie the biological diversity and evolution of life on earth (Raff, 1996). The conception and definition of a gene have evolved over the past century. The term "gene" was coined in 1909 and initially utilized as an abstract concept to explain the heredity of phenotypes between parents and offspring (Gerstein et al., 2007; Portin, 2015; Portin and Wilkins, 2017). Research over the following 50 years identified the physical characteristics of genes: they are regions of chromosomes, composed of DNA, and typically represent blueprints for proteins (Gerstein et al., 2007; Portin, 2015; Portin and Wilkins, 2017). The concept of a gene as a distinct chromosomal sequence that encodes a single functional protein product emerged (Beadle and Tatum, 1941; Crick, 1963; Gaertner and Cole, 1977; Portin and Wilkins, 2017).

The post-genome era has complicated the straightforward view of a gene (Pesole, 2008; ENCODE Project Consortium, 2012). First, genes are not necessarily protein-coding, as extensive roles for non-coding RNAs have been uncovered (Fire et al., 1998; Guil and Esteller, 2012; Tan et al., 2015). Next, alternative splice forms (Kelemen et al., 2013) and translational start sites (Kochetov, 2008; Bazykin and Kochetov, 2011) are common, indicating that a single gene may encode multiple protein isoforms. A related issue is that the presence of alternative transcriptional start and stop sites suggest boundaries of a gene may not be well defined (Pesole,

2008). Genes may also overlap one another, either on opposing DNA strands or in distinct coding frames (Veeramachaneni et al., 2004; Makalowska et al., 2005), and therefore genes do not necessarily represent unique regions on a chromosome. Last, genes have previously been defined while including their upstream regulatory promoters (Pearson, 2006). However, *trans*-acting gene regulatory elements, such as enhancers (Bulger and Groudine, 2010), exist remotely from gene bodies. Together, these issues highlight genes as DNA sequences that frequently encode multiple protein isoforms, may not have clearly-definable start and stop points, and can be associated with an extensive array of both *cis*- and *trans*-acting of regulatory elements.

The complicated picture of gene structure and function represents a paradigm shift in which the gene as a distinct protein-coding element now represents only a subset of all possible cases (Portin, 2015), and as a result, there have been repeated calls to reconsider the definition of a gene (Pearson, 2006; Gerstein et al., 2007; Pesole, 2008; Portin and Wilkins, 2017). However, it is not clear to what extent the potentially complex structures of genes undermine the traditional concept of a gene as the unit of heredity that underlies phenotype. Instead, ever-advancing understanding of the physical structure of genes has perhaps overemphasized the use of the "nominal" gene definition (Griffiths and Stotz, 2006), which focuses on predicted gene sequences without regard to potential phenotype association. As evidence for this, there have been multiple tentative updates to the definition of a gene outlined in the past decade (Gerstein et al., 2007; Pesole, 2008; Portin and Wilkins, 2017) and each includes a requirement that genes produce a "functional" product or associate with a phenotype (Gerstein et al., 2007; Pesole, 2008; Portin and Wilkins, 2017). Thus, there appears to be two questions related to gene definitions that should be considered separately: How do we define the coordinates of gene models? How do we identify functional genome sequences? Furthermore, biological noise may

play a critical role, as alternative splice variants or transcriptional start and end sites may result from non-specific activity of the cellular mechanisms regulating these processes. Thus, it is crucial to not only catalog all observed transcript isoforms, but to also direct attention to determining whether an isoform is functional.

Definition of function

Despite increasing complexity with which we view genes, these sequences are agreed to produce functional products. Thus, a clear definition of what constitutes function within a biological context is needed. Two contrasting definitions of function are frequently debated, the "causal role" and "selected effect" definitions (Cummins, 1975; Amundson and Lauder, 1994; ENCODE Project Consortium, 2012; Graur et al., 2013; Doolittle et al., 2014). Under the causal role definition, a functional genome region is one that is exhibits reproducible biochemical activity, such as transcription, protein-binding, or the presence of particular chromatin states. This definition was invoked by the ENCODE Consortium (2012) to conclude that 80% of the human genome was biochemically functional, which was in turn cited as evidence to disprove the existence of "junk" DNA (Eddy 2013). However, this estimate of the functional proportion of the human genome far exceeds those based on evolutionary conservation (Rands et al., 2014) or mutational load (Graur, 2017). Because of this, the suitability of the causal role definition has come under extensive critique (Doolittle, 2013; Graur et al., 2013; Niu and Jiang, 2013). An alternative definition for function is the selected effect, which requires that functional genome sequences contribute to the survival and reproduction of an organism and thus be under evolutionary selection. As biochemical activity may be the result of noise, the selected effect definition has been suggested to be more suitable for defining biological function (Amundson

and Lauder, 1994; Graur et al., 2013; Doolittle et al., 2014). Given these considerations, I will utilize the selected effect definition of function throughout this dissertation.

Based on the selected effect definition, what data should be considered as evidence that a sequence is functional? Observing a phenotype resulting from mutation of a sequence is considered the gold standard for identifying functional genome regions (Ponting and Belgard, 2010; Niu and Jiang, 2013). In such cases, the presence of a mutant phenotype indicates that a sequence had a role in producing the wild-type phenotype and is likely under selection. However, phenotype data for each region in a genome is not available and has been used to validate the functionality of only a handful of ITRs. Sequence conservation, particularly over long time periods, also represents strong functional evidence and can be assessed genome-wide. However, selective pressure may be weak or positive, resulting in a lack of detectable conservation, particularly among RNA genes (Pang et al., 2006; Ponting, 2017), or a sequence may perform a species-specific function. By contrast, sequence conservation over short periods may result from a lack of time for sequences to significantly diverge. Thus, the presence or absence of conservation cannot be used by itself to classify functional sequences. Last, biochemical activity also represents evidence of functionality, including biochemical signatures such as transcription, transcription factor binding, and the presence of epigenetic marks (e.g. specific DNA methylation and histone mark patterns). However, biochemical activities are subject to noise and, similar to conservation evidence, any individual activity should not be used to define functionality alone.

It has been suggested that an integrative approach that considers phenotype, evolutionary, and biochemical evidence in combination could be used to define functional sequences (Kellis et al., 2014). Such a framework has been criticized for ignoring the evolutionary origin of a

biochemical activity (i.e. the selected effect definition). Nevertheless, recent studies have shown that considering genetic, evolutionary, and biochemical properties in combination effectively distinguishes between sequences that are under selection and those that are not (Tsai et al., 2017). This approach is explored in more detail in this dissertation.

Gain and loss of functional sequences

Functional sequences in genomes are not static and each species harbors a distinct set of genes and regulatory sequences. What mechanisms underlie the changes in functional sequences between genomes? Gene duplication plays a critical role in the evolution of new functional sequences (Ohno et al., 1968; Zhang, 2003; Panchy et al., 2016). Following gene duplication, one gene copy may retain the ancestral function while the additional copy evolves a new function (Hughes, 1994; Zhang et al., 1998). Alternatively, two gene copies may lose complementary subsets of the ancestral gene to become more specialized, a process known as subfunctionalization (Force et al., 1999; Lynch and Force, 2000). Although subfunctionalization can result from a stochastic loss of gene functions, it may also play an important role in optimization of genes by separating conflicting functions into separate gene copies that can evolve independently (referred to as escape from adaptive conflict) (Hittinger and Carroll, 2007; Marais and Rausher, 2008).

While duplication events represent a source of functional diversity, the most common outcome of gene duplication is the pseudogenization of one copy (Li et al., 1981; Maere et al., 2005; Hanada et al., 2008; Moghe et al., 2014). Thus, while duplication can facilitate the evolution of new functions, it more frequently results in the production of non-functional sequences. The process of pseudogenization of duplicate gene copies can take place over millions of years (Lynch and Conery, 2000; Lynch and Conery, 2003). As a result, a subset of

annotated genes may represent sequences undergoing functional decay and *en route* to pseudogene status. For example, a pair of duplicate transcription factors in *A. thaliana*, *DDF1* and *DDF2*, exhibit evidence of highly asymmetric divergence with *DDF1* retaining ancestral functions and *DDF2* losing binding site affinity (Lehti-Shiu et al., 2015). This suggests that *DDF2* could be undergoing functional decay. In a genome-wide survey, 1,939 *A. thaliana* protein-coding genes lack transcriptional evidence and exhibit characteristics that are more consistent with pseudogenes than protein-coding genes (Yang et al., 2011). This represents 7% of the annotated gene space in *A. thaliana*, indicating that functional decay among putative gene annotations may not be uncommon. Further complicating the relationship between pseudogenes and functional decay is the fact that pseudogenes may remain functional as truncated proteins or evolve novel functions at the RNA level post-pseudogenization (Poliseno et al., 2010; Karreth et al., 2015).

In addition to gene duplication, novel genes can evolve *de novo* from genome regions that did not previously contain a functional element (Kaessmann, 2010; Tautz and Domazet-Lošo, 2011). While the probability of a protein-coding gene evolving from non-coding sequence was described as "practically zero" (Jacob, 1977), multiple studies have identified *de novo* gene birth in primates (Johnson et al., 2001; Knowles et al., 2009; Toll-Riera et al., 2009), fruit fly (Levine et al., 2006; Zhou et al., 2008) and yeast (Cai et al., 2008). Intergenic transcription may play a critical role in *de novo* gene evolution, with transient "proto"-genes evolving from intergenic transcripts that may come under positive selection and evolve into protein-coding genes (Carvunis et al., 2012). This could suggest that tolerating the transcription of non-functional intergenic sequences plays a role in the evolution novel functional sequences.

Identifying the role of a gene

What role is a gene performing in the cell or for the organism? Evidence can be provided by expression patterns (Eisen et al., 1998; Spellman et al., 1998; Lee et al., 2004a; Uygun et al., 2016), metabolic profiling (Raamsdonk et al., 2001; Sumner et al., 2003; Hirai et al., 2007), biochemical analyses (Martzen et al., 1999; Chen et al., 2003), and gene networks and interactions (Bork et al., 2004; Lee et al., 2004b; Lee et al., 2010; Arabidopsis Interactome Mapping Consortium, 2011). Beyond these tools, assessing the biological consequences resulting from disruption of a gene (i.e. mutant phenotypes) provides potent clues to the function of a gene. Owing to this, comprehensive datasets of mutant phenotype data for each annotated gene in a genome are available for multiple model systems (Winzeler et al., 1999; Kamath et al., 2003; Boutros et al., 2004; Kim et al., 2010).

Plant systems, however, lag behind in the availability of well-curated phenotype data. Only 13% of annotated genes in the model plant *A. thaliana* are associated with a mutant phenotype (Kuromori et al., 2006; Lloyd and Meinke, 2012; Savage et al., 2013), despite the presence of near-saturation mutagenesis resources (Kuromori et al., 2009). This is due in part to long generation times among plant species. For example, *A. thaliana*, a relatively fast-growing plant, has a generation time of 6 weeks (Meyerowitz, 1989), longer than those of other eukaryotic models such as yeast, fruit fly, or nematode. In addition, pervasive duplication of plant genes has resulted in extensive functional overlap between paralogs that masks the consequences of gene disruption, making the identification of phenotypes-of-interest difficult. Computational approaches may play a critical role in streamlining costly and time-consuming experimental analysis by providing candidate predictions of gene-gene interactions (You et al.,

2010; Upstill-Goddard et al., 2013), genetic redundancy (Chen et al., 2010), and phenotype genes (Seringhaus et al., 2006; Yuan et al., 2012; Musso et al., 2014).

Machine learning as a tool in biology

Advancements in biotechnology have brought with them massive quantities of biological sequencing data. In response, biologists have begun to adopt computer science techniques to manipulate and mine these data for valuable biological insights. One notable data mining technique is machine learning (Tarca et al., 2007): computational algorithms designed to recognize patterns in data and make predictions. Briefly, two examples of machine learning algorithms are decision trees (e.g. Random Forests), which are constructed as flow chart-like structures based on a set of training data, and support vector machines (SVM), which use multi-dimensional planes to separate sets of labeled training data. Studies based on machine learning have made major impacts in a wide array of fields in biology. Early machine learning analyses were successful in identifying translation initiation sites in *Escherichia coli* (Stormo and Schneider, 1982) and predicting secondary structure of proteins (King and Sternberg, 1990). Machine learning methods have been widely applied in biomedical cancer research, where they effectively discriminate between healthy and diseased tissue (Furey et al., 2000; Guyon et al., 2002; Hilario et al., 2003; Wang et al., 2005) and classify tumor severity (Shipp et al., 2002; Ye et al., 2003) on the basis of expression and proteomic data. In more recent basic research, machine learning methods have been used to predict mutant phenotypes (Yuan et al., 2012; Zhang et al., 2013), generate gene functional networks (Bassel et al., 2011), classify stress-responsive gene expression (Zou et al., 2011), assess functional overlap between gene pairs (Chen et al., 2010), and characterize the features of gene duplicates (Moghe et al., 2014). Throughout this dissertation I make extensive use of machine learning techniques, specifically to

generate prediction models capable of classifying the functionality of a sequence and the essentiality of genes.

Dissertation outline and significance

Genes produce the functional protein and RNA products required by a cell, and thus represent a primary target for basic research and biological engineering. Given their fundamental importance, the identification of the number, location, and role of all genes within a genome represent critical and fundamental tasks in genome research. Toward the identification of genic (i.e. functional) sequences, Chapters 2 and 3 outline three key advancements: 1) systematic identification of the biochemical, evolutionary, and structural characteristics shared among functional genome regions, 2) establishment of a machine learning framework that accurately distinguishes functional and non-functional sequences, and 3) detailed evaluation of the evolutionary dynamics of ITRs, which represent functionally ambiguous genome regions. Overall, these chapters describe how functional genome sequences are distinct from non-functional ones and outline how these differences can be used to distinguish functional and noisy biochemical activity.

Once functional sequences have been identified, the next question is what role they are performing in a cell. Toward this objective, Chapter 4 describes a machine learning framework that effectively distinguishes between genes with lethal and non-lethal mutant phenotypes. This resulted in a catalog of characteristics shared among essential genes, a gene set has been the target of research as they highlight the minimal gene set and biological processes that are required for life. Moreover, successful prediction of genes with lethal phenotypes provides a proof-of-concept to show that machine learning approaches represent a promising system for prioritizing genes for large-scale phenotyping analysis in plants. Importantly, predictions of both

functional genome regions and essential genes show translational potential, indicating that the ‘omics data available for model systems can be successfully leveraged to provide insight into potential gene function in non-model systems.

CHAPTER 2: DEFINING THE FUNCTIONAL SIGNIFICANCE OF INTERGENIC TRANSCRIBED REGIONS [1]

[1] The work described in this chapter has been submitted for publication:

John P. Lloyd, Zing Tsung-Yeh Tsai, Rosalie P. Sowers, Nicholas L. Panchy, Shin-Han Shiu (2017) Defining the functional significance of intergenic transcribed regions. *Submitted.*

ABSTRACT

With advances in transcript profiling, the presence of transcriptional activities in intergenic regions has been well established. However, whether intergenic expression reflects transcriptional noise or activity of novel genes remains unclear. We identified intergenic transcribed regions (ITRs) in 15 diverse flowering plant species and found that the amount of intergenic expression correlates with genome size, a pattern that could be expected if intergenic expression is largely non-functional. To further assess the functionality of ITRs, we first built machine learning classifiers using *Arabidopsis thaliana* as a model that accurately distinguish functional sequences (phenotype genes) and non-functional ones (pseudogenes and unexpressed intergenic regions) by integrating 93 biochemical, evolutionary, and sequence-structure features. Next, by applying the models genome-wide, we found that 4,427 ITRs (38%) and 796 annotated ncRNAs (44%) had features significantly similar to benchmark protein-coding or RNA genes and thus were likely parts of functional genes. Approximately 60% of ITRs and ncRNAs were more similar to non-functional sequences and should be considered transcriptional noise. The predictive framework established 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. 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. Thus the identification of 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 study represents the gold standard by which the functional significance of genomic regions, including ITRs, can be confirmed (Ponting and Belgard, 2010; Niu and Jiang, 2013). In *Mus musculus* (mouse), at least 25 ITRs with loss-of-function mutant phenotypes have been identified (Sauvageau et al., 2013; Lai et al., 2015), 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 rerio* (Ulitsky et al., 2011). In human, 162 long intergenic non-coding

RNAs 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 but not limited to *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 may 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 considered functional cannot depend solely on the fact that it is expressed. In addition to being biochemically active, the genomic region must be under selection. This line of logic has revived the classical ideas on differentiating "causal role" and "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 (ENCODE Project Consortium, 2012) to classify ~80% of the human genome as having biochemical functions. This finding has been used as evidence to disprove the presence of junk

DNA that is not under natural selection (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, selected effect functionality is advocated to be a more suitable definition for a genomic region with discernible activity (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.

Functional ITRs represent genic sequences that have not been identified with conventional gene finding programs. Such 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 remain unidentified. Due to the debate on the definitions of function post-ENCODE, Kellis et al. (2014) suggested that evolutionary, biochemical, and genetic evidences provide complementary information to define functional genomic regions. Integration of chromatin accessibility, transcriptome, and conservation evidence was 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 or if the features that define functional genomic regions in human are applicable in other species. In plants, even though 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 functionality of intergenic transcription, we first identified ITRs in 15 flowering plant species with 17-fold genome size differences and evaluated the relationship

between the prevalence of intergenic expression and genome size. Next, we determined whether 93 evolutionary, biochemical, and sequence-structure features could distinguish functional sequences (phenotype genes) and non-functional ones (pseudogenes) using *A. thaliana* as a model. We then jointly considered all 93 features to establish functional gene prediction models using machine learning methods. Given that phenotype genes were composed of protein-coding sequences, prediction models were also generated by considering benchmark RNA genes to ensure that functional predictions were not exclusive to protein-coding genome regions. Finally, we applied the models to ITRs and annotated ncRNAs to determine whether these functionally ambiguous sequences are more similar to known functional or likely non-functional sequences.

RESULTS AND DISCUSSION

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

Transcription of an unannotated, intergenic region could be due to non-functional transcriptional noise or the activity of a novel gene. If noisy transcription occurs due to random landing of RNA Pol II or spurious regulatory signals, a naïve expectation is that, as genome size increases, the coverage of intergenic expression would increase accordingly. By contrast, we expect that the extent of expression for genic sequences will not be significantly correlated with genome size because larger plant genomes do not necessarily have more genes (r^2=0.01; p=0.56; see Methods). 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 coverage of intergenic expression occurring within 15 flowering plant species.

We first identified genic and intergenic transcribed regions using leaf transcriptome data from 15 flowering plants with 17-fold differences in genome size (Supplemental Table 2.1). 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 gene annotation and had no significant translated sequence similarity to plant protein sequences. As expected, the coverage of expression originating from annotated genic regions had no significant correlation with genomes size (r^2=0.03; p=0.53; **Fig. 2.1A**). In contrast, the coverage of intergenic expression occurring was significantly and positively correlated (r^2=0.30; p=0.04; **Fig. 2.1B**). 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 (**Fig. 2.1B**), 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 allow benchmark functional and non-functional genomic regions to be distinguished in the model plant, *Arabidopsis thaliana*. 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). Because their mutations have significant growth and/or developmental impact and likely contribute to reduced fitness, these phenotype genes can be considered functional under the selected effect definition (Neander, 1991). For benchmark 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).

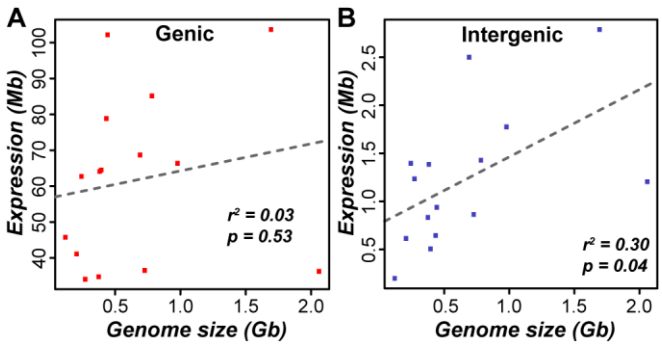


Figure 2.1 Relationship between genome size and transcription coverage. Transcription coverage is shown for (*A*) annotated genic regions and (*B*) intergenic regions excluding any annotated features. Each dot represents one of 15 flowering plant species. Mb: megabase. Gb: gigabase. Dotted lines: linear model fits. r^2: square of Pearson's correlation coefficient.

We evaluated 93 gene or gene product features for their ability to distinguish between phenotype genes and pseudogenes. These features were grouped into seven categories, including chromatin accessibility, DNA methylation, histone 3 (H3) marks, sequence conservation, sequence-structure, transcription factor (TF) binding, and transcription activity. Feature values (Supplemental Table 2.2) 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. 2.2A**). Despite the strong performance of transcription activity-related features, the presence of expression (i.e. presence of transcript) evidence was a poor predictor of functionality (AUC-ROC=0.58; **Fig. 2.2A**). This is because 80% of pseudogenes were considered expressed in ≥1 of 51 RNA-seq datasets, demonstrating that presence of transcripts should not be used by itself as evidence of functionality. Sequence conservation, DNA methylation, TF binding, and H3 mark features were also fairly distinct between phenotype genes and pseudogenes (median AUC-ROC ~0.7 for each category; **Fig. 2.2B-E**). We also observed high performance variability within feature categories (see Supplemental Information). By contrast, chromatin accessibility and sequence-structure features were largely uninformative (median AUC-ROC=0.51 and 0.55, respectively; **Fig. 2.2F,G**). The poor performance of chromatin accessibility features is likely because the DNase I hypersensitive site (DHS) datasets were sparse, as only 2-6% of phenotype gene and pseudogene sequences overlapped a DHS

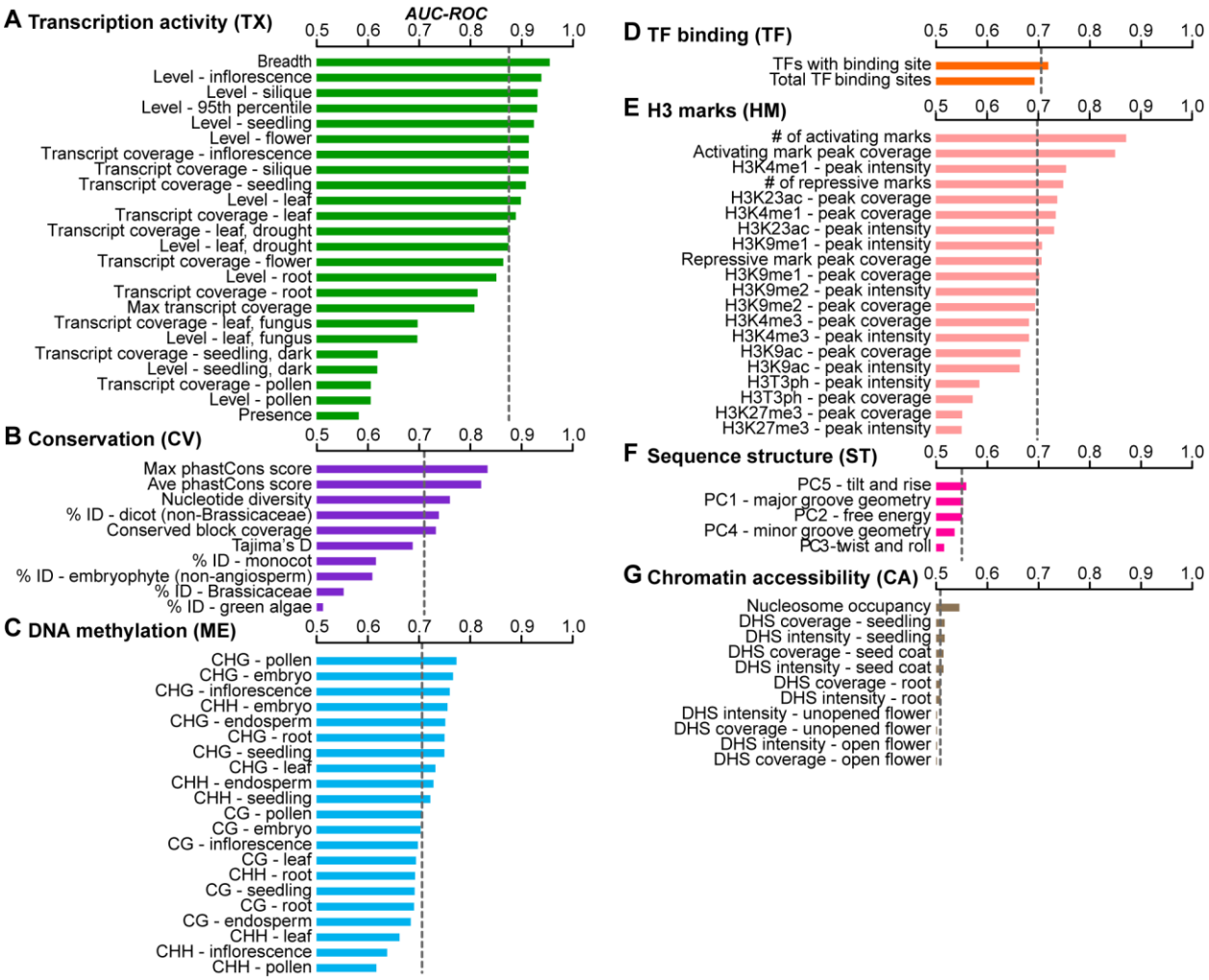


Figure 2.2 Predictions of functional and non-functional sequences based on single features. Prediction performance is measured using Area Under the Curve - Receiver Operating Characteristic (AUC-ROC). Features include those in the categories of (*A*) transcription activity, (*B*) sequence conservation, (*C*) DNA methylation, (*D*) transcription factor (TF) binding, (*E*) histone 3 (H3) marks, (*F*) sequence structure, and (*G*) chromatin accessibility. AUC-ROC ranges in value from 0.5 (equivalent to random guessing) to 1 (perfect predictions). Dotted lines: median AUC-ROC of features in a category.

peak. Further, median nucleosome occupancy of phenotype genes (median normalized nucleosome occupancy = 1.22) is only slightly lower than that of pseudogenes (median = 1.31; Mann Whitney U test, p < 2e-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 (r^2=0.99, p<3e-16).

The differences between genes and pseudogenes in transcription, conservation, and epigenetic features and functional genomic regions suggested 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 function predictions based on single features. We first considered expression breadth of a sequence, the best predicting single feature of functionality. Despite high AUC-ROC (0.95; **Fig. 2.2A**), the false positive rate (FPR; % of pseudogenes predicted as phenotype genes) was 21% when only expression breadth was used, while the false negative rate (FNR; % of phenotype genes predicted as pseudogenes) was 4%. Similarly, the best-performing H3 mark- and sequence conservation-related features (**Fig. 2.2B,E**) had FPRs of 26% and 32%, respectively, and also incorrectly classified at least 10% of phenotype genes as pseudogenes. Thus, error rates are high even when considering well-performing single features, 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. 2.3A**) and pseudogenes (**Fig. 2.3B**) were 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 for phenotype gene and pseudogenes in combination using random forest (referred to as the full model; see Methods). The phenotype gene and pseudogene sequences and associated conservation, biochemical, and sequence-structure features were separated into distinct training and testing sets and the full model was generated and validated using independent data subsets (cross-validation). The full model provided more accurate predictions (AUC-ROC=0.98; FNR=4%; FPR=10%; **Fig. 2.3C**) than any individual feature (**Fig. 2.2**; Supplemental Table 2.3). An alternative measure of performance based on the precision (proportion of predicted functional sequences that are truly functional) and recall (proportion of truly functional sequences predicted correctly) values among predictions generated by the full model also indicated that the model was performing well (**Fig. 2.3D**). When compared to the best-performing single feature (expression breadth), the full model had a similar FNR but half the FPR (10% compared to 21%). Thus, the full model is highly capable of distinguishing between phenotype genes and pseudogenes.

We next determined the relative contributions of different feature categories in predicting phenotype genes and pseudogenes and 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. 2.2**). 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. 2.3C**). Particularly, the transcription activity feature category model performed

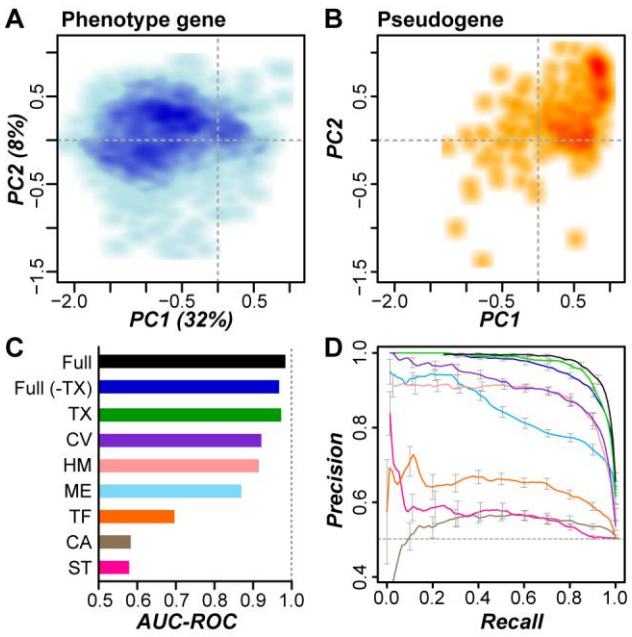


Figure 2.3 Predictions of functional and non-functional sequences based on multiple features. Smoothed scatter plots of the first two principle components (PCs) of (*A*) phenotype gene and (*B*) pseudogene features. The percentages on the axes in *(A)* indicate the feature value variation explained by the associated PC. (*C*) AUC-ROC values of function prediction models built when considering all features (Full), all except transcription activity (TX)-related features (Full (-TX)), and all features from each category. The category abbreviations follow those in **Fig. 2.2**. (*D*) Precision-recall curves of the models with matching colors from (*C*). The models were built using feature values calculated from 500 bp sequence windows.

almost as well as the full model (FNR=6%, FPR=12%). We emphasize that the breadth and level of transcription are the causes of the strong performance of the transcription activity-only model, not the presence of expression evidence. To evaluate whether the strong performance of the full model is being driven solely by transcription activity-related features, we also built a function prediction model did not consider these features (full (-TX), **Fig. 2.3 C,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, but 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. 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.

To provide a measure of the potential functionality of any sequence in the *A. thaliana* genome, including ITRs and ncRNAs, we utilized the confidence score from the full model as a "functional likelihood" value (see Methods) (Tsai et al., 2017). The functional likelihood (FL) score ranges between 0 and 1, with high values indicating that a sequence is more similar to phenotype genes (functional) and low values indicating a sequence more closely resembles pseudogenes (non-functional). FL values for all genomic regions examined in this study are available in Supplemental Table 2.4. As expected, phenotype genes had high FL values (median=0.97; **Fig. 2.4A**) and pseudogenes had low values (median=0.01; **Fig. 2.4B**). To call sequences as functional or not, we defined a threshold FL value (0.35) by maximizing the F-measure (see Methods). Using this threshold, 96% of phenotype genes (**Fig. 2.4A**) and 90% of

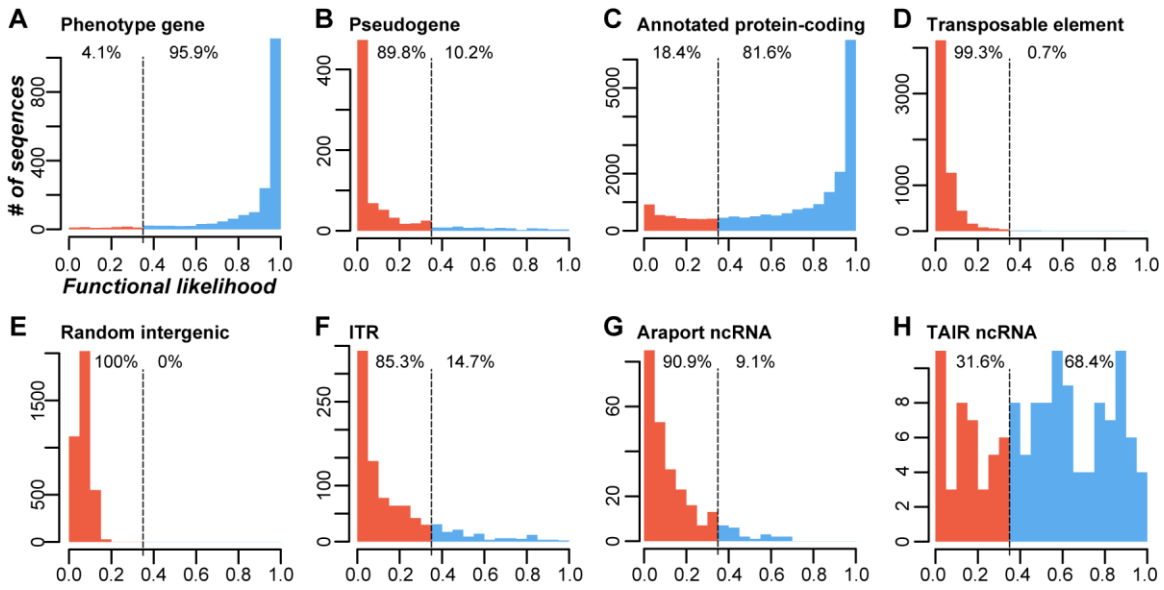


Figure 2.4 Functional likelihood distributions based on the full model. (*A*) Phenotype genes. (*B*) Pseudogenes. (*C*) Annotated protein-coding genes. (*D*) Transposable elements. (*E*) Random unexpressed intergenic sequences. (*F*) Intergenic transcribed regions (ITR). (*G*) Araport11 ncRNAs. (*H*) TAIR10 ncRNAs. The full model was established using 500 bp sequence windows. Higher and lower functional likelihood values indicate greater similarity to phenotype genes and pseudogenes, respectively. Vertical dashed lines indicate the threshold for calling a sequence as functional or non-functional. The percentages to the left and right of the dashed line indicate the percent of sequences predicted as functional or non-functional, respectively.

pseudogenes (**Fig. 2.4B**) 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 (TEs), and unexpressed intergenic regions. Most annotated protein-coding genes not included in the phenotype gene dataset had high FL scores (median=0.86; **Fig. 2.4C**) and 80% were predicted as functional. The features exhibited by low-scoring protein-coding genes and high-scoring pseudogenes are discussed in the Supplemental Information. By contrast, the FLs were low for both TEs (median=0.03, **Fig. 2.4D**) and unexpressed intergenic regions (median=0.07; **Fig. 2.4E**), 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 FL measure provides a useful metric to distinguish between phenotype genes and pseudogenes. In addition, the FLs of annotated protein-coding genes, TEs, and unexpressed intergenic sequences agree with *a priori* expectations regarding the functionality of these sequences.

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

We sought to apply functional prediction models to ITRs, which often exhibit narrow expression patterns (Supplemental Fig. 2.1A). However, given the association between transcription activity features and functional predictions (**Fig. 2.2A; Fig. 2.3C)**, we first 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 FLs (median=0.96) as those with phenotypes under standard growth conditions (median=0.97; U test, p=0.38, Supplemental Fig. 2.2A). Thus, our model can capture

conditionally functional sequences. Next, we evaluated FL distributions among sequences with different breadths of expression. For this comparison, we focused on non-stress, single-tissue expression datasets (Supplemental Table 2.5), which was distinct from the expression breadth feature in the prediction model that considered all datasets. While phenotype genes were better predicted than pseudogenes among sequences with the same number of tissues with expression evidence (U tests, all $p < 1.7E-06$; Supplemental Fig. 2.2B), 65% of the 62 phenotype genes expressed in ≤3 tissues were predicted as non-functional. Further, there was a significant correlation between the number of tissues with expression evidence and FL values of all sequences in our analysis ($r^2=0.77$; $p < 2E-16$). Consistent with misclassifications among narrowly-expressed phenotype genes, a key difference between 80 pseudogenes predicted as functional (high-FL) and 683 pseudogenes predicted as non-functional was that high-FL pseudogenes were more highly and broadly expressed (**Fig. 2.5**). Thus, the function prediction model is biased against narrowly-expressed sequences, regardless of whether they are functional or not.

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. 2.3; Supplemental Table 2.4). 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. 2.1C), indicating that the tissue-agnostic model is more suitable for predicting the functionality of narrowly-expressed sequences than the full model, although there was an increase in FPR (from

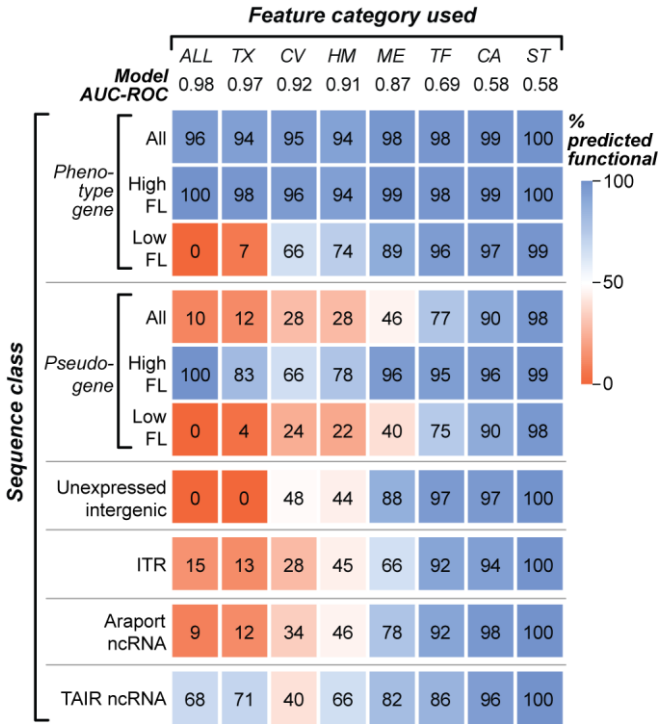


Figure 2.5 Proportion of sequence classes predicted as functional in the full and single-category models. Percentages of sequence classes that are predicted as functional in models based on all features and the single category models, each using all features from a category (abbreviated according to **Fig. 2.2**). The models are sorted from left to right based on performance (AUC-ROC). The colors of and numbers within the blocks indicate the proportion sequences predicted as functional by a given model. Phenotype gene and pseudogene sequences are shown in three sub-groups: all sequences (All), and those predicted as functional (high functional likelihood (FL)) and non-functional (low FL) in the full model. ITR: intergenic transcribed regions.

10% to 15%). We next sought to evaluate the FL of ITR and annotated ncRNA sequences utilizing both the full model and the tissue-agnostic model, as these sequences were often narrowly-expressed (Supplemental Fig. 2.5A).

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

A subset of ITRs and ncRNAs likely represent novel genes or unannotated exon extensions of known genes (Johnson et al., 2005; Moghe et al., 2013). Nevertheless, most ITRs and ncRNAs are functionally ambiguous, as they are predominantly identified by the presence of expression evidence and few have been characterized genetically. To evaluate the functionality of ITRs and ncRNAs, we applied both the full and tissue-agnostic models to 895 ITRs, 136 TAIR ncRNAs, and 252 Araport long ncRNAs (referred to as Araport ncRNAs; see Methods) that do not overlap with any annotated genome features. Consistent with previous studies (Moghe et al., 2013), ITRs and ncRNAs in our dataset were narrowly and weakly expressed and poorly conserved compared to phenotype genes, and ITRs in particular had biochemical characteristics that were generally more consistent with pseudogenes (Supplemental Fig. 2.4). The median FLs based on the full model were low (0.09) for both ITRs (**Fig. 2.4F**) and Araport ncRNAs (**Fig. 2.4G**), and only 15% and 9% of these sequences were predicted as functional, respectively. By contrast, TAIR ncRNAs had a significantly higher median FL value (0.53; U tests, both p<5e-31; **Fig. 2.4H**) and 68% were predicted as functional. The higher proportion of functional TAIR ncRNA predictions compared to ITRs and Araport ncRNAs could be best explained by differences in features from the transcription activity category (**Fig. 2.5**; Supplemental Fig. 2.4). We also note that a greater proportion of ITRs and Araport ncRNAs are predicted as functional when considering only DNA methylation or H3 mark features (**Fig. 2.5**).

However, these two category-specific models also had higher false positive rates (unexpressed intergenic sequences and pseudogenes, **Fig. 2.5**). Thus, single feature-category models do not provide additional support for the functionality of most Araport ncRNAs and ITRs.

We next applied the tissue-agnostic model that is less biased against narrowly-expressed sequences (Supplemental Fig. 2.2C) to ITRs and TAIR/Araport ncRNAs that were generally narrowly-expressed (Supplemental Fig. 2.1A). Compared to the full model, around twice as many ITRs (30%) and Araport ncRNAs (19%) but a similar number of TAIR ncRNA (67%) were predicted as functional. Considering the union of the full and tissue-agnostic model predictions, 268 ITRs (32%), 57 Araport ncRNAs (23%), and 105 TAIR ncRNAs (77%) were likely functional. ITRs and annotated ncRNAs closer to annotated genes tended to be predicted as functional (Supplemental Fig. 2.5A). Using the 95th percentile of intron lengths for all genes as a threshold to call ITRs and annotated ncRNAs as proximal or distal to neighboring genes, 57% of likely functional and 35% of likely non-functional ITRs and ncRNAs were proximal to neighboring genes, respectively (FET, $p < 2E-09$). To assess if a subset these likely functional, proximal ITRs/ncRNAs may be unannotated exons of known genes, we assessed whether they tended to have similar features with their neighbors. Compared to feature similarities between neighboring and random gene pairs (Supplemental Fig. 2.5B-D), likely functional ITRs/ncRNAs were less similar to their neighbors, regardless of proximity (Supplemental Fig. 2.7C,D). Thus, despite their proximity to annotated genes, it remains unclear if some ITRs or annotated ncRNAs represent unannotated exon extensions of known genes or not. In addition, 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. Given the challenge in ascertaining the origin of likely functional, proximal ITRs/ncRNAs, we

instead conservatively estimate that 116 distal, functional ITRs and annotated ncRNAs may represent fragments of novel genes.

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

Thus far, we predicted the majority of ITR and annotated ncRNA sequences as non-functional. We demonstrated that the full model was able to predict conditional phenotype genes (Supplemental Fig. 2.2A) and the tissue-agnostic model was more effective than the full model in predicting narrowly expressed phenotype genes (Supplemental Fig. 2.2B,C). Thus, conditional or tissue-specific functionality do not fully explain why the majority of ITRs and ncRNAs are predicted as non-functional. However, the function prediction models so far were built by contrasting protein-coding genes with pseudogenes and it remains possible that these protein-coding gene-based models can not accurately predict RNA genes. To evaluate this possibility, we generated a tissue-agnostic model using features calculated from a randomly-selected 100 bp sequence within a phenotype protein-coding gene or pseudogene body (for features, see Supplemental Table 2.6). The reason for using 100 bp sequences is that most RNA genes are too short to be considered by earlier models, which were based on 500 bp sequences. In addition, features from the tissue agnostic model are more suitable for RNA gene prediction as annotated RNA genes tend to be more narrowly expressed than phenotype genes (U tests, all $p < 2e\text{-}05$; Supplemental Fig. 2.1B). The 100 bp tissue-agnostic model performed similarly to the full 500 bp model in distinguishing between phenotype protein-coding genes and pseudogenes, except with higher FNR (AUC-ROC=0.97; FNR=13%; FPR=5%; Supplemental Fig. 2.6), but only predicted three out of six RNA genes with documented mutant phenotypes (phenotype RNA genes) as functional (Supplemental Fig. 2.6I). Further, other RNA Pol II-transcribed RNA genes

exhibited mixed predictions from the 100 bp tissue-agnostic model, as 15% of microRNA (miRNA) primary transcripts (Supplemental Fig. 2.6J), 73% of small nucleolar RNAs (snoRNAs; Supplemental Fig. 2.6K), and 50% of small nuclear RNAs (snRNAs; Supplemental Fig. 2.6L) were predicted as functional. Although the proportion of phenotype RNA genes predicted as functional (50%) is significantly higher than the proportion of pseudogenes predicted as functional (5%, FET, $p < 0.004$), this finding suggests that a model built with protein-coding genes has a substantial FNR for detecting RNA genes.

To determine whether the suboptimal predictions by the phenotype protein-coding gene-based models are because RNA genes belong to a class of their own, we next built a multi-class function prediction model aimed at distinguishing four classes of sequences: benchmark RNA genes (n=46), phenotype protein-coding genes (1,882), pseudogenes (3,916), and randomly-selected, unexpressed intergenic regions (4,000). Benchmark RNA genes include six phenotype RNA genes and 40 high-confidence miRNA primary transcript sequences (see Methods). RNA phenotype genes exhibit a phenotype when mutated and are likely under selection, fulfilling the selected effect definition of functionality. However, the lack of a sizeable sample of RNA genes with documented phenotypes required that we also include annotated RNA genes with no phenotype information. In addition, unexpressed intergenic sequences were included to provide another set of likely non-functional sequences distinct from pseudogenes. Expression breadth and tissue-specific features were excluded from the four-class model and 100 bp sequences were used. In the four-class model, 87% of benchmark RNA genes, including all six phenotype RNA genes, were predicted as functional sequences (65% RNA gene-like and 22% phenotype protein-coding gene-like; **Fig. 2.6A**). In addition, 95% of phenotype genes were predicted as functional (**Fig. 2.6B**), including 80% of narrowly expressed genes, an increase of 22% over the 500 bp

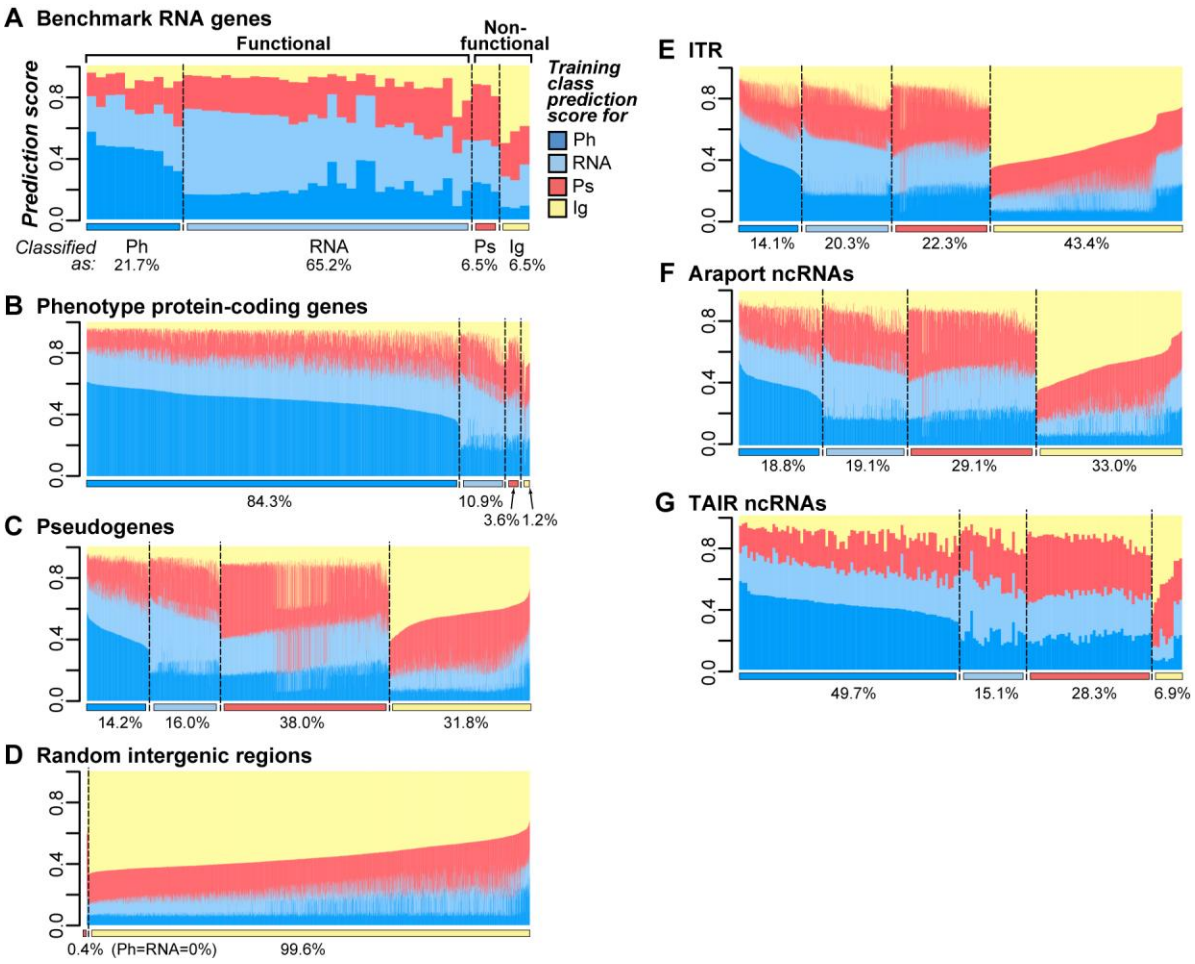


Figure 2.6 Function predictions based on a four-class prediction model. (*A*) Stacked bar plots indicate the prediction scores of benchmark RNA genes for each of the four classes: dark blue - phenotype protein-coding gene (Ph), cyan - RNA gene (RNA), red - pseudogene (Ps), yellow – random intergenic sequence (Ig). A benchmark RNA gene is classified as one of the four classes according to the highest prediction score. The color bars below the chart indicate the predicted class, with the same color scheme as the prediction score. Sequences classified as Ph or RNA were considered functional, while those classified as Ps or Ig were considered non-functional. Percentages below a classification region indicate the proportion of sequences classified as that class. (*B*) Phenotype protein-coding gene prediction scores. (*C*) Pseudogene

Figure 2.6 (cont'd)

prediction scores. (*D*) Random unexpressed intergenic region prediction scores. Note that no sequence was predicted as functional. (*E*) Intergenic transcribed region (ITR), (*F*) Araport11 ncRNA regions. (*G*) TAIR10 ncRNA regions.

tissue-agnostic model (Supplemental Fig. 2.2C). For the remaining two sequence classes, 70% of pseudogenes (**Fig. 2.6C**) and 100% of unexpressed intergenic regions (**Fig. 2.6D**) were predicted as non-functional (either pseudogenes or unexpressed intergenic sequences). Thus, the four-class model improves prediction accuracy of RNA genes and narrowly expressed genes. However, the inclusion of RNA genes in the model has significantly increased the ambiguity in pseudogene classification.

Since the four-class model was able to distinguish benchmark RNA genes from non-functional sequence classes, we next evaluated whether ITRs and annotated ncRNAs resemble functional sequences with the four-class model. Note that the 100 bp model used here allowed us to evaluate an additional 10,938 ITRs and 1,406 annotated ncRNAs. We found that 34% of ITR, 38% of Araport ncRNA, and of 65% TAIR ncRNAs were predicted as functional sequences (**Fig. 2.6E-G**). More specifically, 20% or fewer ITR and annotated ncRNA sequences were most similar to RNA genes (**Fig. 2.6E-G**), suggesting that most are not functioning as miRNA, which represented the majority of benchmark RNA sequences. However, other potential roles as RNA regulators, including *cis*-acting (Guil and Esteller, 2012) and competitive endogenous (Tan et al., 2015) regulatory functions, should be further studied. To provide an overall estimate of the proportion of likely-functional ITRs and annotated ncRNAs, we considered the predictions from the four-class model (**Fig. 2.6**), the full model (**Fig. 2.3,2.4**), and the tissue-agnostic models (Supplemental Fig. 2.3,2.6). Based on support from at least one of the four models, we classified 4,437 ITRs (38%) and 796 annotated ncRNAs (44%) as functional, as they resembled either phenotype protein-coding or RNA genes. Our findings lend support that they are likely parts of novel or annotated genes. Meanwhile, we find that a substantial number of ITRs (62%) and annotated ncRNAs (56%) are predicted as non-functional. Moreover, at least a third of ITRs

(**Fig. 2.6E**) and Araport ncRNAs (**Fig. 2.6F**) most closely resemble unexpressed intergenic regions. Thus, we show that the majority of ITRs and annotated ncRNA regions resemble non-functional genomic regions, and therefore could represent regions of noisy transcription.

CONCLUSION

Discerning the location of functional regions within a genome represents a key goal in genomic biology. Despite advances in computational gene finding, it remains challenging to determine whether intergenic transcribed regions (ITRs) represent functional or noisy biochemical activity. We established robust function prediction models based on the evolutionary, biochemical, and structural characteristics of phenotype genes and pseudogenes. The prediction models accurately define functional and non-functional regions and are applicable genome-wide. These results echo recent findings that human phenotype genes could be distinguished from pseudogenes (Tsai et al., 2017). Given that function predictions were successful in both plant and metazoan model systems, integrating the evolutionary and biochemical features of known genes will likely be applicable to any species. The next step will be to test whether function prediction models can be applied across species, which could ultimately allow the phenotype data and 'omics resources available in model systems to effectively guide the identification of functional regions in non-models.

Expression data was highly informative to functional predictions. We found that the prediction model based on only 24 transcription activity-related features performs nearly as well as the full model that integrates additional information including conservation, H3 mark, methylation, and TF binding data. In human, use of transcription data from cell lines also produced highly accurate predictions of functional genomic regions (AUC-ROC=0.96) (Tsai et al., 2017). Despite the importance of transcription data, we emphasize that the presence of expression evidence is an extremely poor predictor. Taken together, these results indicate that function prediction models can be established in any species, model or not, with a modest

number of transcriptome datasets (e.g. 51 in this study and 19 in human). One caveat of the current model is that narrowly-expressed phenotype genes are frequently predicted as pseudogene and broadly-expressed pseudogenes tend to be called functional. To improve the function prediction model, it will be important to explore additional features unrelated to transcription. Because few phenotype genes are narrowly-expressed (5%) in the *A. thaliana* training data, more phenotyping data for narrowly expressed genes will be crucial as well.

Upon application of the function prediction models genome-wide, we found that 4,427 ITRs and 796 annotated ncRNAs in *A. thaliana* are likely functional. Assuming each entry equals a novel gene, this estimate represents a 19% increase in annotated gene space (excluding annotated ncRNAs) for the model plant. However, considering the high false positive rates (e.g. 10% for the full and 31% for the four-class model), this is most likely an overestimate of the number of novel genes contributed by functional ITRs and annotated ncRNAs. In addition, we emphasize that the majority of ITRs and ncRNAs resemble pseudogenes and random unexpressed intergenic regions. Similarly, most human ncRNAs are more similar to non-functional sequences than they are to protein coding and RNA genes (Tsai et al., 2017). Furthermore, the significant relationship between the amount of intergenic expression occurring in a species and the size of a genome is consistent with the interpretation that intergenic transcripts are generally non-functional. Thus, instead of assuming any expressed sequence must be functionally significant, we advocate that the null hypothesis should be that it is not, particularly considering that most ITRs and annotated ncRNAs have not been experimentally characterized. The machine learning framework we have described provides an approach for distinguishing between functional and noisy biochemical activity, and will help defining the gene space in a genome.

METHODS

Identification of transcribed regions in leaf tissue of 15 flowering plants

RNA-sequencing (RNA-seq) datasets were retrieved from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov/sra/) for 15 flowering plant species (Supplemental Table 2.1). 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 v.11 (www.phytozome.net) (Goodstein et al., 2012) or Oropetium Base v.01 (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 was not available. Only one end from paired-end read datasets were utilized in downstream processing. Reads were trimmed to be rid of low scoring ends and residual adaptor sequences using Trimmomatic v0.33 (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:20) (Bolger et al., 2014) and mapped to genome sequences using TopHat v2.0.13 (default parameters except as noted below) (Kim et al., 2013). Reads ≥20 nucleotides in length that mapped uniquely within a genome were used in further analysis.

For each species, thirty million mapped reads were randomly selected from among all datasets and assembled into transcript fragments using Cufflinks v2.2.1 (default parameters except as noted below) (Trapnell et al., 2010), while correcting for sequence-specific biases during the sequencing process by providing an associated genome sequence with the -b flag. The expected mean fragment length for assembled transcript fragments in Cufflinks was set to 150 from the default of 200 so that expression levels in short fragments would not be overestimated.

The 1st and 99th percentile of intron lengths for each species were used as the minimum and maximum intron lengths, respectively, for both the TopHat2 and Cufflinks steps. Intergenic transcribed regions (ITRs) were defined by transcript fragments that did not overlap with gene annotation and did not have significant six-frame translated similarity to plant protein sequences in Phytozome v.10 (BLASTX E-value < 1E-05). The correlation between assembled genome size and gene counts was determined with data from the first 50 published plant genomes (Michael and Jackson, 2013).

Phenotype data sources

Mutant phenotype data for *A. 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. (Lloyd et al., 2015). Phenotype genes used in our analyses were those whose disruption resulted in lethal or visible defects under standard laboratory growth conditions. Genes with documented mutant phenotypes under standard conditions were considered as a distinct and non-overlapping category from other annotated protein-coding genes. We identified six RNA genes with documented loss-of-function phenotypes through literature searches (Supplemental Table 2.7): *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). Conditional phenotype genes were those belonging to the Conditional phenotype group as described by Lloyd and Meinke (Lloyd and Meinke, 2012). Loss-of-function mutants of these genes exhibited phenotype only under stress conditions.

Arabidopsis thaliana genome annotation

A. thaliana protein-coding gene, miRNA gene, snoRNA gene, snRNA gene, ncRNA region, pseudogene, and transposable element annotations were retrieved from The Arabidopsis Information Resource v.10 (TAIR10; www.arabidopsis.org) (Berardini et al., 2015). Additional miRNA gene and lncRNA region annotations were retrieved from Araport v.11 (www.araport.org). A primary difference between the TAIR ncRNAs and Araport lncRNAs (referred to as Araport ncRNAs in the Results & Discussion section) is the date in which they were annotated. For example, 221 ncRNAs were present in the v.7 release of TAIR, which dates back to 2007 (TAIR10 contains 394 ncRNA annotations) (Swarbreck et al., 2008; Lamesch et al., 2012; Berardini et al., 2015). However, Araport lncRNAs were annotated in the past five years (Krishnakumar et al., 2015). Thus, that TAIR ncRNAs are generally more highly and broadly expressed is likely a result of the less sensitive transcript identification methods available for early TAIR releases. A pseudogene-finding pipeline (Zou et al., 2009) was used to identify additional pseudogene fragments and count the number of disabling mutations (premature stop or frameshift mutations). Genes, pseudogenes, and transposons with overlapping annotation were excluded from further analysis. Overlapping lncRNA annotations were merged for further analysis. When pseudogenes from TAIR10 and the pseudogene-finding pipeline overlapped, the longer pseudogene annotation was used.

A. thaliana ITRs analyzed include: (1) the Set 2 ITRs in Moghe et al. (Moghe et al., 2013), (2) the novel transcribed regions from Araport v.11, and (3) additional ITRs from 206 RNA-seq datasets (Supplemental Table 2.5). Reads were trimmed, mapped, and assembled into transcript fragments as described above, except that overlapping transcript fragments from across datasets were merged. ITRs analyzed did not overlap with any TAIR10, Araport11, or

pseudogene annotation. Overlapping ITRs from different annotated subsets were kept based on a priority system: Araport11 > Set 2 ITRs from Moghe et al. (Moghe et al., 2013) > ITRs identified in this study. For each sequence entry (gene, ncRNA, pseudogene, transposable element, or ITR), a 100 and 500 base pair (bp) window was randomly chosen for calculating feature values and subsequent model building steps. Feature descriptions are provided in the following sections. The feature values for randomly selected 500 and 100 bp windows are provided in Supplemental Tables 2.2 and 2.6, respectively. Additionally, non-expressed intergenic sequences were randomly-sampled from genome regions that did not overlap with annotated genes, pseudogenes, transposable elements, or regions with genic or intergenic transcript fragments (100 bp, n=4,000; 500 bp, n=3,716). All 100 and 500 bp windows described above are referred to as sequence windows throughout the Methods section.

Sequence conservation and structure features

There were 10 sequence conservation features examined. The first two were derived from comparisons between *A. thaliana* accessions including nucleotide diversity and Tajima's D among 81 accessions (Cao et al., 2011) using a genome matrix file from the 1,001 genomes database (www.1001genomes.org). The python scripts are available through GitHub (https://github.com/ShiuLab/GenomeMatrixProcessing). The remaining eight features were derived from cross-species comparisons, three based on multiple sequence alignments and five based on pairwise alignments. Three multiple sequence alignment-based features were established using aligned genomic regions between *A. thaliana* and six other plant species (*Glycine max, Medicago truncatula, Populus trichocarpa, Vitis vinifera, Sorghum bicolor,* and *Oryza sativa*) (Li et al., 2012a), which are referred to as conserved blocks. For each conserved block, the first feature was the proportion of a sequence window that overlapped a conserved

block (referred to as coverage), and the two other features were the maximum and average phastCons scores within each sequence window. The phastCons score was determined for each nucleotide within conserved blocks (Li et al., 2012a). Nucleotides in a sequence window that did not overlap with a conserved block were assigned a phastCons score of 0. For each sequence window, five pairwise alignment-based cross-species conservation features were the percent identities to the most significant BLASTN match (if E-value<1E-05) in each of five taxonomic groups. The five taxonomic groups included the *Brassicaceae* family ($n_{species}$=7), other dicotyledonous plants (22), monocotyledonous plants (7), other embryophytes (3), and green algae (5). If no sequence with significant similarity was present, percent identity was scored as zero.

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

Transcription activity features

We generated four multi-dataset and 20 individual dataset transcription activity features. To identify a set of RNA-seq datasets to calculate multi-dataset features, we focused on the 72 of 206 RNA-seq datasets each with ≥20 million reads (see above; Supplemental Table 2.5). Transcribed regions were identified with TopHat2 and Cufflinks as described in the RNA-seq analysis section except that the 72 *A. thaliana* RNA-seq datasets were used. Following transcript

assembly, we excluded 21 RNA-seq datasets because they had unusually high RPKM (Reads Per Kilobase of transcript per Million mapped reads) values (median RPKM value range=272~2,504,294) compared to the rest (2~252). The remaining 51 RNA-seq datasets were used to generate four multi-dataset transcription activity features including: expression breadth, 95th percentile expression level, maximum transcript coverage, and presence of expression evidence (for values see Supplemental Table 2.2, 2.6). Expression breadth was the number of RNA-seq datasets that have ≥1 transcribed region that overlapped with a sequence window. The 95th percentile expression level was the 95th percentile of RPKM values across 51 RNA-seq datasets where RPKM values were set to 0 if there was no transcribed region for a sequence window. Maximum transcript coverage was the maximum proportion of a sequence window that overlapped with a transcribed region across 51 RNA-seq datasets. Presence of expression evidence was determined by overlap between a sequence window and any transcribed region in the 51 RNA-seq datasets.

In addition to features based on multiple datasets, 20 individual dataset features were derived from 10 datasets: seven tissue/organ-specific RNA-seq datasets including pollen (SRR847501), seedling (SRR1020621), leaf (SRR953400), root (SRR578947), inflorescence (SRR953399), flower, (SRR505745) and silique (SRR953401), and three datasets from non-standard growth conditions, including dark-grown seedlings (SRR974751) and leaf tissue under drought (SRR921316) and fungal infection (SRR391052). For each of these 10 RNA-seq datasets, we defined two features for each sequence window: the maximum transcript coverage (as described above) and the maximum RPKM value of overlapping transcribed regions (referred to as Level in **Fig. 2.2**). If no transcribed regions overlapped a sequence window, the maximum RPKM value was set as 0. For the analysis of narrowly- and broadly-expressed phenotype genes

and pseudogenes (Supplemental Fig. 2.2B,C), we used 28 out of 51 RNA-seq datasets generated from a single tissue and in standard growth conditions to calculate the number of tissues with evidence of expression (tissue expression breadth). In total, seven tissues were represented among the 28 selected RNA-seq datasets (see above; Supplemental Table 2.5), and thus tissue expression breadth ranges from 0 to 7 (note that only 1 through 7 are shown in Supplemental Fig. 2.2B,C due to low sample size of phenotype genes in the 0 bin). The tissue breadth value is distinct from the expression breadth feature used in model building that was generated using all 51 datasets and considered multiple RNA-seq datasets from the same tissue separately (range: 0-51).

Histone 3 mark features

Twenty histone 3 (H3) mark features were calculated based on eight H3 chromatin immunoprecipitation sequencing (ChIP-seq) datasets from SRA. The H3 marks examined include four associated with activation (H3K4me1: SRR2001269, H3K4me3: SRR1964977, H3K9ac: SRR1964985, and H3K23ac: SRR1005405) and four associated with repression (H3K9me1: SRR1005422, H3K9me2: SRR493052, H3K27me3: SRR3087685, and H3T3ph: SRR2001289). Reads were trimmed as described in the RNA-seq section and mapped to the TAIR10 genome with Bowtie v2.2.5 (default parameters) (Langmead et al., 2009). Spatial Clustering for Identification of ChIP-Enriched Regions v.1.1 (Xu et al., 2014) was used to identify ChIP-seq peaks with a false discover rate ≤ 0.05 with a non-overlapping window size of 200, a gap parameter of 600, and an effective genome size of 0.92 (Koehler et al., 2011). For each H3 mark, two features were calculated for each sequence window: the maximum intensity among overlapping peaks and peak coverage (proportion of overlap with the peak that overlaps maximally with the sequence window). In addition, four multi-mark features were generated.

Two of the multi-mark features were the number of activating marks (0-4) overlapping a sequence window and the proportion of a sequence window overlapping any peak from any of the four activating marks (activating mark peak coverage). The remaining two multi-mark features were the same as the two activating multi-mark features except focused on the four repressive marks.

DNA methylation features

Twenty-one DNA methylation features were calculated from bisulfite-sequencing (BS-seq) datasets from seven tissues (pollen: SRR516176, embryo: SRR1039895, endosperm: SRR1039896, seedling: SRR520367, leaf: SRR1264996, root: SRR1188584, and inflorescence: SRR2155684). BS-seq reads were trimmed as described above and processed with Bismark v.3 (default parameters) (Krueger and Andrews, 2011) to identify methylated and unmethylated cytosines in CG, CHH, and CHG (H = A, C, or T) contexts. Methylated cytosines were defined as those with ≥5 mapped reads and with >50% of mapped reads indicating that the position was methylated. For each BS-seq dataset, the percentage of methylated cytosines in each sequence window for CG, CHG, and CHH contexts were calculated if the sequence window had ≥5 cytosines with ≥5 reads mapping to the position. To determine whether the above parameters where reasonable, we assessed the false positive rate of DNA methylation calls by evaluating the proportion of cytosines in the chloroplast genome that are called as methylated, as the chloroplast genome has few DNA methylation events (Ngernprasirtsiri et al., 1988; Zhang et al., 2006). Based on the above parameters, 0-1.5% of cytosines in CG, CHG, or CHH contexts in the chloroplast genome were considered methylated in any of the seven BS-seq datasets. This indicated that the false positive rates for DNA methylation calls were low and the parameters were reasonable.

Chromatin accessibility and transcription factor binding features

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

Single-feature prediction performance

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

equal to the threshold and thus falsely predicted as functional. FNR were calculated as the percentage of positive (functional) cases with values below the threshold and thus falsely predicted as non-functional.

Binary classification with machine learning

For binary classification (two-class) models that contrasted phenotype genes and pseudogenes, the random forest (RF) implementation in the Waikato Environment for Knowledge Analysis software (WEKA) (Hall et al., 2009) was utilized. Three types of two-class models were established, including the full model (500 bp sequence window, **Fig. 2.3C**,**D** and **Fig. 2.4**), tissue-agnostic models (500 bp, Supplemental Fig. 2.3; 100 bp, Supplemental Fig. 2.6), and single feature category models (**Fig. 2.3C,D**). For each model type, we first generated 100 balanced datasets by randomly selecting equal numbers of phenotype genes (positive examples) and pseudogenes (negative examples). For each of these 100 datasets, 10-fold stratified cross-validation was utilized, where the model was trained using 90% of sequences and tested on the remaining 10%. Thus, for each model type, a sequence window had 100 prediction scores, where each score was the proportion of 500 random forest trees that predicted a sequence as a phenotype gene in a balanced dataset. The median of 100 prediction scores was used as the functional likelihood (FL) value (Supplemental Table 2.4). The FL threshold to predict a sequence as functional or non-functional was defined based on maximum F-measure as described in the previous section. We tested multiple -K parameters (2 to 25) in the WEKA-RF implementation, which alters the number of randomly-selected features included in each RF tree (Supplemental Table 2.8), and found that 15 randomly-selected features provided the highest performance based on AUC-ROC (calculated and visualized using the ROCR package) (Sing et al., 2005). Feature importance was assessed by excluding one feature at a time to determine the

associated reduction in prediction performance (Supplemental Table 2.9). All leave-one-out models performed well (AUC-ROC >0.97), indicating that no single feature was dominating the function predictions and/or many features are correlated (Supplemental Fig. 2.7). Binary classification models were also built using all features from 500 bp sequences (equivalent to the full model) with the Sequential Minimal Optimization - Support Vector Machine (SMO-SVM) implementation in WEKA (Hall et al., 2009). The results of SMO-SVM models were highly similar to the full RF results: *PCC* between the FL values generated by RF and SMO-SVM=0.97; AUC-ROC of SMO-SVM=0.97; FPR=12%; FNR=3%. By comparison, the full RF model had AUC-ROC=0.98, FPR=10%, FNR=4%.

Tissue-agnostic models were generated by excluding the expression breadth feature and 95th percentile expression level and replacing all features from RNA-seq, BS-seq, and DHS datasets that were available in multiple tissues. For multiple-tissue RNA-seq data, the maximum expression level across 51 RNA-seq datasets (in RPKM) and maximum coverage (as described in the transcription activity section) of a sequence window in any of 51 RNA-seq datasets were used. For multi-tissue DNA methylation features, minimum proportions of methylated cytosines in any tissue in CG, CHG, and CHH contexts were used. For DHS data, the maximum peak intensity and peak coverage was used instead. In single feature category predictions, fewer total features were used and therefore lower –K values (i.e. the number of random features selected when building random forests) were considered in parameter searches (Supplemental Table 2.8).

Multi-class machine learning model

For the four-class model, benchmark RNA gene, phenotype protein-coding gene, pseudogene, and random unexpressed intergenic sequences were used as the four training classes. Benchmark RNA genes consisted of six RNA genes with documented loss-of-function

phenotypes and 40 high-confidence miRNA genes from miRBase (www.mirbase.org) (Kozomara and Griffiths-Jones, 2014). We generated 250 datasets with equal proportions (larger classes randomly sampled) of training sequences. Two-fold stratified cross-validation was utilized due to the low number of benchmark RNA genes. The features included those described for the tissue-agnostic model and focused on 100 bp sequence windows. The RF implementation, *cforest*, in the *party* package of R (Strobl et al., 2008) was used to build the classifiers. The four-class predictions provide prediction scores for each sequence type: an RNA gene, phenotype protein-coding gene, pseudogene, and unexpressed intergenic score (Supplemental Table 2.4). The prediction scores indicate the proportion of random forest trees that classify a sequence as a particular class. Median prediction scores from across 100 balanced runs were used as final prediction scores. Scores from a single balanced dataset models sum to 1, but not the median from 100 balanced runs. Thus, the median scores were scaled to sum to 1. For each sequence window, the maximum prediction score among the four classes was used to classify a sequence as phenotype gene, pseudogene, unexpressed intergenic, or RNA gene.

ACKNOWLEDGEMENTS

I wish to thank Christina Azodi, Ming-Jung Liu, Gaurav Moghe, Bethany Moore, and Sahra Uygun and for providing processed data and discussion.

APPENDIX

Supplemental information

Variability in single feature performance within feature categories

Within each feature category, there was a wide range of performance between features (**Fig. 2.2**, Supplemental Table 2.3) and there were clear biological or technical explanations for features that perform poorly. For the transcription activity category, 17 out of 24 features had an AUC-ROC performance >0.8, including the best-performing feature, expression breadth (AUC-ROC=0.95; **Fig. 2.2A**). However, five transcription activity-related features performed poorly (AUC-ROC<0.65), including the presence of expression (transcript) evidence (AUC-ROC=0.58; **Fig. 2.2A**). 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. 2.2B**). On the other hand, identity to best matching nucleotide sequences found in *Brassicaceae* and algal species were not informative (AUC-ROC=0.55 and 0.51, respectively; **Fig. 2.2B**). This was because 99.8% and 95% of phenotype genes and pseudogenes, respectively, had a potentially homologous sequence within the *Brassicaceae* family and only 3% and 1%, respectively, in algal species. Thus, *Brassicaceae* genomes were too similar and algal genomes too dissimilar to *A. thaliana* to provide meaningful information. H3 mark features also displayed high variability. The most informative H3 mark features were based on the number and coverage of activation-related marks (AUC-ROC=0.87 and 0.85, respectively; **Fig. 2.2E**), consistent with the notion that histone marks are often jointly associated with active genomic sequences to 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) were largely indistinct between phenotype genes and pseudogenes (AUC-ROC range: 0.55-0.59; **Fig. 2.2E**).

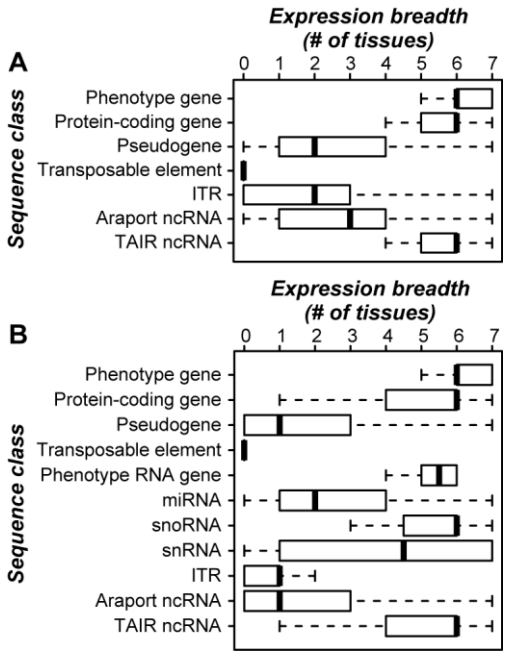
Features of misclassified sequences

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-functional likelihood (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. 2.5**). 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 functional sequences.

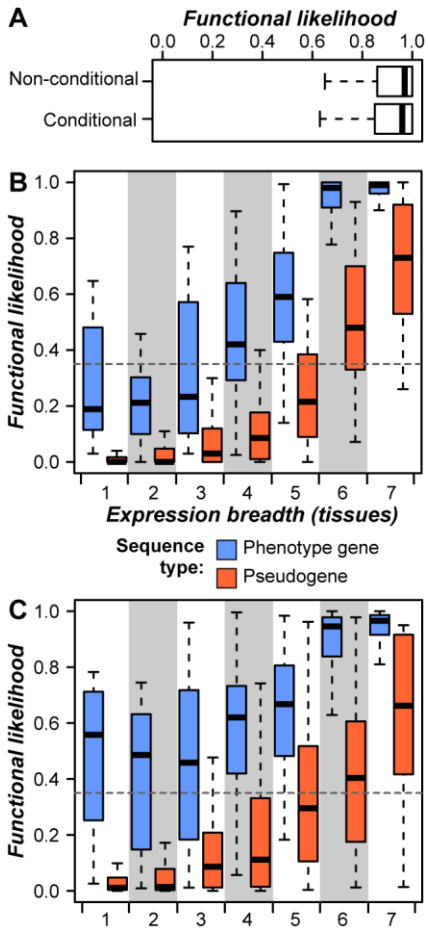
We also predict 80 pseudogenes (10%) to be functional (high-FL pseudogenes). A significantly higher proportion of high-FL pseudogenes came from existing genome annotation as 19% of annotated pseudogenes were classified as functional, compared to 4% of pseudogenes identified through a computational pipeline (FET, $p < 1.5E\text{-}10$) (Zou et al., 2009). 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) were 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 (Poliseno et al., 2010; Karreth et al., 2015). 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.

Among protein-coding genes without phenotype information, we predict 20% 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 these decaying genes represented only 7% of all *A. thaliana* annotated protein-coding genes, they made up 45% of protein-coding genes predicted as non-functional (Fisher's Exact Test (FET), $p < 1E\text{-}11$).



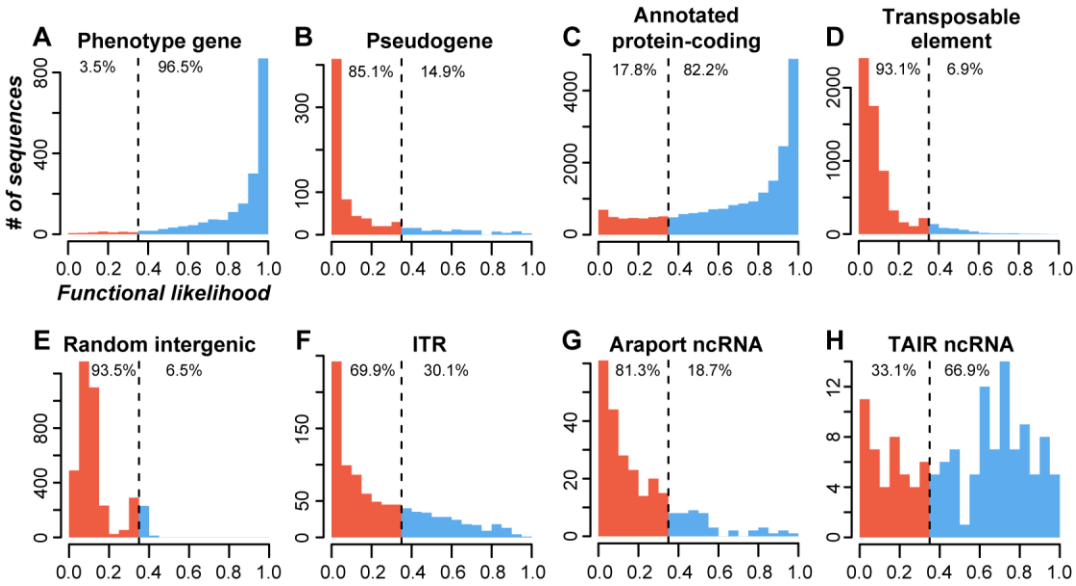
Supplemental Figure 2.1 Expression breadth distributions of sequence classes. (*A*) Based on 500 bp feature regions. (*B*) Based on 100 bp feature regions.



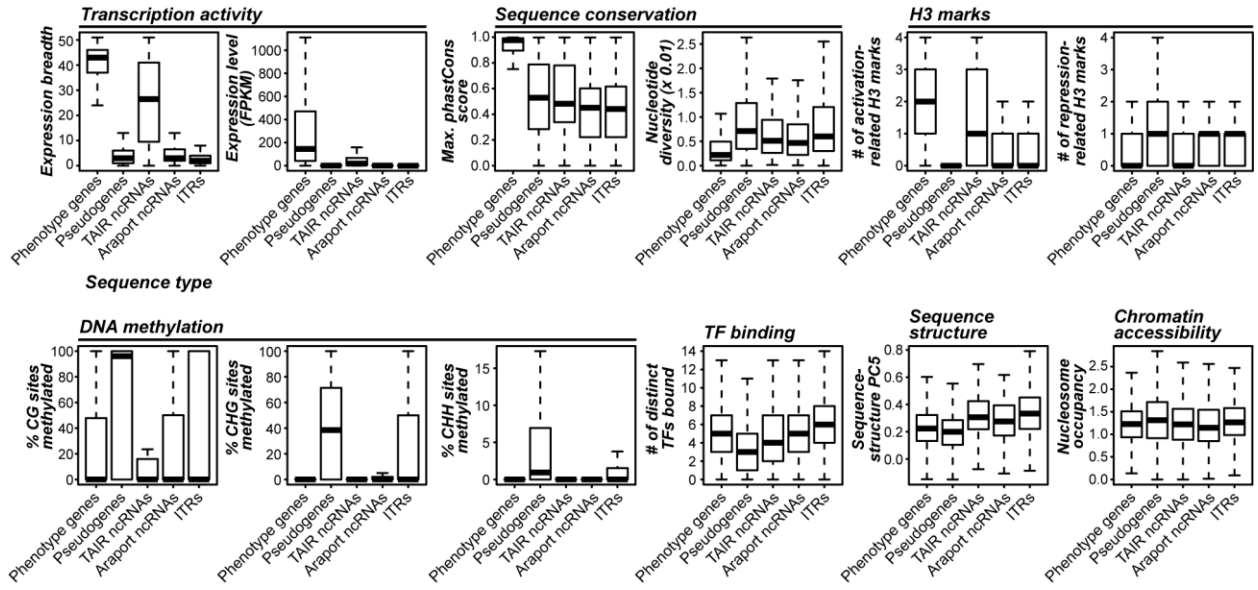
Supplemental Figure 2.2 Impacts of conditional phenotypes and expression breadth on the function prediction model. (*A*) Functional likelihood distributions of phenotype genes with mutant phenotypes under standard growth conditions (non-conditional) and non-standard growth conditions such as stressful environments (conditional) based on the 500 bp full model. Feature values were calculated from a random 500 bp region from within the sequence body. Higher and lower functional likelihood values indicate a greater similarity to phenotype genes and pseudogenes, respectively. (*B*,*C*) Distributions of functional likelihood scores for phenotype genes (blue) and pseudogenes (red) for sequences with various breadths of expression for (*B*) the 500 bp full model and (*C*) the 500 bp tissue-agnostic model generated by excluding the

Supplemental Figure 2.2 (cont'd)

expression breadth and features available from multiple tissues. The tissue-agnostic model is aimed toward minimizing the effects of biochemical activity occurring across multiple tissues and predicts a greater proportion of narrowly-expressed phenotype genes as functional compared to the full model.

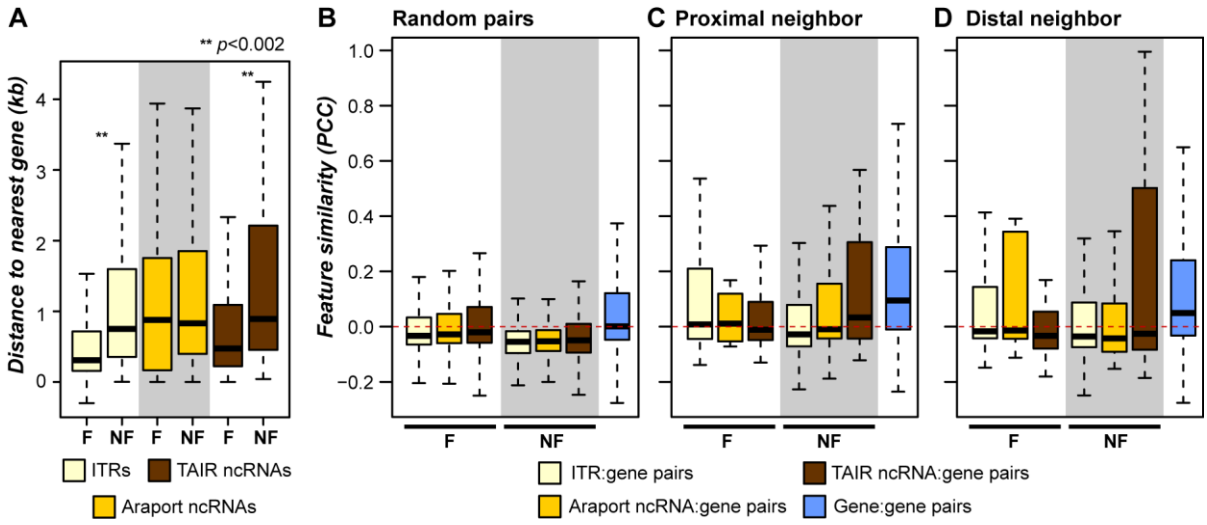


Supplemental Figure 2.3 Distributions of functional likelihood scores based on the 500 bp tissue-agnostic model. (*A*) Phenotype genes. (*B*) Pseudogenes. (*C*) Annotated protein-coding genes. (*D*) Transposable elements. (*E*) Random unexpressed intergenic sequences. (*F*) Intergenic transcribed regions (ITR). (*G*) Araport11 ncRNAs. (*H*) TAIR10 ncRNAs. Vertical dashed lines display the threshold to define 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.

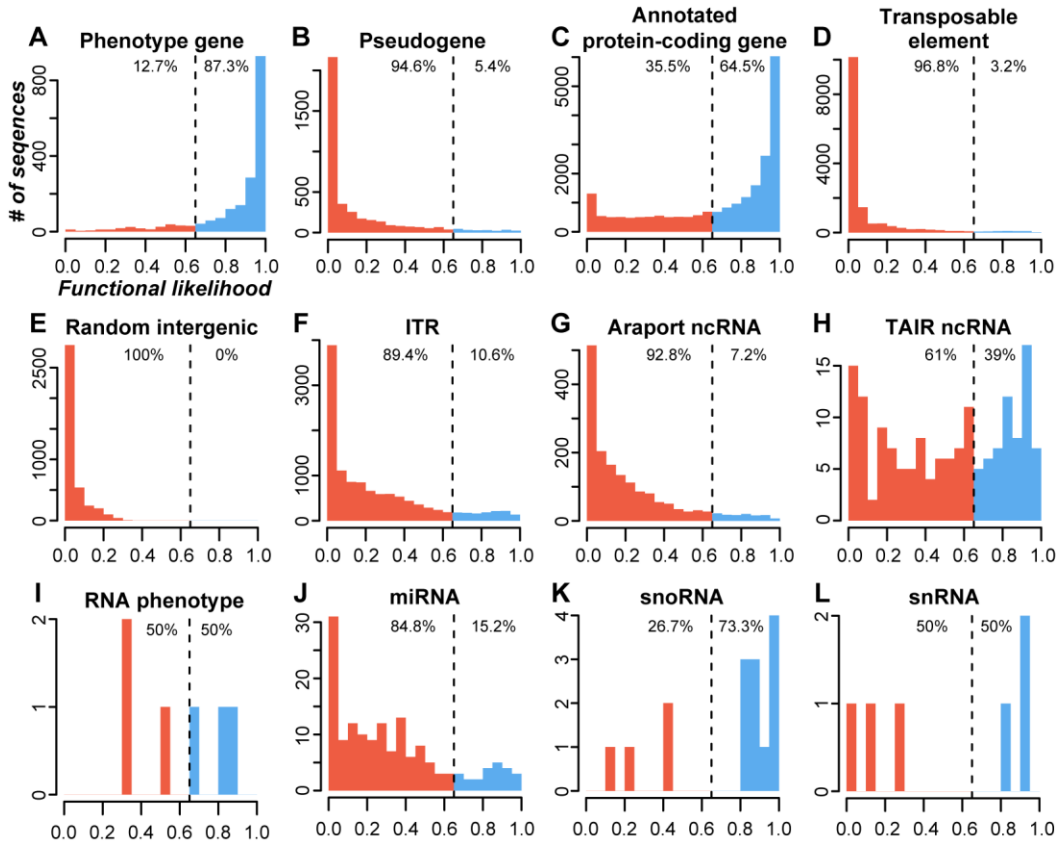


Supplemental Figure 2.4 Distributions of 12 example features from 7 feature categories.

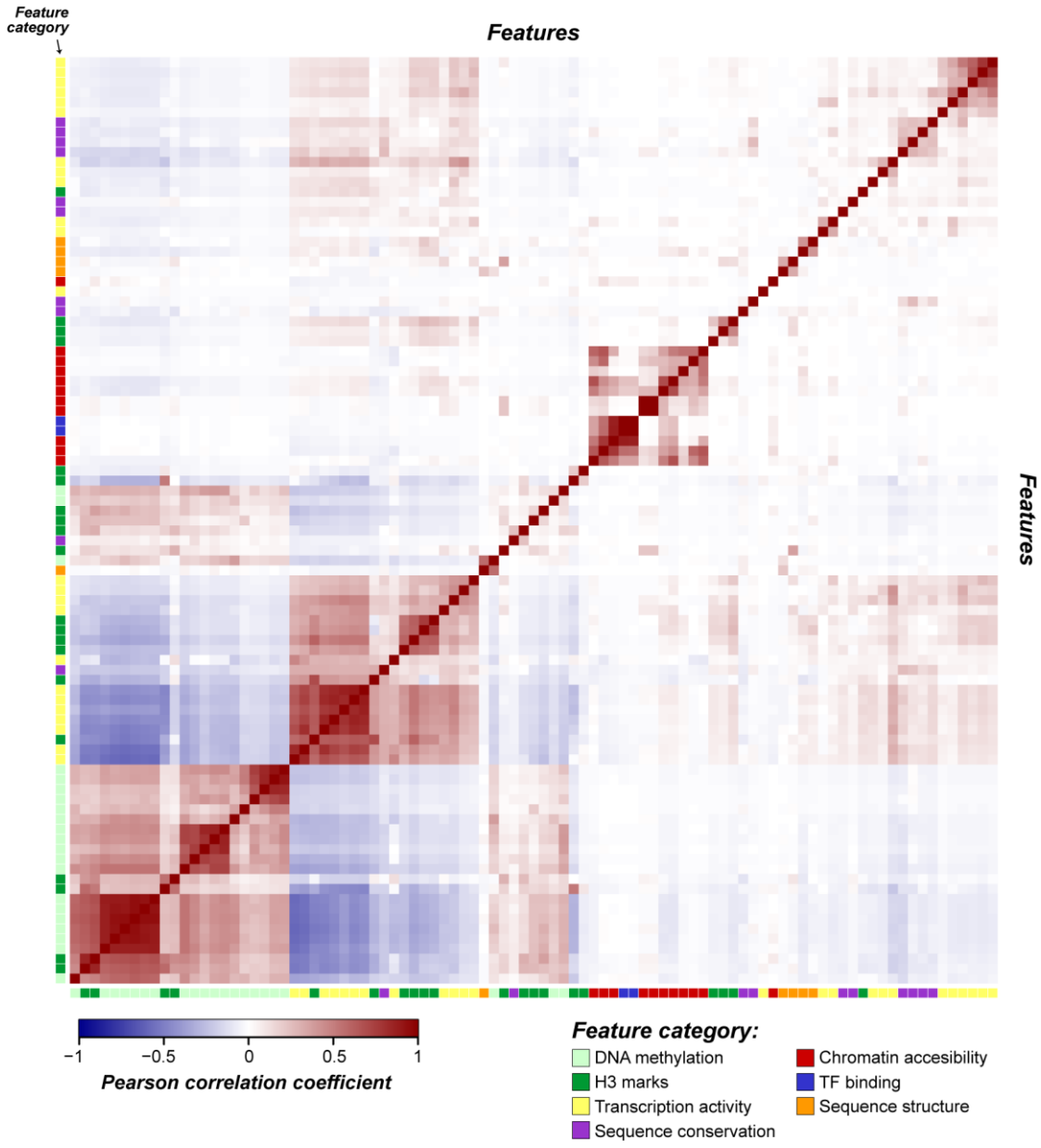
Sequence classes include phenotype genes, pseudogenes, TAIR- and Araport-annotated ncRNAs and intergenic transcribed regions (ITR).



Supplemental Figure 2.5 Distance of ITRs and annotated ncRNA regions to and feature similarity with neighboring genes. (*A*) Distance from intergenic transcribed regions (ITRs) and annotated ncRNAs to the closest neighboring gene. ITR and ncRNA sequences are separated by whether they are predicted as functional (F) or non-functional (NF) by the 500 bp full model. (*B*) Feature similarity based on Pearson's Correlation Coefficients (PCC) between random pairs of ITRs, Araport11 ncRNAs, TAIR10 ncRNAs, or annotated genes. (*C*) Feature similarity between proximal neighbors (within 95th percentile (456 bp) of intron lengths), and (*D*) Feature similarity between distal neighbors (>456 bp). Pairs involving ITRs and annotated ncRNAs were divided by whether the ITR or ncRNA sequence was predicted as functional (F) or non-functional (NF) by the full model. Feature values were quantile normalized prior to calculating correlations.



Supplemental Figure 2.6 Distributions of functional likelihood scores based on the 100 bp tissue-agnostic model. (*A*) Phenotype genes. (*B*) Pseudogenes. (*C*) Protein-coding gene. (*D*) transposable elements. (*E*) Random unexpressed intergenic sequences. (*F*) Intergenic transcribed regions (ITR). (*G*) Araport11 ncRNAs. (*H*) TAIR10 ncRNAs. (*I*) RNA genes with loss-of-function mutant phenotypes. (J) MicroRNAs, (*K*) Small nucleolar RNAs, (*L*) Small nuclear RNAs. The tissue-agnostic model was built with 100 bp features and while excluding the expression breadth and tissue-specific features. Higher functional likelihood values indicate greater similarity to phenotype genes while lower values indicate similarity to pseudogenes. Vertical dashed lines display the threshold to define 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 2.7 Correlation between features used in functional predictions.

Colors within the heatmap indicate pairwise correlation between two features. Colors on the left-most and bottom-most edges indicate the associated feature category (see **Fig. 2.2**). Feature values were quantile normalized prior to calculating correlation.

CHAPTER 3: CROSS-SPECIES AND POST-DUPLICATION EVOLUTIONARY DYNAMICS OF INTERGENIC TRANSCRIBED REGIONS INDICATE WIDESPREAD NOISY TRANSCRIPTION

ABSTRACT

Extensive transcriptional activity occurring in unannotated, intergenic regions of genomes has generated spirited debate over whether intergenic transcription represents the activity of novel genes or noisy expression. A comprehensive investigation of the evolutionary histories of intergenic transcribed regions (ITRs) in multiple species would be informative in this debate but is currently lacking. Here, we evaluated the cross-species and post-duplication sequence and expression conservation of ITRs in four grass (Poaceae) species. The majority of ITRs in our analysis are species-specific sequences and ITR orthologs, when present, were usually not expressed. However, orthologous pairs of ITRs showed similarity in tissue expression patterns approaching that of annotated exon pairs, suggesting that ITRs with expression conservation are enriched for functional sequences. Similarly, ITR duplicates tend to be recent and not expressed, with few examples ITR duplicates retained in synteny from ancient whole genome duplication events. In addition, function prediction models were established in *Oryza sativa* (rice) to classify transcribed regions as likely-functional by integrating evolutionary and biochemical features and were highly capable of distinguishing between benchmark functional and non-functional sequences. Prediction models classified 2,754 rice ITRs (38%) as functional and confirmed that ITRs with expression conservation and those with ancient retained duplicates frequently represent functional sequences. However, evaluations of evolutionary histories primarily highlight ITRs as short-lived sequences with unstable expression, which is consistent with computational function predictions that suggest a majority of ITRs are not under selection.

INTRODUCTION

Advances in sequencing technology have uncovered pervasive transcription occurring throughout unannotated, intergenic space in eukaryotic genomes, including metazoan (ENCODE Project Consortium, 2012; Brown et al., 2014; Boeck et al., 2016), fungal (Nagalakshmi et al., 2008), and plant systems (Yamada et al., 2003; Stolc et al., 2005; Nobuta et al., 2007; Moghe et al., 2013; Krishnakumar et al., 2015). While intergenic transcription was initially thought to be primarily associated with nearby genes (van Bakel et al., 2010), a host of possible new functions indicate many intergenic transcripts may represent the activity of novel, independent genes. These include roles as competitive endogenous RNAs (Tan et al., 2015), *cis*-acting regulatory transcripts (Guil and Esteller, 2012) or small protein-coding regions that are frequently missed by gene finding programs (Hanada et al., 2013). Despite these exciting possibilities, intergenic transcribed regions (ITRs) may also represent the products of noisy transcription, resulting from imperfect regulation of the cellular machinery that controls gene expression (Struhl, 2007). In addition, ITRs may provide the raw materials from which novel genes may evolve *de novo* (Carvunis et al., 2012). Thus, distinguishing between functional intergenic transcripts and those that are the products of noisy transcription represents a difficult but critical task in genome biology.

The foundational step in identifying novel genes via intergenic transcription is to establish what it means to be functional. Throughout this article, we utilize the selected effect definition of function, which requires that a biochemical activity, particularly transcription, have been molded through evolutionary selection and contribute to organismal fitness (Amundson and Lauder, 1994; Graur et al., 2013; Doolittle et al., 2014). This definition stands in contrast to the

casual role definition that considers any reproducible biochemical activity to be functional (Cummins, 1975; Amundson and Lauder, 1994). Under the causal role definition, intergenic transcription would be considered functional *de facto*. However, the possibility of noisy transcription indicates that transcription activity by itself should not be used as evidence that a genome region is under selection. Moreover, pseudogenes (i.e. non-functional gene remnants) can be expressed (Zou et al., 2009; Pei et al., 2012), indicating that transcriptional activity may persist when biological function has been lost. Given these considerations, the selected effect definition of function is frequently considered as more appropriate within biological contexts (Amundson and Lauder, 1994; Graur et al., 2013; Doolittle et al., 2014).

Given the selected effect definition, sequence conservation, particularly over long time periods, represents strong evidence for functionality. Sequence conservation over short time periods, however, may be due to insufficient time for mutations to accumulate, rather than selective pressure. By contrast, lack of observable conservation does not necessarily indicate a lack of function, as this can be due to weak or positive evolutionary selection (Pang et al., 2006; Ponting, 2017). Thus, it is critical to ask not only whether a sequence is conserved, but how informative the presence or absence of sequence conservation is for the potential functionality of a sequence. As a result, it is unclear whether the lack of sequence conservation among most ITRs in *Arabidopsis thaliana*, for example (Moghe et al., 2013), indicates that they are primarily non-functional. In addition to sequence conservation, duplication histories could prove informative to sequence functionality, as gene duplications represented critical evolutionary events that played a fundamental role in shaping the functional content of genomes (Cui et al., 2006; Soltis et al., 2009; Rutter et al., 2012). However, the duplication histories of ITRs are unknown. Instead of relying on a single line of evidence to define functional sequences, such as sequence

conservation, an approach that integrates genetic (i.e. phenotype), evolutionary, and biochemical data has been suggested (Kellis et al., 2014). Based on this framework, recent studies have shown that integrating evolutionary, transcription, epigenetic, and structural characteristics was highly effective at distinguishing between sequences that were under selection and those that were not (Lloyd et al., 2017 [preprint]; Tsai et al., 2017). Uniting the evolutionary histories of ITRs with robust function predictions based on the integration of evolutionary and biochemical signatures could provide valuable insight into the functional content of ITR sequences.

In this study, we investigate the evolutionary dynamics and potential functionality of ITRs in four grass (Poaceae) species: *Oryza sativa* (rice), *Brachypodium distachyon*, *Sorghum bicolor* (sorghum), and *Zea mays* (maize). The four species are oriented phylogenetically as two species pairs (**Fig. 3.1A**), one pair that diverged 15 million years ago (MYA): maize and sorghum (Skendzic et al., 2007; Liu et al., 2014), and another that diverged 47 MYA: rice and *B. distachyon* (Massa et al., 2011). All four species have shared ancient whole genome duplications (WGDs) (Paterson et al., 2004; Tang et al., 2010) and exhibit recent small- and large-sale duplication events (**Fig. 3.1A**) (Swigoňová et al., 2004). We utilized this system to assess sequence and expression conservation of ITRs between species and following duplication. To determine the extent to which evolutionary histories inform the likely functionality of a sequence we also generated function prediction models in rice using established data integration approaches (Lloyd et al., 2017 [preprint]; Tsai et al., 2017). Function prediction models were applied genome-wide to predict candidate functional ITRs, which were then compared among ITRs with varying evolutionary histories.

RESULTS AND DISCUSSION

Identification and classification of Poaceae transcribed regions

To investigate the evolutionary dynamics and potential functionality of intergenic transcripts, we focused on four Poaceae species (**Fig. 3.1A**), each with a set of 10 to 14 developmentally-matched transcriptome datasets. In each of the four species, we identified transcribed regions and classified them according to overlap with exon, intron, and pseudogene annotation. Transcribed regions that did not overlap gene or pseudogene annotation were considered intergenic. To provide an overview of intergenic transcribed regions (ITRs) in the four Poaceae species, we first evaluated their prevalence, putative protein-coding and repetitive element content, and expression characteristics. With regard to the prevalence of ITRs, intergenic transcripts account for only 4-7% of mapped reads and 6-12% of transcribed regions in each species (**Fig. 3.1B**). By contrast, 92-96% of mapped reads overlap an annotated exon or intron (**Fig. 3.1B**), indicating that expression is primarily originating from genic regions. Similarly, 0.4-3.3% of mappable intergenic space (see Materials and Methods) was covered by transcribed regions in each species (**Fig. 3.1C**). Thus, intergenic transcription in the Poaceae is rare relative to genic expression, consistent with previous findings in *Arabidopsis thaliana* (Moghe et al., 2013) and *Homo sapiens* (van Bakel et al., 2011), and only a small fraction of intergenic space is expressed.

Despite the relative scarcity of intergenic transcription, we identified between 7,000 and 16,000 ITRs in each species (**Fig. 3.1B**, right panel). To investigate the content of Poaceae ITRs, we identified ITRs with protein similarity and those that were highly repetitive (see Materials and Methods) and found that the majority of ITRs in each species (54-77%; **Fig. 3.1D**) did not

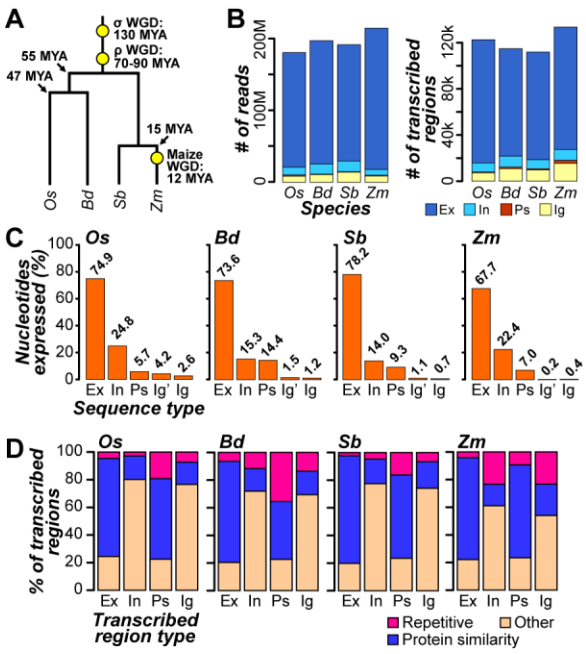


Figure 3.1 Transcriptome content in four Poaceae species. (A) Phylogenetic relationships between the four species. Whole genome duplication (WGD) events are marked with yellow circles. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize; Ex: exon, In: intron, Ps: pseudogene, Ig: intergenic. MYA: millions of years ago. (B) Number of reads mapping to (left panel) and transcribed regions overlapping (right panel) exon (Ex; dark blue), intron (In; cyan), pseudogene (Ps; red), and intergenic (Ig; yellow) genome regions. (C) Percent of nucleotides annotated as exon, intron, pseudogene, and intergenic overlapped by transcribed regions. Ig' represents the proportion of intergenic space covered by transcribed regions that also overlap genic or pseudogenic regions. (D) Percent of transcribed regions that were classified as highly repetitive (pink), with protein similarity (blue), or neither of these (Other; tan).

exhibit hallmarks of either protein-coding or repetitive sequences. Thus, ITRs primarily represent non-protein-coding sequences. However, only 56-94 ITRs in each species (<1%) contained canonical RNA gene domains, indicating that few ITRs may be functioning as microRNAs (miRNAs), small nucleolar RNAs, or small nuclear RNAs.

Given the lack of sequence similarity to known protein-coding or RNA gene regions, ITRs could represent the noisy transcription of random intergenic sequences. We next assessed whether the expression properties of ITRs are consistent with those expected of noisy transcripts. Compared to transcribed regions that overlap exons (ETRs), ITRs in all four species are shorter (Supplemental Fig. 3.1A; Mann Whitney U tests, all p<7e-105) and expressed at lower levels (Supplemental Fig. 3.1B; U tests, all p<8e-311). Further, 49-60% of ITRs in each species are expressed in a single tissue (Supplemental Fig. 3.1C), compared to 10-17% of ETRs (Fisher's Exact Tests (FETs), all p<3e-10). These characteristics are consistent with descriptions of noisy transcripts as short and abortive (Struhl, 2007). However, we found that 76-88% of intergenic transcribed regions were reproducible across replicate leaf transcriptome datasets (Supplemental Fig. 3.1D). While these proportions are lower than those of ETRs (93-95%; FET, all p<3e-10), they far exceed the expected proportions if transcripts were randomly distributed across mappable intergenic space (expected proportions: 2-21%; all p<7e-10). This is indicative of the presence of hotspots in intergenic space where transcription is likely to originate, either due to the presence of truly functional sequences or spurious regulatory signals that increase the likelihood of noisy transcription occurring within a region.

Intergenic transcripts have also been suggested to be associated with nearby genes (van Bakel et al., 2010), either as unannotated exon extensions or products of run-on transcription. We expected that ITRs would be disproportionately originating from the active chromatin regions

surrounding genes, however ITRs were further from genes when compared to a random background (Supplemental Fig. 3.2A). Nevertheless, at least 70% of ITRs in rice, B. distachyon, and sorghum and 35% of ITRs in maize were within the 95th percentile of intron lengths to neighboring genes (Supplemental Fig. 3.2B), indicating that a substantial proportion of ITRs could potentially represent unannotated extensions of genes. Consistent with this notion, ITRs that were within ~500 nucleotides to a gene exhibited increased expression correlation with neighboring ETRs compared to those that are further away (Supplemental Fig. 3.2C). While this could suggest that ITRs nearby genes represent gene extensions, a similar pattern is observed among transcribed regions that overlap pseudogenes. Thus, this may also be explained by the regulation of genes influencing the expression patterns of nearby, but unrelated, transcripts. Overall, we find 43,301 ITRs across four Poaceae species. These ITRs are primarily not protein-coding sequences and their expression characteristics are consistent with noisy expression. To further evaluate their characteristics and potential functionality, we next investigate the evolutionary histories of ITRs both between species and following duplication.

Cross-species sequence conservation of ITRs

Sequence conservation due to selective pressure is a hallmark frequently exhibited by functional genome regions. However, among ITRs in rice and *B. distachyon*, fewer than 15% were conserved across species (**Fig. 3.2A**), conservation rates that are lower than those of exon, intron, and pseudogene transcribed regions (FET, all p<2e-9). Nevertheless, randomly-selected, unexpressed intergenic regions in both species were conserved at lower rates compared to ITRs (<7%; FET, both p<8e-10). Similar patterns were observed among maize and sorghum transcribed regions, except that an increased proportion of sequences were conserved overall (**Fig. 3.2A**). In particular, 2-3 times as many ITRs in maize and sorghum were conserved

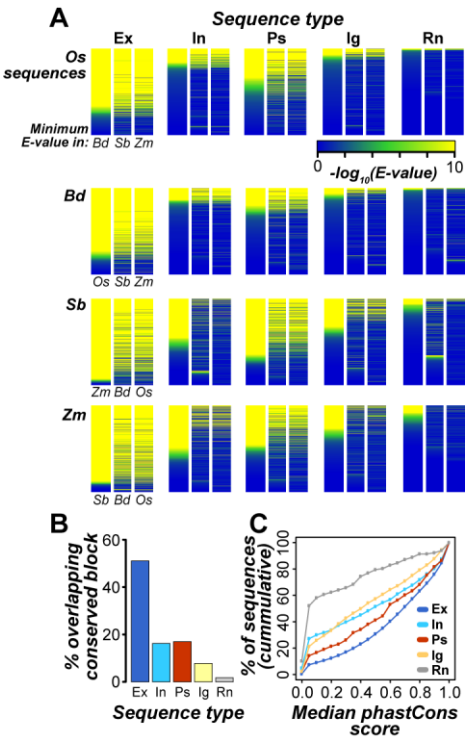


Figure 3.2 Sequence conservation of transcribed regions. (A) Heatmaps of cross-species sequence similarity, where colors indicate the -$\log_{10}$(E-value) for best match within a species. Columns were ordered by evolutionary distance (see **Fig. 3.1A**) then by smaller to larger genome size. Sequence type and species abbreviations are the same as in **Fig. 3.1**. (B) Percent of sequences that overlap a block of conserved nucleotides that are present in all four species. (C) Cumulative percentages of sequences that have at most a median phastCons score indicated on the x-axis. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize; Ex: exon, In: intron, Ps: pseudogene, Ig: intergenic.

compared to those in rice and *B. distachyon* (FET, all p<2e-9), a pattern that is likely due to a more recent divergence between maize and sorghum (**Fig. 3.1A**) (Skendzic et al., 2007; Massa et al., 2011; Liu et al., 2014). This highlights that sequence conservation is not always a result of selection, but instead can be due to insufficient time for sequences to mutate beyond recognition. As a result, reliance on sequence conservation among closely-related lineages to detect functional sequences may have significant false positive error rates. We also note that ITRs with protein similarity in all species have higher rates of conservation that those without (Supplemental Table 3.1; FET, all p<6e-5). Nevertheless, ITRs that lack protein similarity were conserved more frequently than random intergenic sequences (FET, all p<9e-10), suggesting that a subset of ITRs may be functioning at the RNA level (Mercer et al., 2009).

Given that sequence conservation is not always be due to a history of selection, we next investigated potential selective pressure acting on ITRs by calculating phastCons scores (Siepel et al., 2005) within blocks of conserved nucleotides that were present in all four species (see Materials and Methods). As conserved nucleotide blocks (CNBs) and phastCons scores were generated in relation to the rice genome, we focused solely on rice sequences for this analysis. A lower proportion of rice ITRs overlapped CNBs (7%) compared to ETRs (51%; FET, p<2e-9; **Fig. 3.2B**). However, only 2% of random intergenic regions overlapped CNBs (FET, p<4e-10), providing additional evidence that ITRs are more frequently conserved than unexpressed intergenic regions. Although sequences within CNBs were present in all four Poaceae species, we observed significant differences in the phastCons scores among ETRs, ITRs, and random intergenic sequences. ETRs had higher average phastCons scores (median=0.69) than ITRs (median=0.40; U test, p<9e-43), and ITRs were more strongly conserved than random intergenic sequences (median=0.03; U test, p<3e-10). Overall, we find that a minority of ITRs are

conserved across species and are therefore considered lineage specific. However, the proportion of conserved ITRs exceeds that of random intergenic regions, suggesting a subset of ITRs likely represent functional sequences that are under selection.

Cross-species expression conservation of ITRs

Conservation of expression may offer additional information into the likely functionality of a sequence compared to sequence conservation alone. For ITRs that are conserved in sequence across species, we evaluated whether expression was also conserved. For this analysis, we focused on two sets of cross-species homologs: transcribed regions overlapping CNBs described in the previous section and pairs of orthologous sequences present in cross-species syntenic gene blocks (see Materials and Methods). First, we analyzed the expression states of CNBs that were exonic (n=13,616), intronic (n=125), or intergenic (n=525) in all four species. The majority of intergenic CNBs were not expressed in any species (85%; **Fig. 3.3A**). Among 79 intergenic CNBs that had evidence of expression, transcription was restricted to a single species in 57 cases (78%), a pattern that is more similar to conserved intron sequences than exons (**Fig. 3.3A**). Similarly, 18-44% of syntenic ITR orthologs exhibited conserved expression, rates that are lower than those of syntenic ETR orthologs (fig 3B; FET, all p<2e-10). Thus, ITRs with sequence conservation often do not exhibit expression conservation. Among ITRs, approximately twice as many syntenic orthologs between rice and *B. distachyon* were expressed (~40%) compared to those between maize and sorghum (~20%; fig 3B; FET, all p<6e-4), indicating that sequences maintained over longer periods of time are more likely to both be expressed. This could suggest that expression is lost more quickly than sequence similarity and, as a result, the presence of expression conservation may prove useful in classifying likely-functional sequences, particularly between closely-related species.

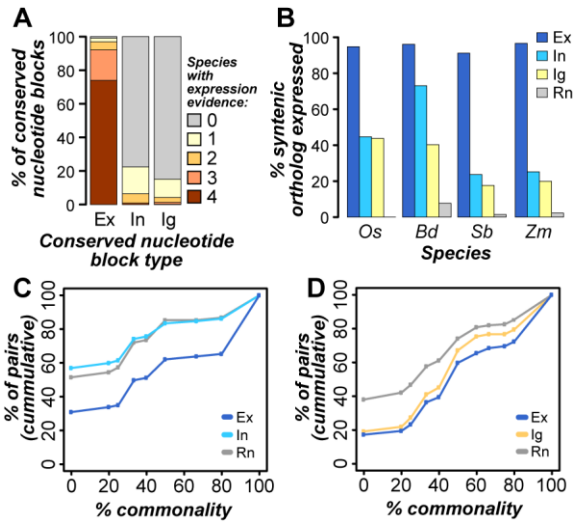


Figure 3.3 Expression conservation of transcribed regions. (A) Percentages of exon (Ex), intron (In), and intergenic (Ig) conserved nucleotides blocks that have evidence of expression from 0, 1, 2, 3, or all 4 species in our analysis. (B) Percentages of syntenic orthologs with expression evidence. Syntenic orthologs were identified by conservation across pairs of cross-species syntenic blocks. (C,D) Commonality in tissue expression between reciprocal pairs of cross-species homologous exons (Ex), introns (In; panel C), and intergenic (Ig; panel D) transcribed regions and randomly-generated expression vectors (Rn). Exon transcript pairs represent a subsample of all exons with breadths of expression that match either intron (C) or intergenic (D) transcript pairs. Similarly, random expression vectors were generated with breadths that matched the distribution from intron or intergenic pairs. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize.

If orthologous sequences with conserved expression are both functional, they may perform similar roles in both species and show similar expression patterns across tissues. To evaluate this possibility, we first investigated similarity in tissue expression among pairs of transcribed, orthologous intron sequences and found that the proportion of expressed tissues in common between intron orthologs (mean % commonality=27%) was no different than random (mean=26%; **Fig. 3.3C**; U test, p=0.33) and lower than that of orthologous ETR pairs (mean=49%; U test, p<2e-18). This may suggest that a substantial proportion of conserved intron expression is due to capture of transcripts prior to splicing or misregulated splicing events from neighboring exon regions, rather than selection. By contrast, orthologous ITRs had similarity in tissue expression (mean % commonality=48%) that exceeded random (mean=36%; **Fig. 3.3D**; U test, p<6e-6), but was lower than that for orthologous ETRs (mean=53%; U test, p=0.04). However, the difference in % commonality between orthologous ITRs and ETRs was small. ITRs with cross-species expression conservation may frequently represent functional sequences performing similar roles in different species. Nevertheless, few ITRs exhibit sequence or expression conservation across species, suggesting that most either function in a species-specific manner, evolve too quickly for orthologous sequences to be detected, or are products of noisy transcription.

Post-duplication sequence and expression conservation of ITRs

Plant genomes harbor a rich history of large- and small-scale duplication events that have contributed to the adaptive evolution of these lineages (Cui et al., 2006; Rizzon et al., 2006; Hanada et al., 2008; Soltis et al., 2009). Thus, we next asked whether ITRs are present as duplicates, and if so, whether ITR duplicates tend to be retained over time. We first assessed the prevalence of ITR duplicates and found that 40-50% of ETRs, ITRs, and random intergenic

sequences were duplicated in rice, *B. distachyon*, and sorghum and 70-80% were duplicated in maize (Supplemental Fig. 3.3). Although duplication rates between sequence types could be significantly different (see Supplemental Fig. 3.3), almost all differences were less than 5%, indicating that the presence of a duplicate is not informative to whether a sequence is likely part of an annotated gene. We next estimated the timing of duplication using per-base substitution rates (*K*) and found that ITR duplicates in all species were generated from more recent duplication events compared to those of ETRs (**Fig. 3.4A**; U tests, all p<6e-127). Thus, duplication patterns are distinct between ITRs and ETRs and could be associated with the functionality of a sequence.

To further investigate the duplication patterns of ITRs, we evaluated putative duplicate retention of ITRs present in WGD-associated syntenic genome regions (see Materials and Methods; Supplemental Fig. 3.4). Among 6,531 maize ITRs present in syntenic blocks associated with the 12 million-year-old (MYO) maize WGD (**Fig. 3.1A**; Supplemental Fig. 3.4), 12% exhibit a retained syntenic duplicate. Fewer retained duplicates are identified for the 5,783 ITRs from all four species present in syntenic blocks associated with the >70MYO ρ / σ WGDs (**Fig. 3.1A**; Supplemental Fig. 3.4), as only 63 exhibited a retained syntenic duplicate (1.1%), including 26 duplicates that were also intergenic (0.5%). We considered that ITRs may frequently represent RNA genes and sequence similarity over >70 million years could be difficult to detect. However, in rice 11 of 18 tRNA genes (61%) and 3 of 7 miRNA genes (43%) had identifiable ρ / σ WGD-associated duplicates, retention rates that surpass that of rice ITRs (2.2%; both p<5e-5). This suggests ITRs exhibit low duplicate retention rates over time. However, we cannot rule out the possibility that many ITRs present in WGD-associated syntenic regions arose after WGD events. Additionally, ITRs may have been disproportionately

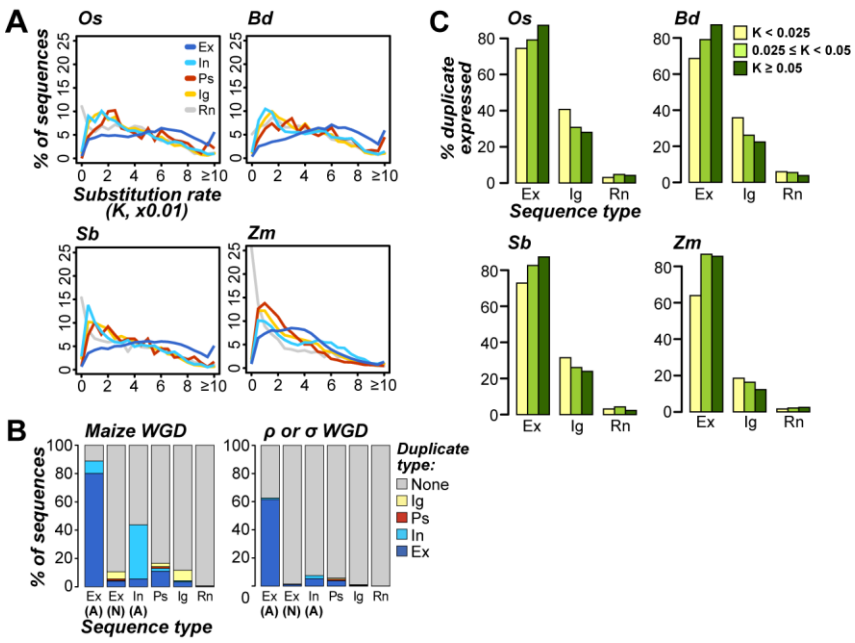


Figure 3.4 Duplication characteristics of transcribed regions. (A) Distributions of per-base substitution rates (*K*) between a sequence and its top within-species BLASTN match based on a generalized time reversible model. (B) Syntenic duplicates from the maize whole genome duplication (WGD; left panel) and ρ or σ WGDs (right panel). 'None' indicates that no syntenic duplicate was identified. Exon transcribed regions (Ex, x-axis) were categorized by whether they overlapped genes anchoring syntenic blocks, '(A)', and those that were between anchor genes, '(N)'. Only intron transcribed regions (In, x-axis) that overlapped anchor genes are shown. (C) Percent of expressed duplicates among sequence types binned by duplication time, which was estimated using *K*. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize; Ps: pseudogene, Ig: intergenic.

transposed into syntenic blocks or their duplicates transposed out, resulting in a lack of observable synteny. Nevertheless, ITRs feature WGD-related duplicates at greater rates than those of random intergenic regions (**Fig. 3.4B**, FET, both $p<0.002$). Thus, ITRs and their duplicates from WGD events could be under selective maintenance and represent functional sequences. Consistent with this notion, 92% of maize ITRs with a retained syntenic duplicate from the recent maize WGD event were conserved across species, compared to 46% of those without an identifiable duplicate (FET, $p<3e-10$). Similarly, 26 of 26 ITRs in all four species with an intergenic duplicate retained from the ρ / σ WGD events were conserved across species. Thus, duplicate retention over long periods is a likely a strong indicator of sequence functionality.

Last, we assessed whether ITR expression was conserved following duplication. Among duplicates of ITRs in each species, 18-34% were expressed, compared at least 79% of ETR duplicates (FET, all $p<6e-10$). In addition, recent (i.e. lower K) ITR duplicates are more likely to be expressed than older duplicates (**Fig. 3.4C**; U tests, all $p<3e-5$). These results might suggest highly dynamic expression among ITRs, with frequent expression state gains and losses. However, fewer duplicates of random intergenic sequences were expressed (2-7% in each species; FET, all $p<4e-10$) compared to ITR duplicates, indicating that the presence of expression between an ITR and its duplicate is likely due to the conservation of an ancestral expression state, rather than two independent gains of expression. Together with high reproducibility of ITR expression across replicate transcriptome datasets (Supplemental Fig. 3.1D), this indicates that intergenic expression is driven in part by the presence of particular regulatory signals (e.g. *cis*-regulatory elements or chromatin state) that may persist following duplication. Overall, the duplication patterns among ITRs indicate that few ITRs exhibit long-

term duplicate retention and may feature highly dynamic expression states, characteristics that could be indicative of sequences that are generally not under selection. Nevertheless, we find rare examples of ITRs with ancient retained duplicates that could represent functional sequences. To more deeply explore the functionality of ITR sequences, particularly those that lack long-term sequence conservation or duplicate retention, we utilized a data integration approach (Lloyd et al., 2017 [preprint]; Tsai et al., 2017) that jointly considers evolutionary and biochemical characteristics of a sequence to classify its functionality.

Predicting the functionality of rice transcribed regions

Sequence and expression conservation over long periods of time and following duplication provide glimpses into the potential functionality of ITR sequences. To provide a more robust estimation of the functionality of transcribed regions, we next utilized a data integration approach that considers the evolutionary and biochemical characteristics of known functional and non-functional sequences (Lloyd et al., 2017 [preprint]; Tsai et al., 2017). This approach consists of four steps: establishing a set of benchmark functional and non-functional sequences, identifying characteristics associated with benchmark sequences, integrating these characteristics via statistical learning methods to generate function prediction models, and applying function prediction models to all transcribed regions, including ITRs. Due to data availability, we focused solely on rice sequences for this approach. We first established two sets of benchmark functional sequences: 513 transcribed regions that overlap exons of genes with documented loss-of-function phenotypes (referred to as Phen-TRs) and 19 transcribed regions that overlapped high-confidence miRNA gene annotations (referred to as miRNA-TRs). Benchmark non-functional sequences consisted of 743 transcribed regions that overlap pseudogenes (referred to as Pseu-TRs), the disabled remnants of ancient genes. We next

identified a set of 44 evolutionary and biochemical features representing five categories: transcription activity (n=12), sequence conservation (n=3), DNA methylation (n=16), histone marks (n=12), and nucleosome occupancy (n=1) (Supplemental Table 3.2; see Materials and Methods). A wide variety of features in multiple categories were informative for distinguishing between benchmark functional and non-functional sequences (Supplemental Table 3.2) and we moved forward with data integration methods to consider all features in combination.

We utilized the random forest machine learning algorithm to jointly consider all 44 evolutionary and biochemical features and generate function prediction models. Two models were established: a binary model trained on Phen-TR and Pseu-TR sequences and a three-class model trained on Phen-TR, miRNA-TR, and Pseu-TR sequences. Models were generated under cross-validation; independent sets of sequences were used for training and testing (see Materials and Methods). The resulting binary prediction model was highly capable of distinguishing between Phen-TRs and Pseu-TRs (**Fig. 3.5A**; Supplemental Fig. 3.5A; AUC-ROC =0.92). We utilized prediction scores from the random forest classifier and a threshold identified by maximizing F-measure to classify sequences as likely functional or non-functional. Based on this approach, we correctly classified 85% of Phen-TRs as functional and 86% of pseudogenes as non-functional. However, only 3 of 19 miRNA-TR sequences (21%) were predicted as functional, indicating substantial false negative rates for RNA gene classification based on the binary model. By comparison, 15 of 19 miRNA-TRs (79%) were classified as either miRNA-TRs or Phen-TRs based on the three-class prediction model (**Fig. 3.5B**; Supplemental Fig. 3.5B; see Materials and Methods), indicating that the three-class model is better suited to predicting functionality among RNA genes. Similarly, 82% of Phen-TRs were predicted as functional based

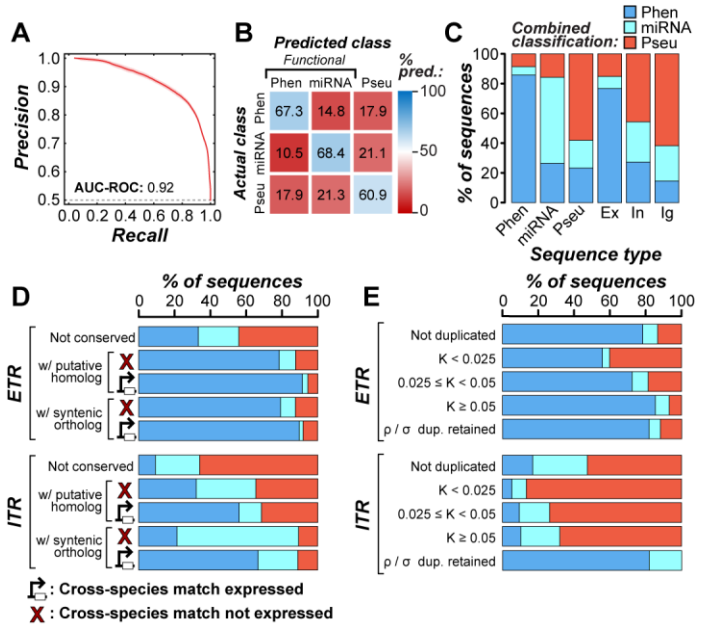


Figure 3.5 Function predictions of transcribed regions. (A) Precision-recall curve of the best-performing random forest classifier based on a binary model distinguishing between transcribed regions that overlap exons of phenotype genes and pseudogenes. (B) Confusion matrix heatmap of actual and predicted classes from a 3-class model trained with transcribed regions that overlap exons of phenotype genes (Phen), high-confidence miRNAs (miRNA), and pseudogenes (Pseu). Predictions as Phen or miRNA were considered putatively functional. (C) Distributions of function predictions among sequence types based on the combination of binary (A) and 3-class (B) prediction models. Combined predictions were based on a priority system: Phen > miRNA > Pseu. (D,E) Distributions of function predictions among conserved (D) and duplicated (E) sequences. Colors represent the same classifications as in (C). (D) Sequences with putative homologs were those with any significant cross-species similarity, while those with syntenic orthologs exhibited significant conservation across a cross-species syntenic gene block. (E) Duplicated ETRs and ITRs were separated according to duplication timing, which was estimated

Figure 3.5 (cont'd)

using per-base substitution rates (*K*). Duplicates from the ρ and σ whole genome duplications represent examples of ancient duplicate retention. Ex: exon, In: intron, Ig: intergenic, ETR: exon transcribed region, ITR: intergenic transcribed region.

on the three-class model. However, only 61% of pseudogenes were predicted as non-functional (compared to 85% based on the binary model). While a small subset of pseudogenes may represent truly functional sequences (Zou et al., 2009; Poliseno et al., 2010; Karreth et al., 2015), this result likely reflects the difficultly in distinguishing between RNA genes and pseudogenes using the feature set established here. We established a final set of combined predictions by merging the classifications from the binary and three-class function prediction models (**Fig. 3.5C**).

We next applied the binary and three-class function prediction models to all transcribed regions within the rice genome: ETRs that did not overlap a known phenotype gene, intron transcribed regions, and ITRs. Among ETRs, 85% were predicted as functional, either as Phen-TRs (77%) or miRNA-TRs (8%). The 15% of ETRs that were predicted as non-functional were enriched for short ETRs (Supplemental Fig. 3.6A; U test; p<9e-83) and those that had minor overlap with an exon (U test, p<2e-39; Supplemental Fig. 3.6B,C). We considered that intron transcribed regions would exhibit similarity to RNA genes, given their location within annotated genes (i.e. potentially-functional genome regions) and lack of protein-coding potential. However, only 27% of intron transcribed regions were classified as miRNA-TRs and an additional 17% were classified as Phen-TRs, indicating that the function prediction models can distinguish between transcribed regions overlapping sequences that encode a final gene product and those that are spliced out. Last, we applied the prediction models to ITRs, and found that 62% were predicted as non-functional. Consistent with previous results (Lloyd et al., 2017 [preprint]; Tsai et al., 2017), the lack of similarity among ITRs to benchmark functional sequences suggests that the majority are the result of transcriptional noise.

Comparison of function predictions and evolutionary histories

With sets of ITRs predicted as functional and non-functional, we next investigated the extent to which the evolutionary histories of a sequence inform sequence functionality. As expected, ITR sequences that were conserved across species were more frequently predicted as functional compared to those that were not (**Fig. 3.5D**; FET, p<5e-10), which is due in part to conservation features used in model training. However, despite at least of 47 million years of sequence conservation, 34% of conserved ITRs that were predicted as non-functional. On the other hand, 35% of ITRs without sequence conservation were predicted as functional, suggesting a subset of ITRs may be performing species-specific roles. Overall, relying on sequence conservation alone results in significant false positive and false negative function predictions among ITRs.

We next investigated whether the presence of expression conservation in addition to sequence conservation more accurately predicted functional sequences. Surprisingly, a similar proportion of ITRs with expression conservation and sequence conservation only were predicted as functional (**Fig. 3.5D**; FET, both p>0.37). However, ITRs with expression conservation were more likely to be classified as Phen-TRs, rather than miRNA-TRs (FET, both p<0.005). This pattern could be explained by the miRNA-TR sequence set containing false positive functional benchmarks, as the miRNA annotations we used lack strong functional evidence provided by genetic analysis. Consequently, sequences predicted as miRNA-TR may be more frequently truly non-functional compared to those predicted as Phen-TRs (**Fig. 3.5B**; Supplemental Fig. 3.5B). Thus, the increased proportion of Phen-TR classifications among ITRs with expression conservation suggests they are enriched for functional sequences relative to ITRs that only exhibit sequence conservation. Of particular note, 89% of ITRs with syntenic, expressed

orthologs were predicted as functional, with 67% classified as Phen-TRs, suggesting that expression conservation of a high-confidence ortholog provides strong evidence for sequence functionality among ITRs.

Last, we investigated the relationship between function predictions and duplication status. We found that 47% of single-copy ITRs were predicted as functional compared to 24% of those that were duplicated (**Fig. 3.5E**; FET, p<5e-10). A similar pattern is observed for ETRs (**Fig. 3.5E**), suggesting that a greater proportion of duplicate sequences may be non-functional and undergoing decay. Among duplicated sequences, ITRs with older duplicates (i.e. higher K) are more frequently predicted as functional compared to those with more recent duplicates (**Fig. 3.5E**; U test, p<6e-11), indicating that the retention of a duplicate over time is at partially informative for the functionality of a sequence. This notion is supported by the 11 ITRs with a retained duplicate from the >70 MYO ρ or σ WGD events, as all were predicted as functional (**Fig. 3.5E**). Thus, combining function predictions with evaluation of evolutionary histories confirms that retention of ancient duplicates, as well as cross-species sequences and expression conservation, provide strong evidence that a sequence is likely to be functional. However, the overall number of ITRs that exhibit sequence and expression conservation is low.

CONCLUSION

We provide a comprehensive, multi-species investigation of the potential functionality of intergenic transcribed regions (ITRs) through evaluation of the cross-species and post-duplication evolutionary histories of these sequences in four grass (Poaceae) species. In summary, ITRs are primarily species-specific and the lack of conserved transcription across species indicates that expression among ITRs may be a recent evolutionary event. This is further reflected by the lack of transcription evidence among ITR duplicates, suggesting that the expression of ITRs is frequently gained and lost. We also found very few examples of ITRs with ancient retained duplicates, suggesting there may be little adaptive potential through neo- or sub-functionalization of ITR sequences. Overall, the evolutionary dynamics explored here are consistent with the notion that intergenic expression primarily represents transcriptional noise.

Robust function prediction models were established in rice based on 44 evolutionary and biochemical characteristics of transcribed regions that overlap phenotype genes, high-confidence miRNAs, or pseudogenes. These models distinguished between benchmark functional and non-functional with high accuracy. Based on model predictions, 2,754 ITRs in rice (38%) were classified as functional. However, high rates of false positive predictions (FPR=42%) suggest this is likely an overestimate of the proportion of functional ITRs. Function predictions could be improved in part by identification of a stronger set of benchmark RNA genes, as the lack of RNA genes with genetic evidence of function required us to use unconfirmed miRNA gene annotations. Using a similar data integration framework, ~40% of ITRs in *Arabidopsis thaliana* were also predicted as functional (Lloyd et al., 2017 [preprint]), and thus this percentage may represent a general baseline for the proportion of candidate functional ITRs in any species.

However, both rice and *A. thaliana* have relatively small genomes among plants and it will be interesting to see whether a species with a much larger genome (e.g. maize) contains more or fewer ITRs that are likely functional.

Function predictions confirmed that ITRs with both sequence and expression conservation and those with ancient retained duplicates are likely to represent functional sequences. Prediction models also classified 35% of rice ITRs that lack sequence conservation as functional. A similar pattern can be seen in *A. thaliana*, as a greater proportion of ITRs are predicted as functional (Lloyd et al., 2017 [preprint]) than are conserved across species (Moghe et al., 2013). These patterns indicate that it may not be uncommon for ITRs to function in species-specific roles. The data integration framework applied here and described in previous studies (Lloyd et al., 2017 [preprint]; Tsai et al., 2017) can identify likely-functional sequences when sequence conservation cannot be effectively utilized, e.g. species-specific function or weak selective pressure (Ponting, 2017). Ultimately, the majority of ITRs lack extensive cross-species or post-duplication sequence conservation and similarity to benchmark functional sequences, and thus we conclude that intergenic transcription primarily represents transcriptional noise.

METHODS

Gene, pseudogene, and random intergenic annotation

The MAKER-P (r1065) genome annotation pipeline was used to reannotate the *Oryza sativa* (rice) Nipponbare (IGRSP-1.0 v7), *Brachypodium distachyon* (v1.2) and *Sorghum bicolor* (sorghum; v2.1) genome assemblies as previously described (Cantarel et al., 2008; Campbell et al., 2014a) . Repeats were masked using default parameters in RepeatMasker (v 4.0.3) for the sorghum and *B. distachyon* genomes, and a custom repeat library was created for rice using a method described previously (Campbell et al. 2014). RepeatMasker was run within the MAKER pipeline to mask repetitive elements. To aid in gene prediction, expressed sequence tag (EST) evidence was provided by transcriptome assemblies generated from selected publically available data from the National Center for Biotechnology Information Sequence Read Archive (NCBI-SRA) (Supplemental Table 3.3) using Trinity (version 2014) with a minimum contig length of 150 bp and the Jaccard clip option (Haas et al., 2013). Protein evidence was provided by the SwissProt plant protein dataset (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_plants.dat.gz), with the rice, sorghum, or *B. distachyon* protein sequences removed. For *Zea mays* (maize; v. 3.21), we utilized the MAKER-P gene annotation described by Law et al. (2015). Pseudogenes were identified in each species using a pipeline similar to that described by Zou et al. (Zou et al., 2009). We considered genome regions at least 40 contiguous ambiguous nucleotides (Ns) as likely-unmappable, a length that represented the size of reads in RNA-sequencing (RNA-seq) datasets used in our analyses (see below).The length of likely-unmappable were excluded when calculating the size of total exon, intron, pseudogene, and intergenic space in **Fig. 3.1B**. Random

sets of intergenic coordinates with equal length distributions as intergenic transcribed regions (ITRs, see below) were identified in each species. Random intergenic coordinates were sampled so that they did not overlap one another, transcribed regions, or likely-unmappable genome regions (as described above). We further filtered random intergenic coordinates to remove those that contained an ambiguous nucleotide, a step that removed 4-9% of random coordinates in each species.

Identification and classification of transcribed regions

Multiple developmentally-matched RNA-seq datasets from rice, *B. distachyon*, sorghum, and maize (n = 11, 11, 10, and 14 respectively) were retrieved from the Sequence Read Archive at the National Center for Biotechnology Information (Supplemental Table 3.4) (Davidson et al., 2011; Davidson et al., 2012). Reads were trimmed of low-scoring ends and Illumina adapter sequences using Trimmomatic v.0.33 (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20). Reads ≥20 nucleotides in length and with an average Phred score ≥20 were mapped to associated genomes using TopHat v.2.1.0 (-i 5000; -I 50000; all other parameters default) (Kim et al., 2013). Transcribed region coordinates were identified by assembling unique-mapped reads with Cufflinks v.2.2.1 (-min-intron-length 5000; -max-intron-length 50000; -m 150; --no-effective-length-correction) (Trapnell et al., 2010) while providing an associated genome sequence via the -b flag to correct for sequence-specific biases. Transcribed regions from across datasets that overlapped with one another by at least 1 nucleotide were merged. Merged transcribed regions were classified based on overlap with annotated exon, intron, and pseudogene sequences through a priority system: exon > intron > pseudogene (e.g. a transcribed region that overlapped both an exon and an intron was classified as exonic). Transcribed regions that did not overlap with any gene or pseudogene annotation were classified as intergenic.

Transcribed regions were further categorized as likely protein coding or repetitive. Likely protein-coding transcribed regions were defined as those that contained a protein domain from Pfam v. 31.0 (Finn et al., 2016) in any translated frame or had significant translated sequence similarity to a plant protein annotated in Phytozome v.10 (BLAST v. 2.2.26; BLASTX E-value < 1e-05) (Goodstein et al., 2012). By contrast, the presence of canonical RNA gene domains were identified within ITRs by searching for significant matches (E<1e-5) with Rfam domain covariance models (v.12) (Nawrocki et al., 2015) using Infernal v.1.1 (Nawrocki and Eddy, 2013)

A set of likely repetitive transcribed regions were defined based on the presence of repeat-associated Pfam protein domains or high numbers of duplicate sequences. To identify repeat-associated protein domains and set duplicate thresholds to call sequences as highly repetitive, a set of benchmark interspersed repetitive elements were identified in each unmasked genome using RepeatMasker v.4.0.5 (-nolow -norna -qq) with the default RepeatMasker repeat library and a custom library of rice repeats generated using previously described methods (Campbell et al. 2014). Transcribed regions that contained a Pfam v.31.0 protein domain that was significantly enriched among interspersed repeats relative to exon transcribed regions (Fisher's exact test; adjusted $p<0.05$; Benjamini-Hochberg procedure) (Benjamini and Hochberg, 1995) were considered likely-repetitive. The number of duplicate sequences required to call a sequence as highly-repetitive was calculated based on the duplicate counts of long terminal repeats (a subset of the interspersed repetitive elements identified by RepeatMasker) and transcribed regions that overlapped exons, through F-measure maximization. Resulting duplicate count thresholds to consider a transcribed region as repetitive were 10, 15, 33, and 202 for rice, *B. distachyon*, sorghum, and maize, respectively. For all duplication-related analysis, we

excluded sequences that were classified as repetitive, as they were considered duplicated by definition.

Sequence and expression conservation

Significant sequence matches were identified using BLAST searches (BLAST v. 2.2.26; BLASTN E-value < 1e-05) by searching nucleotide sequences of transcribed regions against repeat-masked whole genome sequences. BLAST searches were performed within-species to identify duplicates and cross-species to identify putative homologs. Per base substitution rates (*K*) between within-species query and match sequences were calculated with PHAST (Hubisz et al., 2011) using a generalized time reversible model after aligning matched regions with EMBOSS-Needle v.6.5.7.0. Blocks of conserved nucleotides present across all four species were identified using the LASTZ/MULTIZ paradigm with rice as the target genome (Blanchette et al., 2004; Harris, 2007; Hupalo and Kern, 2013) (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto). Among all conserved nucleotide blocks (CNBs; n=60,801), we identified those that were exonic (n=13,616), intronic (n=125), or intergenic (n=525) in all four species and were only overlapped only by transcribed regions of the associated type (e.g. a conserved nucleotide block that was intergenic in all four species and was only overlapped by ITRs). phastCons scores were calculated for each rice nucleotide within a conserved nucleotide block (Siepel et al., 2005), with conserved and non-conserved states estimated using the --estimate-trees option.

Expression conservation was assessed by determining whether a cross-species or within-species match was overlapped by a transcribed region (as described above). Reciprocal matches between exon, intron, and intergenic transcribed regions across CNBs or cross-species syntenic gene blocks (see below) were included in percent commonality calculations (**Fig. 3.3C,D**), which

was calculated as the number of expressed tissues in common between a pair of transcribed regions divided by the total number of expressed tissues. The five tissues that were in common among all four species were used: embryo, endosperm, seed, leaf, and anther. When replicate datasets were available, transcription evidence in a single dataset was required to consider the tissue as expressed. Similarly, expression in either seed datasets – 5 days after pollination (DAP) or 10 DAP – were required. Expected percent commonality is affected by the expression breadth of two transcripts. To control for expected percent commonality, exon percent commonality (**Fig. 3.3C,D**) was calculated using exon transcribed region pairs with expression breadth that matched those of conserved intergenic or intron transcribed regions. For each pair of conserved intron and intergenic transcribed regions, two pairs of exon transcribed regions with the same expression breadths were selected for comparisons. Background expectations of percent commonality (**Fig. 3.3C,D**, Rn) were calculated by randomly selecting tissues to match the expression breadth of intergenic and intron transcribed region pairs. For each intron and intergenic transcribed region pair, 25 random pairs of tissues were selected. The probability that a tissue would be selected for a random set was proportional to how often it appeared among transcribed regions with expression conservation.

Identification of syntenic gene blocks

We identified cross-species and within-species syntenic blocks by identifying of sets of collinear genes using MCScanX v.2 (Wang et al., 2012). Multiple minimum gene pair and maximum gap parameters were tested when generating collinear blocks (Supplemental Table 3.5) and we utilized values of 10 and 10, respectively, for within-species collinear blocks and 10 and 2, respectively, for cross-species blocks. Cross-species blocks were identified between the more-closely-related species pairs: rice/*B. distachyon* and maize/sorghum. The rates of

synonymous substitutions (*Ks*) between homologous protein-coding genes anchoring within-species syntenic blocks were calculated using the yn00 package in the Phylogenetic Analysis by Maximum Likelihood software (Yang, 2007). Syntenic blocks with a median *Ks* ≥0.7 across all anchor genes in the syntenic block in all four species were considered associated with the ρ / σ whole genome duplication (WGD) events (Paterson et al., 2004; Tang et al., 2010), while syntenic blocks in maize the an average Ks < 0.7 were associated with the more recent 12 MYO maize WGD (Supplemental Fig. 3.4) (Swigoňová et al., 2004). An additional 254 gene-pair syntenic block in rice was identified with a median block *Ks* of 0.13, suggesting recent duplication. However, due to the uncertain nature of the origin and timing of the duplication event (Wang et al., 2011), this block was excluded from further analysis. For exon and intron transcribed regions, syntenic duplicates and orthologs were identified as BLASTN matches (E-value<1e-5) that overlapped a homologous anchor gene. For pseudogene transcribed regions, ITRs, and random intergenic sequences, syntenic duplicates and orthologs were BLASTN matches that were present within corresponding block regions circumscribed by homologous anchor genes. To provide annotated RNA gene comparisons, rice tRNA gene sequences marked as reliable were retrieved from tRNAdb (Jühling et al., 2009) and high-confidence rice miRNA gene sequences were retrieved from miRBase (Kozomara and Griffiths-Jones, 2014). To identify the location of tRNA and miRNA gene sequences within the rice genome, we identified perfect alignment between RNA gene sequences and the MSU.v.7 rice genome (BLASTN, 100% identity, full length alignment). We further filtered out tRNA and miRNA sequence alignments that overlapped gene or pseudogene annotation.

Rice function prediction features

Transcription activity features

Twelve transcription-related features were generated for use in rice function prediction models (Supplemental Table 3.2). The first 9 features were FPKM values (referred to as expression level; 'Level' in Supplemental Table 3.2) from each of the 9 tissues represented in the RNA-seq datasets described above. For replicate datasets (leaf and endosperm), expression level was taken as the average FPKM value if a region was expressed in both replicate datasets, and the single FPKM value otherwise. If a transcribed region was not expressed in an RNA-seq dataset, expression level was set to 0. Two additional features were represented by the maximum expression level among all RNA-seq datasets and the median FPKM among datasets where a transcribed region was expressed. The final feature was the expression breadth of a sequence, represented by the total number of tissues in which a sequence was expressed. For the breadth calculations, both seed datasets (5 and 10 DAP) and inflorescence datasets (early and emerging) were considered as a single tissue.

Sequence conservation features

Three sequence-conservation-related features were generated. The first feature was the minimum BLASTN E-value to a sequence in *B. distachyon*, sorghum, or maize. A threshold of 1e-05 was required to consider a match significant. Sequences without a significant cross-species match where given a score of 0. The second sequence-conservation feature was the proportion of a sequence covered by the 4-species conserved nucleotide blocks (described above). The third sequence-conservation feature was the median per-base phastCons score across the proportion of a sequence that overlapped a conserved nucleotide block. If a sequence did not overlap a conserved block the phastCons score was set to 0.

Histone mark and nucleosome occupancy features

Twelve histone mark-related features among 10 histone marks were calculated based on 18 chromatin immunoprecipitation sequencing (ChIP-seq) datasets from NCBI-SRA (Supplemental Table 3.6). Reads were trimmed as described above and mapped to the rice MSU v.7 genome with Bowtie v2.2.5 (default parameters). Spatial Clustering for Identification of ChIP-Enriched Regions v.1.1 (Xu et al., 2014) was used to identify significant ChIP-seq peaks with a non-overlapping window size of 200, a gap parameter of 600, and an effective genome size of 0.68 (Koehler et al., 2011). For datasets with control total histone or protein datasets, a false discovery rate (FDR) ≤0.05 was utilized (see Supplemental Table 3.6). Peaks for histone marks with multiple datasets were merged. The first 10 histone mark-related features were represented by the percent overlap of a sequence with histone mark peaks for each histone marks ('coverage' in Supplemental Table 3.2). The other two features were the number of activation-associated histone marks and repression-associated histone marks with a peak that overlapped a sequence. Eight marks were considered activation-associated and two were considered repression-associated (Supplemental Table 3.6). A single nucleosome occupancy feature was also generated from micrococcal nuclease sequencing (MNase-seq) data. MNase-seq data was generated by Wu et al. (2014) and processed by Liu et al. (2015). The nucleosome occupancy feature was calculated as MNase-seq average read depth across the length of a sequence.

DNA methylation features

Sixteen DNA methylation features were calculated from bisulfite-sequencing (BS-seq) datasets from four tissues: embryo (SRR059000), endosperm (SRR059005), leaf (SRR618545), and panicle (SRR1520042). BS-seq reads were trimmed as described above and processed with Bismark v.3 (default parameters) to identify the number of reads that call cytosines as

methylated and unmethylated in CG, CHG, and CHH (H = A, C, or T) contexts. The first 12 features were represented by the methylation level of a sequence in CG, CHG, and CHH contexts for each of the four tissues. Methylation level of a sequence was calculated as the number of reads that mapped to CG, CHG, or CHH sites that called a cytosine site as methylated divided by the total number of reads mapping to CG, CHG, or CHH sites within the sequence. A minimum of 5 reads across 5 cytosine sites were required to calculate methylation level. Multiple minimum read (range: 1-20) and site (range: 1-10) requirements were tested and found to have little effect on the ability of resulting methylation level features to distinguish between Phen-TR and Pseu-TR sequences (Supplemental Table 3.7). The final four DNA methylation features were represented by whether a sequence exhibited a methylation pattern consistent with gene body methylation (GBM) in each of the four tissues. The GBM pattern is presence of CG methylation and absence of CHG and CHH methylation. Presence or absence of CG, CHG, and CHH states within a sequence were determined by binomial tests of the methylation level of a sequence (as described above) compared to the background methylation level across the whole genome for a given cytosine context. Binomial test *P*-values were corrected for multiple testing using the Benjamini and Hochberg method (FDR $\leq$0.05) (Benjamini and Hochberg, 1995). Sequences that were significantly enriched in CG methylation relative to the genome background and not significantly enriched in CHG and CHH methylation were considered to be gene body methylated.

Machine learning approach

We utilized the random forest algorithm implemented in the Scikit-learn software (Pedregosa et al., 2011) to perform machine learning runs aimed at distinguishing between functional and non-functional transcribed regions in rice. Two sets of benchmark functional

transcribed regions were established: those that overlapped exons of genes with documented loss-of-function phenotypes (referred to as Phen-TRs) (Lloyd et al., 2015; Oellrich et al., 2015) and those that overlapped a set of high-confidence miRNA gene annotations from miRBase (referred to as miRNA-TRs) (Kozomara and Griffiths-Jones, 2014). Benchmark non-functional sequences were represented by transcribed regions that overlapped pseudogene annotation (referred to as Pseu-TRs). Prior to model building, missing data points among features were imputed by randomly-selecting a representative value from another sequence of the same type. For example, an ITR with a missing data point would randomly-select a value for the same feature from another ITR.

Two sets of machine learning-based predictions were generated: a binary prediction trained using Phen-TR and Pseu-TR sequences (**Fig. 3.5A**; Supplemental Fig. 3.5A) and three-class prediction trained using Phen-TR, miRNA-TR, and Pseu-TR sequences (**Fig. 3.5B**; Supplemental Fig. 3.5B). For the binary predictions, 100 balanced datasets that included equal proportions of Phen-TR and Pseu-TR sequences were used for training and 10-fold cross-validation was implemented (i.e. 90% of a dataset was used for training and the held-out 10% used for testing). Parameter sweeps of maximum tree depth (3, 5, 10, and 50) and proportion of random features (10%, 25%, 50%, 75%, square root, and $\log_2$) values were performed, with 3 and 10% providing the highest performance, respectively. For each of the 100 balanced datasets, a prediction score for each transcribed region was generated that was equal to the proportion of 500 random forest trees that predicted the region as Phen-TR; a final prediction score was calculated as the median of all scores. A threshold to predict transcribed regions as likely-functional or non-functional was generated by identifying the threshold with the maximum F-measure, which is calculated as the harmonic mean of precision (proportion of sequences

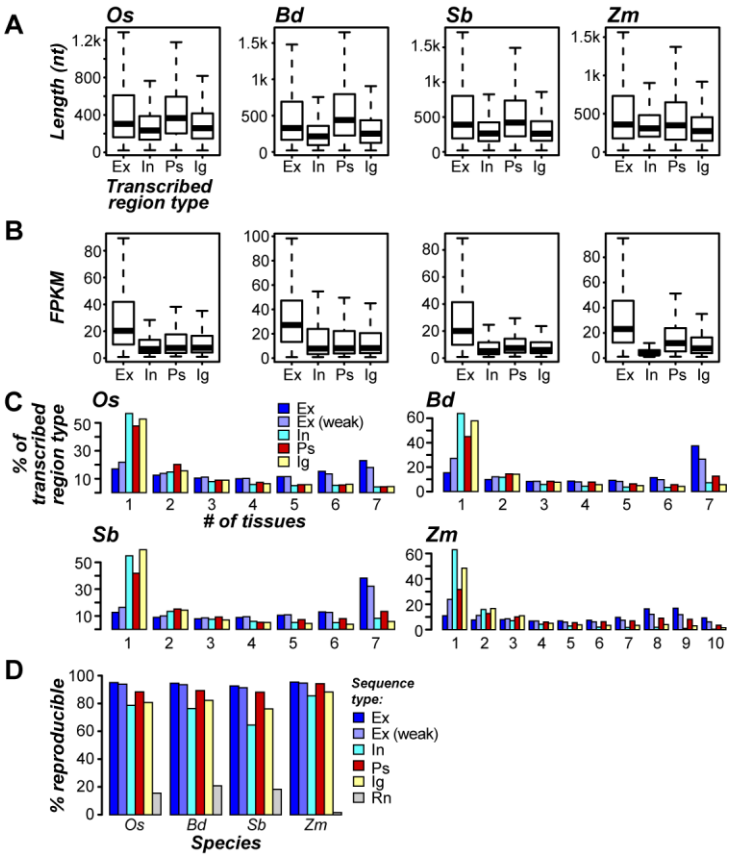
predicted as Phen-TRs that are truly Phen-TR sequences) and recall (proportion of true Phen-TR sequences that are predicted as Phen-TRs). The Python script utilized to generate these predictions (ML_classification.py) is available on GitHub: https://github.com/ShiuLab/ML-Pipeline. Binary predictions were also generated using the support vector machine and logistic regression algorithms implemented in Scikit-learn, but random forest provided the highest performance by AUC-ROC.

Three-class prediction models were generated similarly to binary models, except that 500 balanced datasets were utilized due to low sample sizes among miRNA-TR sequences (n=19). Each of the 500 balanced datasets provided sequences with three prediction scores representing the proportion of 500 random forest trees predicting a sequence as Phen-TR, miRNA-TR, or Pseu-TR, and the final prediction scores were generated by calculating the median all scores. Sequences were classified based on the highest median score. Binary and three-class classifications were merged using a priority system: Phen-TR in either model > miRNA-TR in the 3-class model > Pseu-TR (**Fig. 3.5C**).

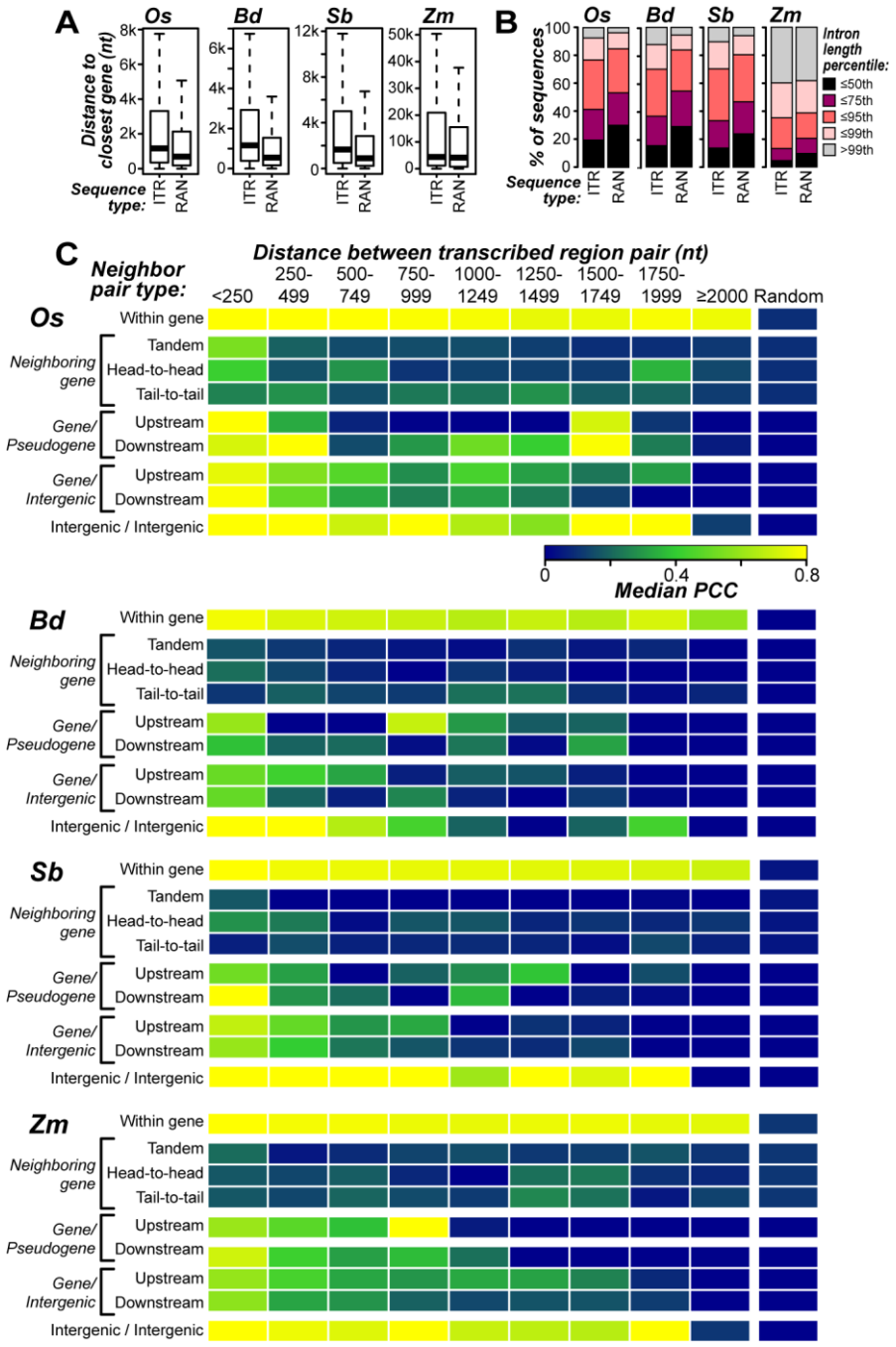
ACKNOWLEDGEMENTS

I wish to thank Rosalie Sowers for processing rice histone mark data, Ning Jiang for providing a rice repeat library, Robin Buell for helpful discussions, and Jack R. S. DeCatters for inspiration and support.

APPENDIX



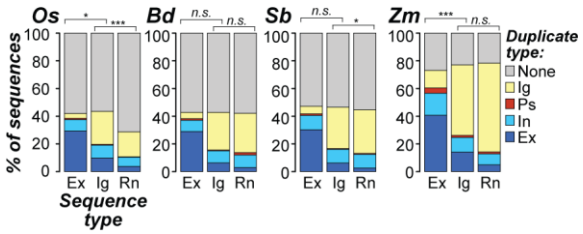
Supplemental Figure 3.1 Expression characteristics of transcribed regions in four Poaceae species. (A) Boxplots of length distributions among transcribed regions that overlap exon (Ex), intron (In), pseudogene (Ps), or intergenic (Ig) regions. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize. Nt: nucleotides. (B) Boxplots of maximum FPKM distributions among all tissues. (C) Distributions of expression breadth (i.e. # of tissues with expression evidence) for transcribed regions. A subset of exons with expression levels ±5% of intergenic transcribed regions (Ex (weak)) are also shown. (D) Percentage of sequences that reproducible across replicate leaf transcriptome datasets. Intergenic sequences were randomly-sampled to determine the background expected reproducibility (Rn).



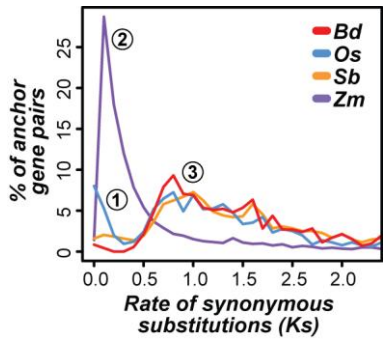
Supplemental Figure 3.2 Distance and expression correlation between intergenic transcribed regions and neighboring genes. (A) Boxplots of distance distributions between genes and intergenic transcribed regions (ITR) or randomly-selected intergenic regions (RAN). Nt: nucleotides. (B) Percentage of ITR and RAN sequences that are within multiple intron length

Supplemental Figure 3.2 (cont'd)

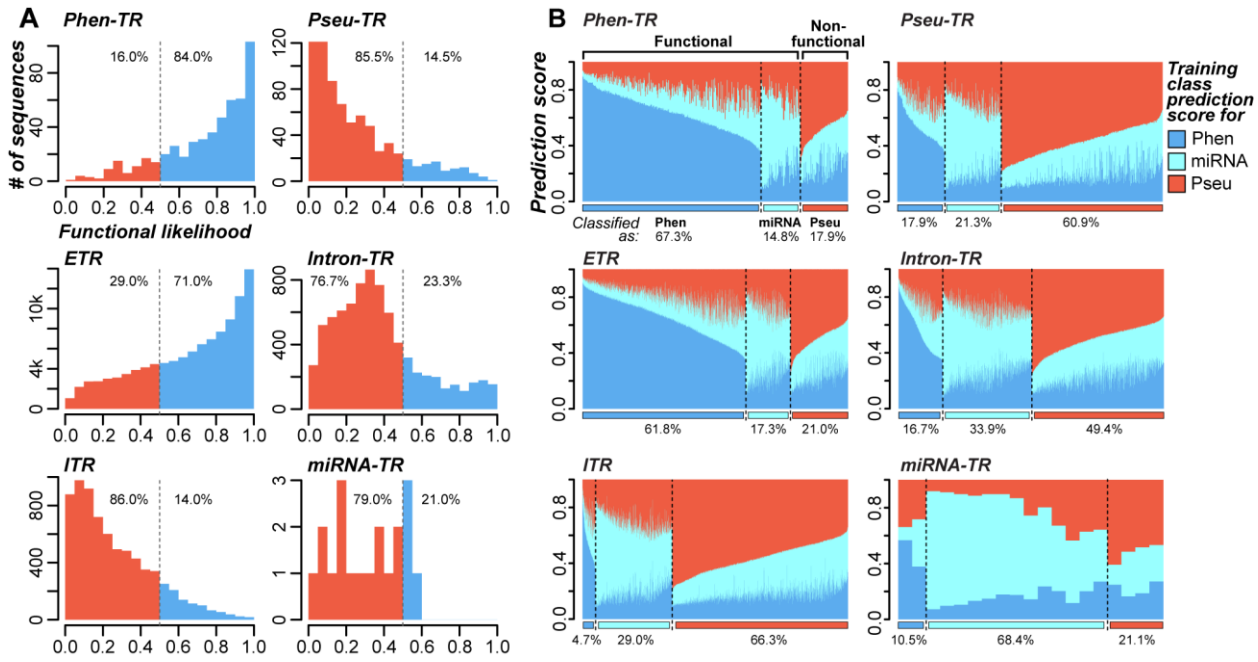
percentiles to the closest neighboring gene. Intron length percentiles were calculated based on the length distributions of annotated introns within each species. (C) Heatmaps of expression correlation between neighboring pairs of transcribed regions. Colors represent the median Pearson's correlation coefficient (PCC) of expression levels (FPKM) across tissues between all transcribed regions pairs within a distance bin. Neighboring transcribed region pairs were classified according to whether they overlapped the same gene (within gene), neighboring genes, gene and pseudogene neighbors (Gene/Pseudogene), genes and surrounding intergenic space (Gene/Intergenic), or neighboring ITRs (Intergenic/Intergenic). Neighboring gene pairs were sub-classified according to whether genes were oriented in the same direction (Tandem), or different directions with proximal 5' (Head-to-head) or 3' (Tail-to-tail) regions. Gene/Pseudogene and Gene/Intergenic pairs were sub-classified according to whether the pseudogene or intergenic transcribed region was upstream or downstream of a gene neighbor. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize.



Supplemental Figure 3.3 Distributions of duplication types. Sequences shown include ETRs (Ex, x-axis), ITRs (Ig, x-axis), and random intergenic (Rn) sequences. Fisher's exact tests were used to test significance between the proportion of sequences that were duplicated (sum of Ex, In, Ps, and Ig duplicate proportions) or not duplicated (None). *: $p<0.05$, ***: $p<0.001$, n.s.: not significant, $p\geq0.05$, In: intron, Ps: pseudogene; *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize.



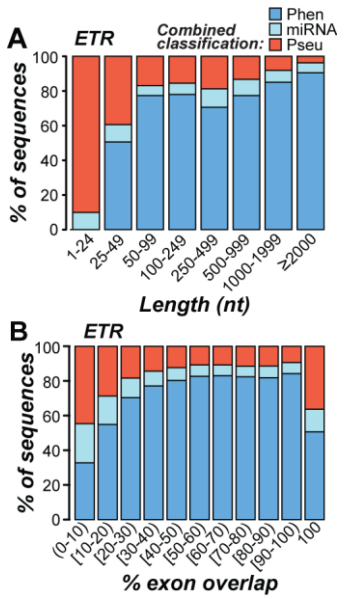
Supplemental Figure 3.4 Distributions of synonymous substitution rates between anchor genes of within-species collinear gene blocks. Circled numbers indicate synonymous substitution rate (*Ks*) peaks associated with (1) a low-*Ks* large-scale duplication in rice, (2) a recent whole genome duplication (WGD) in maize, and (3) the ρ and σ WGD events. Average *Ks* values among ρ and σ duplicates have been estimated at 0.9 and 1.7, respectively (Paterson et al., 2004; Tang et al., 2010). *Ks* distributions for these two events are highly overlapping and cannot be effectively distinguished. Due to uncertain origin and timing of the low-*Ks* rice duplication (1) (Wang et al., 2011), duplicates from this event were not included in further analysis. *Os*: rice, *Bd*: *B. distachyon*, *Sb*: sorghum; *Zm*: maize.



Supplemental Figure 3.5 Prediction score distributions based on binary and three-class function prediction models. (A) Functional likelihood distributions of various sequence classes based on the binary model. TR: transcribed region; Phen: phenotype exon; Pseu: pseudogene; ETR: exon transcribed region; ITR: intergenic transcribed region. Higher and lower functional likelihood values indicate greater similarity to Phen-TRs and Pseu-TRs, respectively. Vertical dashed lines indicate the threshold for calling a sequence as functional or non-functional. The percentages to the left and right of the dashed line indicate the percent of sequences predicted as functional or non-functional, respectively. (B) Stacked bar plots indicate the prediction scores among various sequence types for each of the three classes: dark blue: phenotype exon (Phen), cyan: miRNA, red: pseudogene (Pseu). Sequences were classified as one of the three classes according to the highest prediction score. Color bars below the chart indicate the predicted class, with the same color scheme as the prediction score. Sequences classified as Phen or miRNA were considered putatively functional, while those classified as Pseu were considered

Supplemental Figure 3.5 (cont'd)

non-functional. Percentages below a classification region indicate the proportion of sequences classified as that class.



Supplemental Figure 3.6 Relationship between sequence length, % exon overlap, and function predictions. Distributions of function predictions among exon transcribed regions (ETRs) of various lengths (A) and degrees of overlap with exon annotations (B). Sequences were classified as phenotype exon-like (Phen; dark blue), miRNA-like (cyan), or pseudogene-like (Pseu; red). Parentheses and brackets indicate a value was excluded from or included in a range, respectively. Nt: nucleotides.

CHAPTER 4: CHARACTERISTICS OF PLANT ESSENTIAL GENES ALLOW FOR WITHIN- AND BETWEEN-SPECIES PREDICTION OF LETHAL MUTANT PHENOTYPES [1]

[1] The work described in this chapter has been published:

John P. Lloyd, Alexander E. Seddon, Gaurav D. Moghe, Matthew C. Simenc, Shin-Han Shiu (2015) Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. *Plant Cell* **8**: 2133-2147

ABSTRACT

Essential genes represent critical cellular components whose disruption results in lethality. Characteristics shared among essential genes have been uncovered in fungal and metazoan model systems. However, features associated with plant essential genes are largely unknown and the full set of essential genes remains to be discovered in any plant species. Here we show that essential genes in *Arabidopsis thaliana* have distinct features useful for constructing within- and cross-species prediction models. Essential genes in *A. thaliana* are often single copy or derived from older duplications, highly and broadly expressed, slow evolving, and highly connected within molecular networks compared to genes with non-lethal mutant phenotypes. These gene features allowed the application of machine learning methods that predicted known lethal genes as well as an additional 1,970 likely essential genes without documented phenotypes. Prediction models from *A. thaliana* could also be applied to predict *Oryza sativa* and *Saccharomyces cerevisiae* essential genes. Importantly, successful predictions drew upon many features, while any single feature was not sufficient. Our findings show that essential genes can be distinguished from genes with non-lethal phenotypes using features that are similar across kingdoms and indicate the possibility for translational application of our approach to species without extensive functional genomic and phenomic resources.

INTRODUCTION

In the post-genome era, one major challenge in genetic research is in linking genotypes to phenotypes (Dowell et al., 2010; 1000 Genomes Project Consortium, 2010). Genome-wide phenotype information, obtained through large-scale loss-of-function studies, is available for several eukaryotic models including *Saccharomyces cerevisiae* (Winzeler et al., 1999), *Caenorhabditis elegans* (Kamath et al., 2003), *Drosophila melanogaster* (Boutros et al., 2004) and *Schizosaccharomyces pombe* (Kim et al., 2010). This information allows systematic analysis of genotype-phenotype connections and provides clues on homologous gene functions in species where large-scale loss-of-function analysis cannot be readily applied. By comparison, only a small proportion (~15%) of genes in the model plant *Arabidopsis thaliana* are associated with well-curated phenotype information (Lloyd and Meinke, 2012), despite the availability of powerful reverse genetics resources that allow for the potential of near-saturation mutagenesis studies (Kuromori et al., 2009). This is due in large part to the time and resources required for cultivating and phenotyping mutant populations. While *S. cerevisiae* and *C. elegans* have generation times measured in hours or days, *A. thaliana*, a relatively fast-growing plant, requires 5 to 6 weeks to begin seed production (Meyerowitz, 1989). These difficulties are exacerbated by high gene duplication rates in plants, due to both polyploidization (Soltis et al., 2009) and tandem duplications (Rizzon et al., 2006; Hanada et al., 2008), which result in many genes not exhibiting a phenotype under controlled conditions. Thus, the ability to effectively prioritize gene selection by predicting mutant phenotypes would represent an important step towards streamlining intensive and costly phenotypic analysis in plants.

Among genes with apparent phenotypes when lost, “essential” genes (lethal-phenotype genes) have been the target of focused analysis because they perform functions required for organismal viability and are critical in the investigation of potential drug targets in microbes (Golling et al., 2002; Firon et al., 2003; Kobayashi et al., 2003; Glass et al., 2006; Meinke et al., 2008; Silva et al., 2008). In *S. cerevisiae*, a variety of genomic features are associated with essential genes, including but not limited to singleton status, elevated transcription levels, and strong phylogenetic conservation (Winzeler et al., 1999; Kim et al., 2010). Some of these attributes are shared by lethal-phenotype genes in *C. elegans*, *S. pombe*, and *Mus musculus* (Kamath et al., 2003; Kim et al., 2010; Yuan et al., 2012). Using these features, lethal-phenotype genes have been predicted in *S. cerevisiae* and *M. musculus* (Seringhaus et al., 2006; Acencio and Lemke, 2009; Yuan et al., 2012).

In plants, essential genes tend to be single copy (Mutwil et al., 2010; Lloyd and Meinke, 2012) and have distinct functional biases (Tzafrir et al., 2004; Lloyd and Meinke, 2012). It has also been shown that genes with housekeeping functions that may or may not have lethal phenotypic consequences tend to be present in single copy across many plant species (De Smet et al., 2013). In addition to single copy status, essential genes are often highly-connected in gene functional networks (Mutwil et al., 2010), and genes with embryo-lethal defects tend to be connected with one another in the AraNet functional network (Lee et al., 2010). With these pioneering studies, an outstanding question is what other characteristics plant essential genes possess. For example, although single-copy genes tend to be essential, there are a number of duplicate genes that are essential. Thus, from the gene duplication perspective, it is possible that the extent, timing, and mechanism of duplication may be important. Similarly, one would expect that cross-species conservation, selective pressure, and expression characteristics will be related

to whether a gene is essential or not. Nonetheless, these features have not been evaluated for their relationship with plant essential genes.

Aside from the studies of essential gene features, Mutwil et al. (2010) identified clusters in their gene network with higher proportions of lethal-phenotype genes and predicted six novel essential genes. Although this study established a set of essential gene predictions in plants, the method will miss any essential genes outside of enriched clusters and therefore is not applicable genome wide. One potential solution to this is to predict lethal-phenotype genes based on many gene features beyond simply presence in a co-expression cluster, as this can produce genome-wide and potentially more accurate predictions. A data integration approach that made use of sequence data and expression correlation was successful in predicting functional overlap between *A. thaliana* duplicates, i.e. the absence of a phenotype due to buffering effects from another gene (Chen et al., 2010). Although the prediction of genetic buffering effects represents the opposite extreme of potential mutant phenotypes, a similar methodological framework could be used to predict essentiality or other detectable phenotypes on a genome scale. However, such a framework is not currently available.

To determine the feasibility of large-scale lethal-phenotype gene prediction in *A. thaliana* we collected loss-of-function phenotype data for ~3,500 genes and assessed relationships between phenotype lethality and gene function, copy number, duplication, expression levels and patterns, rate of evolution, cross-species conservation, and network connectivity, many of which were not explored previously in detail. We generated machine learning models to identify additional lethal-phenotype genes on the basis of multiple gene features, including a predictive model based only on sequence-derived features. Finally, as lethal-phenotype genes share many

characteristics between species, we tested whether lethal-phenotype predictions would be possible across species boundaries.

RESULTS AND DISCUSSION

Phenotype classification and functions of genes with lethal phenotypes

To predict lethal mutant phenotypes in *Arabidopsis thaliana*, loss-of-function phenotype descriptions were collected for 3,443 genes (Supplemental Table 4.1) (Kuromori et al., 2006; Ajjawi et al., 2010; Lloyd and Meinke, 2012; Savage et al., 2013), covering 12.7% of *A. thaliana* protein-coding genes. A phenotype was considered "lethal" if it resulted in developmental arrest at the gametophytic, embryonic, seedling, or rosette stage prior to bolting or extreme developmental defects that are expected to significantly affect plant growth in laboratory growth conditions. Under this definition, the loss-of-function phenotypes of 705 (20.5%) genes were considered lethal and the remaining (2,738; 79.5%) were considered non-lethal (Supplemental Table 4.1). Genes displaying lethal and non-lethal mutant phenotypes are referred to as "lethal genes" (essential genes) and "non-lethal genes," respectively. Genes not in our phenotype dataset are referred to as "undocumented genes".

An earlier study demonstrated that genes involved in, for example, RNA synthesis and modification, protein synthesis, and protein degradation tend to have higher essential-to-non-essential gene ratios (Lloyd and Meinke, 2012). However, that study classified genes into 11 categories and included only 5% of *A. thaliana* genes. In addition, despite the differences in ratios, the statistical significance of such differences is unclear. To assess if there is a statistically significant bias in the function of lethal genes and to assess if gene functions may be useful for generating predictions genome-wide, we tested for over- and under-representation of lethal genes in Gene Ontology (GO) categories (see Methods). We identified 28 terms in which lethal genes are significantly over- or under-represented compared to non-lethal genes (Fisher's exact tests

(FET), adjusted $p < 0.05$; Supplemental Fig. 4.1). Lethal genes in our dataset tend to be enriched in the translation, nucleolus, mitochondrion, and plastid categories and are rarely associated with signaling and regulation-related terms (signal transduction, cell communication, kinase and transcription factor activity, and response to endogenous, biotic, and abiotic stimulus). We also found that several basic developmental processes, such as reproduction, pollination, and the cell cycle, tend to be over-represented with lethal genes. In total, 27 GO terms that contain over- or under-represented numbers of lethal genes (not including the embryo development term; see Methods) were used in machine learning predictions of lethal-phenotype genes.

Copy number of lethal genes

In addition to functional bias, the presence or absence of paralogs is correlated with phenotypic severity in fungi (Winzeler et al., 1999; Gu et al., 2003; Kim et al., 2010) as paralogs may compensate for the loss of related genes and buffer the effects of gene loss. It has also been shown that single-copy genes in *A. thaliana* tend to be lethal genes (Mutwil et al., 2010), a trend also observed in another study (Lloyd and Meinke, 2012),. Consistent with these studies, lethal genes in our phenotype dataset are more commonly present as single copy genes than non-lethal genes (FET, $p < 4e\text{-}10$; **Fig. 4.1A**). This result provides additional support for the relationship between lethality and singleton status in plants, with a much larger gene set than in a previous study (Mutwil et al., 2010), and also indicates that gene copy number represents a potentially useful feature for lethal gene prediction. While we expected that lethal genes would be overrepresented in other small paralogous groups, both double- and triple-copy genes have a statistically similar proportion of lethal and non-lethal genes (FET, p=0.29 and 0.11 for double-copy and triple-copy genes, respectively; **Fig. 4.1A**). Thus, the presence of even a single paralog

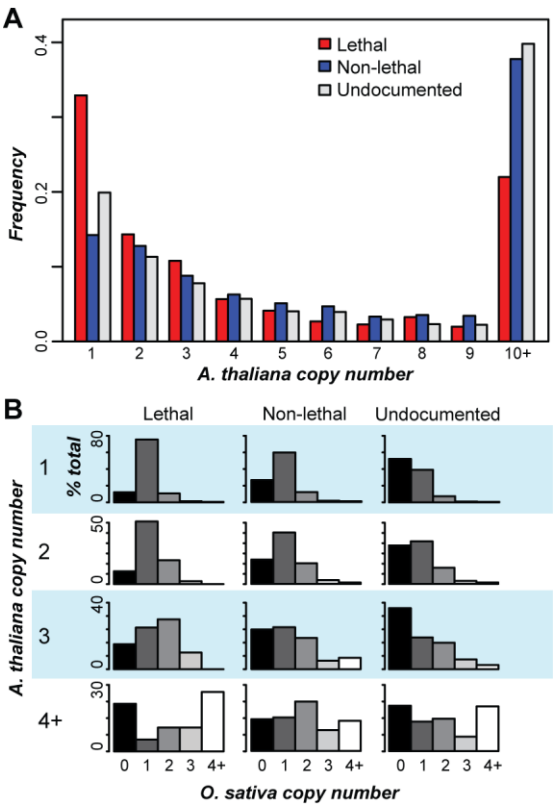


Figure 4.1 Copy number of phenotype genes in *A. thaliana* and *O. sativa*. (A) Frequency distribution of the number of paralogs (copy number) in the sets of lethal, non-lethal, and undocumented (i.e., no documented phenotype) genes. (B) Distributions of orthologous group sizes between *A. thaliana* and *O. sativa*. Rows indicate *A. thaliana* gene copy numbers in the orthologous groups, while columns denote phenotype categories.

provides appreciable functional overlap and therefore reduces the likelihood of lethality following disruption of a gene in laboratory conditions.

As lethal genes are enriched among certain functional categories and tend to be single copy, it is possible that lethal gene duplicates with particular functions were preferentially reduced to single copy. This preferential reduction to single copy appears to be conserved across species. Single-copy *A. thaliana* lethal genes tend to more often have one *Oryza sativa* (rice) ortholog compared to non-lethal and undocumented genes (**Fig. 4.1B**). More lethal *A. thaliana* genes also have readily identifiable homologs in *O. sativa* (87%) compared to non-lethal (77%; FET of lethal vs. non-lethal, $p < 5e\text{-}10$) and undocumented (54%; FET of lethal vs. undocumented, $p < 5e\text{-}10$) genes, which suggests a stronger degree of selective constraint on lethal genes. Considering that there were repeated rounds of whole-genome duplications in both the *A. thaliana* and the rice lineages (Paterson et al., 2004; Cui et al., 2006), the conserved single-copy status of *A. thaliana* lethal genes and their rice orthologs suggests that the loss of lethal gene paralogs compared to non-lethal gene paralogs is not completely random. In addition, this conservation of single-copy status suggests that single-copy rice orthologs are likely lethal-phenotype genes as well. Such cross-species conservation is explored in greater depth in a later section.

Duplication timing of lethal-phenotype genes

Although lethal genes are more likely to be single copy compared to non-lethal and undocumented genes, ~67% of lethal genes have paralogs, raising the question: why do some duplicate genes have a lethal phenotype in null mutant backgrounds? For genes with paralogs, a greater period since duplication may allow for a higher degree of functional divergence, which lessens the ability of duplicates to compensate for the loss of one another. An earlier study found

that essential genes tend to have greater protein sequence divergence from their paralogs (Lloyd and Meinke, 2012). Accordingly, we asked if lethal genes with paralogs (referred to as "lethal gene duplicates") would be the product of older duplication events compared to non-lethal genes with paralogs ("non-lethal gene duplicates"). Using synonymous substitution rate (*Ks*) as a proxy for duplication time, most-similar lethal gene duplicate pairs have significantly higher *Ks* values (median = 1.69) than those of most-similar non-lethal gene duplicate pairs (median = 1.07; Kolmogorov–Smirnov test (KST), $p < 3e\text{-}08$; **Fig. 4.2A**). One possible explanation for the lower median *Ks* among non-lethal gene duplicates is that they tend to be genes that arose after duplication events took place, i.e. lineage-specific genes. To assess this, we eliminated a subset of lineage-specific genes by focusing on genes with homologs in rice and again performed the *Ks* analysis. The results were almost identical to the results based on the full set of lethal and non-lethal genes (median lethal *Ks* = 1.7, median non-lethal *Ks* = 1.03; KST, $p < 5e\text{-}08$), indicating that lineage-specific genes may not fully explain the differences in *Ks* distributions between lethal and non-lethal genes.

Interestingly, the major *Ks* peak for non-lethal gene duplicates coincides with that for duplicates derived from the α whole-genome duplication (WGD) event that took place 50-65 million years ago (**Fig. 4.2B**) (Beilstein et al., 2010). By contrast, the major *Ks* distribution peaks for lethal gene duplicates (**Fig. 4.2A**) coincide with the peak *Ks* for not only duplicates derived from the α but also the much older βγ WGD (**Fig. 4.2B**) (Bowers et al., 2003), contributing to the significantly higher *Ks* values among lethal gene duplicates compared to non-lethal ones. This suggests that lethal gene duplicates may be generated from both WGD events, raising the question of how often lethal-phenotype genes retain their duplicates from these events.

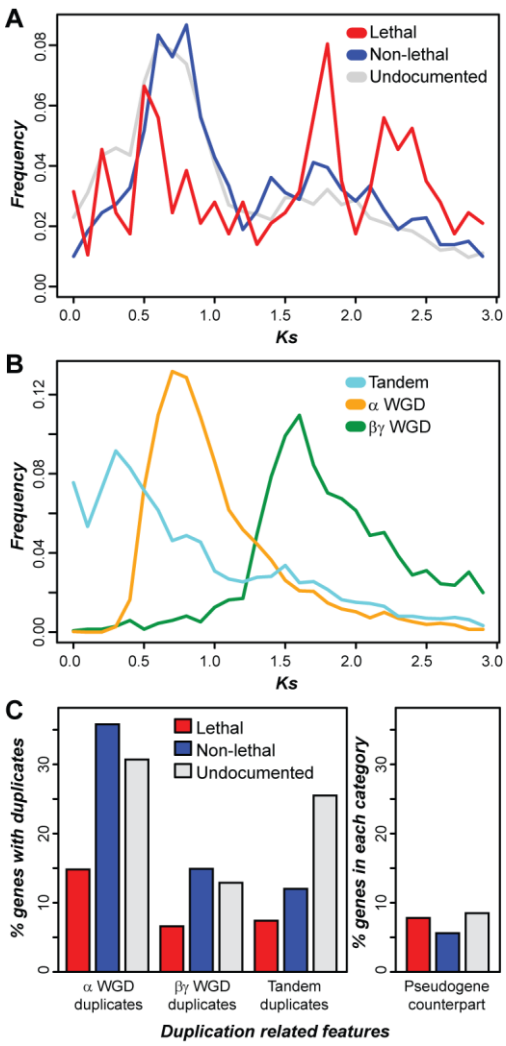


Figure 4.2 Duplication timing and type of *A. thaliana* phenotype genes. (A) Synonymous substitution rate (*Ks*) distributions for gene pairs of lethal, non-lethal, and undocumented genes and their most similar paralog. Gene pairs with higher *Ks* values are expected to be the result of older duplication events. Genes in a pair may not be from the same phenotype category. (B) *Ks* distributions of genes duplicated via tandem, and the α and βγ whole genome duplications (WGD). Some genes are derived from both tandem and whole-genome duplications. (C) Percent of duplicated lethal, non-lethal, and undocumented genes that have a paralog derived from α WGD, βγ WGD, and tandem duplications. Percent of all lethal, non-lethal, and undocumented

Figure 4.2 (cont'd)

genes with significant sequence similarity (percent identity ≥ 40%) to ≥1 pseudogenes is shown in the right-most portion of the panel.

Assuming that duplication rates are similar among all genes, significantly higher *Ks* values among lethal gene duplicates suggest that duplicates of lethal genes are more frequently lost than non-lethal gene duplicates. This is consistent with the finding that lethal genes tend to be single copy (Mutwil et al., 2010; Lloyd and Meinke, 2012; **Fig. 4.1**). In addition, we found that significantly fewer lethal genes with paralogs have retained their duplicates generated during WGD events compared to non-lethal genes (FET, $p < 4e\text{-}10$ and 3e-7 for the α and $\beta\gamma$ events, respectively; **Fig. 4.2C**). This analysis focuses on all possible duplicate pairs and the conclusion remains the same for the α WGD if we examine only the most closely-related paralogs (as in **Fig. 4.2A**; FET, α: $p < 3e\text{-}10$ and $\beta\gamma$: $p = 0.06$; Supplemental Fig. 4.2). Thus, some lethal genes retain their duplicates from WGD events, but the retention rate of lethal genes is lower than that of non-lethal genes. In addition, while a major peak in the lethal gene *Ks* distribution (**Fig. 4.2A**) coincides with the *Ks* peak from the $\beta\gamma$ WGD events (**Fig. 4.2B**), the lethal gene pairs underlying the peak in **Fig. 4.2A** may not necessarily represent duplicates retained from the $\beta\gamma$ WGD event. We also found that pseudogenes resembling lethal genes are more often present compared to those resembling non-lethal genes (FET, $p = 0.03$), although this proportion is not significantly different from that for undocumented genes ($p = 0.54$; **Fig. 4.2C**). In addition to WGD, tandem duplication is another major mechanism that contributes to form paralogous genes (Rizzon et al., 2006; Hanada et al., 2008). We found that duplicate lethal genes are less likely to be present in tandem clusters compared to non-lethal duplicates (FET, $p < 0.01$) and undocumented duplicates (FET, $p < 4e\text{-}10$; **Fig. 4.2C**). Furthermore, the few lethal genes derived from tandem duplications tend to have larger *Ks* values (median = 1.22) compared to non-lethal (median = 0.64; KST, $p = 0.05$) and undocumented (median = 0.69; KST, $p < 0.02$) tandem duplicates. These results indicate that lethal gene duplicates have a significantly higher loss rate after WGD and a

significantly lower proportion of tandem genes compared to non-lethal and undocumented gene duplicates. If lethal gene duplicates cannot be attributed to tandem or whole genome duplications, then what mechanisms were responsible for generating these duplicates? One explanation may be that lethal gene duplicates were generated via WGD, but are not in present in recognizable WGD blocks. However, the α WGD blocks cover ~90% of *A. thaliana* genes (Bowers et al., 2003), and thus the above explanation can only account for few of the lethal gene duplication events. It is also possible that duplicates of lethal genes may be commonly produced through segmental duplication events similar to those found in human (Bailey et al., 2002), but that remains to be verified. In either case, this represents an intriguing question that calls for further study.

Although lethal genes tend to be present as singletons, when lethal gene paralogs are present, they are derived from relatively ancient duplication events, consistent with the interpretation that deletion of a gene with a lesser degree of functional overlap with its paralog(s) due to longer divergence time will result in more severe phenotypic effects. Our findings also identify a number of features that can be used for lethal-phenotype gene prediction, including singleton status, *Ks* with top paralog, presence of duplicates from the α or βγ WGD or tandem duplication events, presence of pseudogene counterparts, and absence of orthologs in other species (**Table 4.1**).

Relationship between phenotype lethality and gene expression

Overrepresentation of lethal genes among older duplicates compared to non-lethal genes suggests a higher degree of functional divergence among lethal gene duplicates. To explore this further, we compared gene expression levels and patterns between lethal and non-lethal genes. Duplicates with a higher degree of expression divergence are expected to perform their

Table 4.1 Features of essential genes in *A. thaliana*.

Category	Feature	Data type	Sign of lethal association[1]	*P*-value[2]	Seq. based feature[3]	Rice[4]	Yeast[4]
Gene duplication	α WGD duplicate retained	Binary	-	3.17E-10	No	No	No
	βγ WGD duplicate retained	Binary	-	3.07E-10	No	No	No
	Pseudogene present	Binary	+	0.035	Yes	Yes	No
	Tandem duplicate	Binary	-	7.93E-06	Yes	Yes	No
	Paralog *Ks*	Numeric	+	2.17E-08	Yes	Yes	Yes
	Gene family size	Numeric	-	1.20E-24	Yes	Yes	Yes
Expression	Median expression	Numeric	+	1.60E-08	No	Yes	Yes
	Expression variation	Numeric	-	0.002	No	Yes	Yes
	Expression breadth	Numeric	+	5.47E-20	No	Yes	No
	Expression correlation	Numeric	NA	0.072	No	No	No
	Expression correlation (*Ks* < 2)	Numeric	-	0.004	No	No	No
Evolution and conservation	Core eukaryotic gene	Binary	+	2.44E-08	No	No	Yes
	Homolog not found in *O. sativa*	Binary	-	4.04E-10	Yes	No	No
	Percent identity in plants	Numeric	+	2.73E-06	Yes	No	No
	Percent identity in metazoans	Numeric	NA	0.254	Yes	No	No
	Percent identity in fungi	Numeric	NA	0.077	Yes	No	No
	A. lyrata homolog *Ka*/*Ks*	Numeric	-	0.012	Yes	No	No
	P. trichocarpa homolog *Ka*/*Ks*	Numeric	-	0.008	Yes	No	No
	V. vinifera homolog *Ka*/*Ks*	Numeric	-	0.003	Yes	No	No
	O. sativa homolog *Ka*/*Ks*	Numeric	-	0.012	Yes	No	No
	P. patens homolog *Ka*/*Ks*	Numeric	-	0.038	Yes	No	No
	Nucleotide diversity	Numeric	-	0.001	No	No	No
	Paralog *Ka*/*Ks*	Numeric	+	2.51E-14	Yes	Yes	Yes
Networks	Expression module size	Numeric	+	1.94E-34	No	No	Yes
	Gene network connections	Numeric	+	9.84E-11	No	No	Yes
	Protein-protein interactions	Numeric	NA	0.72	No	No	No
Miscellaneous	Gene body methylated	Binary	+	3.46E-10	No	No	No
	Paralog percent identity	Numeric	-	2.75E-33	Yes	Yes	Yes
	Protein length	Numeric	+	1.22E-06	Yes	Yes	Yes
	Domain number	Numeric	+	0.023	Yes	Yes	Yes

1. For each binary feature, + and – indicate that the proportion of lethal genes are significantly higher (over-representation) or lower (under-representation) than non-lethal genes, respectively. For each numeric feature, + and - indicate that lethal genes have significantly higher or lower feature values compared to non-lethal genes, respectively. NA indicates that there is no significant difference between lethal and non-lethal genes,

2. *P*-values from Fisher's exact tests (used for binary data) or Kolmogorov-Smirnov tests (used for numeric data).

3. Sequence-based features, where "Yes" indicates that a feature can be derived from genome sequence data.

4. Feature used ("Yes") or not used ("No") in rice or yeast lethal phenotype gene predictions.

molecular functions in more distinct temporal, spatial, and conditional contexts. Because of this, we expect lethal genes may show higher degrees of expression divergence with their paralogs compared to non-lethal genes. Consistent with this expectation, lethal gene duplicate pairs have significantly lower expression correlation (thus higher divergence) compared to non-lethal gene duplicates when $Ks \leq 2$ (KST; $0 < Ks \leq 1$, $p <$ 4e-4; $1 < Ks \leq 2$, $p <$ 0.01; **Fig. 4.3A**). However, older lethal and non-lethal genes show similar degrees of expression correlation with paralogs ($Ks > 2$; KST, $p = 0.35$). These results are also consistent with previous findings in *A. thaliana* that, unlike other eukaryotic species, expression divergence is not correlated with duplication timing (Gu et al., 2002; Ganko et al., 2007).

One potential explanation for the decreasing differences in expression divergence between lethal and non-lethal duplicates over time is that greater divergence in the protein-coding sequences among older duplicates has contributed to a higher degree of biochemical divergence between duplicates. Therefore the presence of a paralog with a similar expression profile no longer buffers against the consequences of gene loss. We found that lethal gene duplicates with $Ks > 2$ have a significantly higher ratio of non-synonymous to synonymous substitution rates compared with non-lethal duplicates (Ka/Ks; KST, $p <$ 6e-10; Supplemental Fig. 4.3), indicating that there is increased divergence at the protein-coding level for older lethal genes compared with non-lethal genes. This raises a question: among duplicates with $Ks < 1$, what underlying mechanisms contribute to the differences in expression correlation between lethal and non-lethal genes? Was there selection pressure driving the expression differences between lethal genes and/or maintaining expression similarity among non-lethals? Alternatively, were the patterns we see predominantly driven by neutral processes such as drift? In this context, the distinction between lethal and non-lethal genes may simply be how paralogs accrued

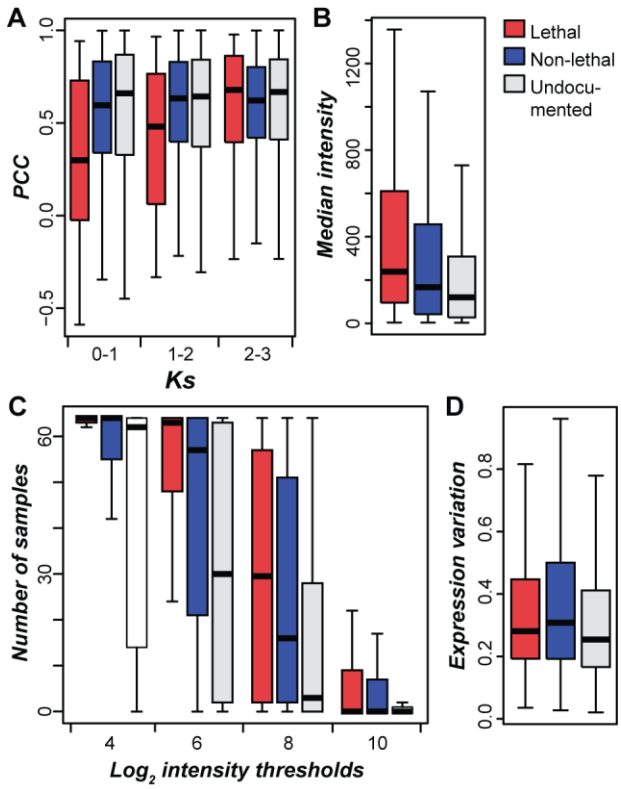


Figure 4.3 Expression characteristics of *A. thaliana* phenotype genes. (A) Box plots of expression correlations (Pearson's Correlation Coefficient, PCC) of paralogous gene pairs involving three gene categories - lethal (red), non-lethal (blue), and undocumented (grey) genes - across AtGenExpress developmental dataset samples (Schmid et al., 2005). Lower expression correlation indicates increased degree of expression divergence for a gene pair. Genes in a paralog pair may or may not be from the same phenotype category. (B) Box plots of median expression levels (array hybridization intensities) of genes in three categories across array experiments. (C) Box plots of numbers of samples where genes in each category were considered expressed according to multiple thresholds. (D) Box plots of expression variation across samples (median absolute deviation/median) in each gene category.

mutations that contribute to expression divergence and have little to do with selection. These possibilities need to be further studied.

In addition to expression divergence, expression level of a gene may affect phenotypic severity. In *S. cerevisiae* and *Mus musculus,* essential genes tended to be expressed at higher levels (Winzeler et al., 1999; Yuan et al., 2012). Consistent with findings in other species, in *A. thaliana* the expression levels of lethal genes across the AtGenExpress developmental expression series (n = 64; Schmid et al., 2005) are significantly higher than those of non-lethal genes (KST, $p < 2e\text{-}8$; **Fig. 4.3B**), suggesting that transcript levels are correlated with gene essentiality. In addition to expression level, lethal genes tend to be more broadly expressed across developmental stages and organs than non-lethal genes (KST, $p < 5e\text{-}19$, 4e-12, 6e-05, and = 0.19 for $\log_2$ intensity thresholds of 4, 6, 8 and 10, respectively; **Fig. 4.3C**). Finally, while lethal genes show a significantly lower degree of expression variation compared to non-lethal genes, the effect size is small (KST, $p < 0.01$; **Fig. 4.3D**). Although lethal genes tend to be highly expressed, 7% are expressed among the bottom third of all genes (defined as weakly-expressed, n= 51; $\log_2$ median intensity ≤ 4.39). Among 15 GO categories significantly overrepresented in weakly-expressed lethal genes compared to highly-expressed ones, 14 are related to transcriptional regulation due to the contribution of the same 15 genes across categories (Supplemental Table 4.2). These genes exhibit a broad spectrum of lethal phenotypes (gametophytic, embryo, and seedling) with notable developmental defects, including cotyledons with leaf-like characteristics (*FUS3*, *LEC1*, *LEC2*), precocious seed development (*FIS2*, *MEA*), and complete loss of the primary root (*STIP*). To summarize, we found that lethal gene duplicates tend to display higher expression divergence when $Ks \leq 2$, and higher protein-sequence divergence when $Ks > 2$. We also found that lethal genes tend to be more highly and

broadly expressed and have lower degrees of expression variation compared to non-lethal genes. Thus, a variety of expression characteristics correlate with phenotype lethality and were incorporated into lethal-phenotype prediction models (**Table 4.1**).

Conservation of lethal genes

Due to their severe phenotypic consequences when lost, lethal genes likely experienced greater selective constraint compared to genes with non-lethal phenotypes. The *Ka*/*Ks* values between *A. thaliana* lethal genes and their homologs in five plant species tend to be significantly lower compared to cross-species non-lethal gene homolog pairs (KST, see figure legend for *p*-values; **Fig. 4.4A**). Similarly, lethal genes have a significantly lower degree of nucleotide diversity among 80 accessions of *A. thaliana* compared to non-lethal genes (KST, $p < 7e\text{-}4$; **Fig. 4.4B**). Both results suggest that lethal genes are experiencing stronger purifying selection. There are two potential confounding factors. First, lethal genes tend to be expressed at higher levels than non-lethal genes (**Fig. 4.3B**) and highly-expressed genes often experience greater selective pressure due to disproportionate effects of toxic protein misfolding (Drummond et al., 2005). Second, expression levels can affect calculations of *Ka*/*Ks* due to codon usage bias (Duret and Mouchiroud, 1999). Thus we analyzed the relationship between the *Ka*/*Ks* values and median expression levels. Consistent with our expectation, we found a negative correlation between median expression levels of lethal genes and *Ka*/*Ks* values in each of the five plant species (median Pearson Correlation Coefficient, PCC = -0.23). However this relationship explains only a minor component of the variation in selective pressure experienced by lethal genes (r^2 values range from 0.03 to 0.08). Thus, our finding that lethal genes are experiencing stronger negative selection is not simply due to their higher expression levels.

Similar to the *Ka/Ks* based comparison, lethal genes have significantly higher sequence identities to their best matches in other plant lineages compared to non-lethal genes (**Fig. 4.4C**). Although no significant difference in sequence identity is noted between lethal and non-lethal genes when considering their best matches in animal and fungal species, a significantly higher proportion of lethal genes (25%) are present in orthologous clusters consisting of genes from seven diverse eukaryotes ("core eukaryotic genes"; see Methods) compared to non-lethal genes (15.7%; FET, $p < 3e\text{-}8$). Lethal genes tend to be the result of older duplications and are present in fewer copies than non-lethal genes. As any set of genes with these features may be highly conserved, we assessed the effects of copy number and duplication age on the sequence conservation of lethal genes. We found that both timing of duplication (*Ks* value, $r^2 = 0.01$) and gene copy number ($r^2 = 0.03$) explain little of the variation in protein conservation across the plant kingdom for lethal genes. These results, along with those from the above analysis of the relationship between expression level and *Ka*/*Ks*, show that correlation between features that are expected to be dependent can be far from perfect, highlighting the need to consider them jointly for lethal-phenotype gene predictions.

Together with our finding that lethal genes tend to have homologs in *O. sativa* (**Fig. 4.1B**), the results from **Fig. 4.4** indicate a higher degree of conservation for lethal genes. Because evolutionary rate values and protein conservation metrics could prove useful in a prediction context, they were included in later lethal gene prediction (**Table 4.1**). However, we should emphasize that the *Ka/Ks*, nucleotide diversity, and cross-species sequence identity distributions between lethal and non-lethal genes overlap substantially, i.e. the effect sizes are small despite significant differences (**Fig. 4.4**). One explanation is that the non-lethal genes studied here are those with observable phenotypes in loss-of-function backgrounds. These non-lethal genes thus

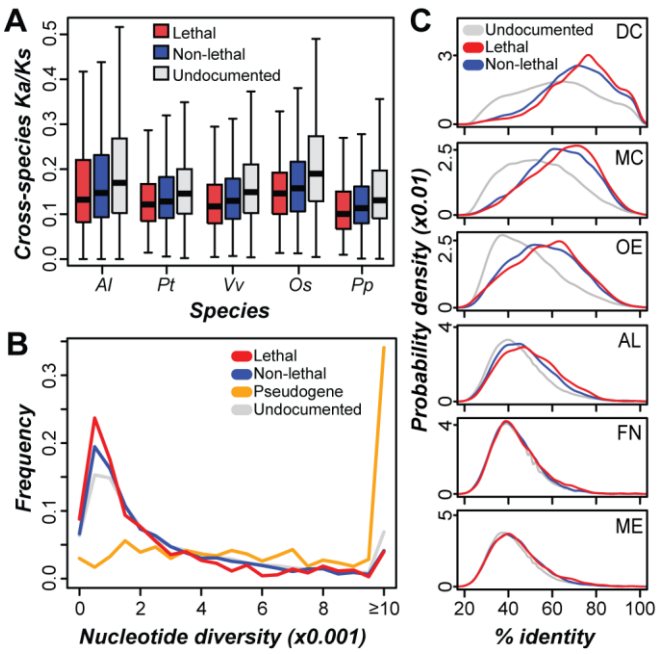


Figure 4.4 Evolutionary rate and cross-species protein conservation of *A. thaliana* phenotype genes. (A) Ratios of non-synonymous substitutions (*Ka*) to synonymous substitutions (*Ks*) between *A. thaliana* genes and homologs in the same OrthoMCL cluster from *Arabidopsis lyrata* (*Al*; KST of lethal vs. non-lethal, $p<0.02$), *Populus trichocarpa* (*Pt*; $p<0.01$), *Vitis vinifera* (*Vv*; $p<0.01$), *O. sativa* (*Os*; $p<0.02$), and *Physcomitrella patens* (*Pp*; $p<0.04$). Lower *Ka*/*Ks* values are indicative of stronger negative selection pressure. (B) Distributions of nucleotide diversity for lethal, non-lethal, pseudo-, and undocumented genes. Higher nucleotide diversity values indicate higher degree of sequence polymorphism between *A. thaliana* accessions. (C) Probability density distributions of median % identity of lethal, non-lethal, and undocumented genes to top BLASTP matches in dicotyledonous plants (DC; lethal vs. non-lethal KST, $p<2e-6$), monocotyledonous plants (MC; $p<7e-4$), other embryophytic plants (OE; $p=0.05$), algae (AL; $p<7e-6$), fungi (FN; $p=0.07$), and metazoans (ME; $p=0.25$).

are likely subjected to strong selection, although not as strong as the selection against lethal gene mutations. We should also note that none of the examined characteristics that distinguish between lethal and non-lethal genes are perfect. As a result, multiple characteristics are considered jointly in statistical learning models for predicting lethal-phenotype genes (described in a later section).

Network connectivity of lethal-phenotype genes

In *S. cerevisiae*, proteins that are highly connected in physical protein-protein interaction networks tend to be essential (Jeong et al., 2001). Similarly, analyses of *S. cerevisiae* and *A. thaliana* gene networks based on functional relatedness between genes have demonstrated that highly connected genes tend to have severe loss-of-function phenotypes (Lee et al., 2010; Mutwil et al., 2010), and identification of co-expression clusters in *A. thaliana* that are enriched in lethal phenotype genes was useful in selecting and validating six novel essential genes (Mutwil et al., 2010). Further, an *A. thaliana* gene functional network (AraNet; Lee et al., 2010) was used to demonstrate that genes with embryo-lethal phenotypes tend to be connected with one another in the gene network (Lee et al., 2010). However, no formal prediction of essential genes using AraNet data has been performed. To verify that the relationship between network connectivity and gene phenotype lethality also exists in our phenotype dataset, we examined co-expression networks established in this study, connections in the AraNet gene network, and protein-protein interaction data (Arabidopsis Interactome Mapping Consortium, 2011). We found that lethal genes in *A. thaliana* tend to be found in larger co-expression modules (median size = 19) than those containing non-lethal genes (median = 13; KST, $p < 2e\text{-}34$; **Fig. 4.5A**). In addition, lethal genes tend to be co-expressed (PCC > 0.86; 99^{th} percentile of all pairwise co-expression coefficients) with a greater number of genes (median = 20) than non-lethal genes

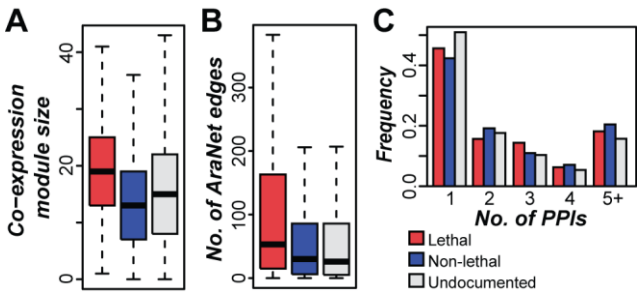


Figure 4.5 Network connectivity of *A. thaliana* phenotype genes. (A) Co-expression module sizes of lethal, non-lethal, and undocumented genes. Modules represent groups of genes clustered via K-means clustering (K=2000) based on expression similarity across *A. thaliana* development samples in AtGenExpress (Schmid et al., 2005). (B) Number of edges connected to genes in the three categories with a log likelihood ≥ 1 in the AraNet network (Lee et al., 2010). Higher numbers of edges indicate increased connectivity within the network. (C) Distributions of the numbers of genes with 1, 2, 3, 4, or ≥5 protein-protein interactions (PPIs) (Arabidopsis Interactome Mapping Consortium, 2011).

(median = 5; KST, $p < 8e\text{-}8$). Similarly, *A. thaliana* lethal genes tend to have a greater number of interactions (median = 53) in the AraNet gene functional network than non-lethal genes (median = 30; KST, $p < 1e\text{-}10$; **Fig. 4.5B**). These results corroborate previous findings based on analysis of co-expression networks (Mutwil et al., 2010) and indicate that high interactivity in gene networks may be useful for establishing lethal-phenotype gene predictions.

In contrast to the relationship between gene essentiality and centrality in gene networks, connectivity within a physical protein-protein interaction network does not seem to be correlated with phenotypic severity in *A. thaliana* (Lloyd and Meinke, 2012). There remains no clear relationship between our updated phenotype data and protein-protein interactions (KST, $p = 0.73$; **Fig. 4.5C**). It remains to be determined if this is due to the lower coverage in the *A. thaliana* interactome map (12% of proteins compared to 30% in yeast; Jeong et al., 2001). Taken together, the higher connectivity among phenotype-lethal genes is consistent with the interpretation that their disruption may interfere with the function of many other genes. One additional possibility is low-interacting genes may play more specialized roles than high-interacting ones. Thus, low-interacting genes would tend not to have strong phenotypic consequences when mutated.

Prediction of lethal genes using a machine learning framework

Based on analysis of functional annotation, gene copy number, duplicate retention patterns, gene expression, evolutionary rates, cross-species conservation, and network connectivity (**Table 4.1**), we have identified a wide variety of genomic features correlated with phenotype lethality. In addition to these features, genes encoding longer proteins with a larger number of domains and those with CG gene body methylation (Takuno and Gaut, 2012) are more likely to exhibit lethal phenotypes upon disruption (**Table 4.1**). As these features do not

correlate perfectly with whether disruption of a gene results in a lethal or non-lethal phenotype, it raises the question as to whether a meaningful prediction of phenotype lethality is feasible if they are jointly considered. In addition, it remains unclear how these disparate features would differ in their contribution to *A. thaliana* lethal-phenotype gene prediction.

To address these questions, we applied machine learning methods that have been used for essential gene predictions in budding yeast (Seringhaus et al., 2006; Acencio and Lemke, 2009) and mouse (Yuan et al., 2012). A matrix of genes with a documented phenotype and their associated values for different features (Supplemental Table 4.3) was used as input for six machine learning classifiers (see Methods; **Fig. 4.6A**). To build the classifiers, 90% of our dataset was used for model building (training phase) and 10% was held out for testing the accuracy of the predictive model (validation phase). The model building process was repeated 10 times so that every gene in our phenotype dataset was held out of the model building exactly one time (10-fold cross-validation). We should emphasize that the training data were completely independent from the validation data. Performance was evaluated by calculating the Area Under the Curve - Receiver Operating Characteristic (AUC-ROC), where the AUC-ROC of a model based on random guessing is ~0.5 and that of a perfect model is 1.0. Using the best performing classifier, Random Forest (Ho, 1995), the lethal gene prediction model AUC-ROC is 0.81, which is significantly better than random guessing (**Fig. 4.6A**; see methods). To provide an alternative interpretation of model performance, we also examined the precision (proportion of predicted genes that are truly lethal) and recall (proportion of true lethal genes recovered) of our model (**Fig. 4.6B**). Based on this analysis, to correctly recover 50% of lethal genes, our precision is at 57%. Because the proportion of lethal genes in our dataset is 0.2, the precision of random guesses is expected to be ~20% (grey line, **Fig. 4.6B**), indicating that our methods perform

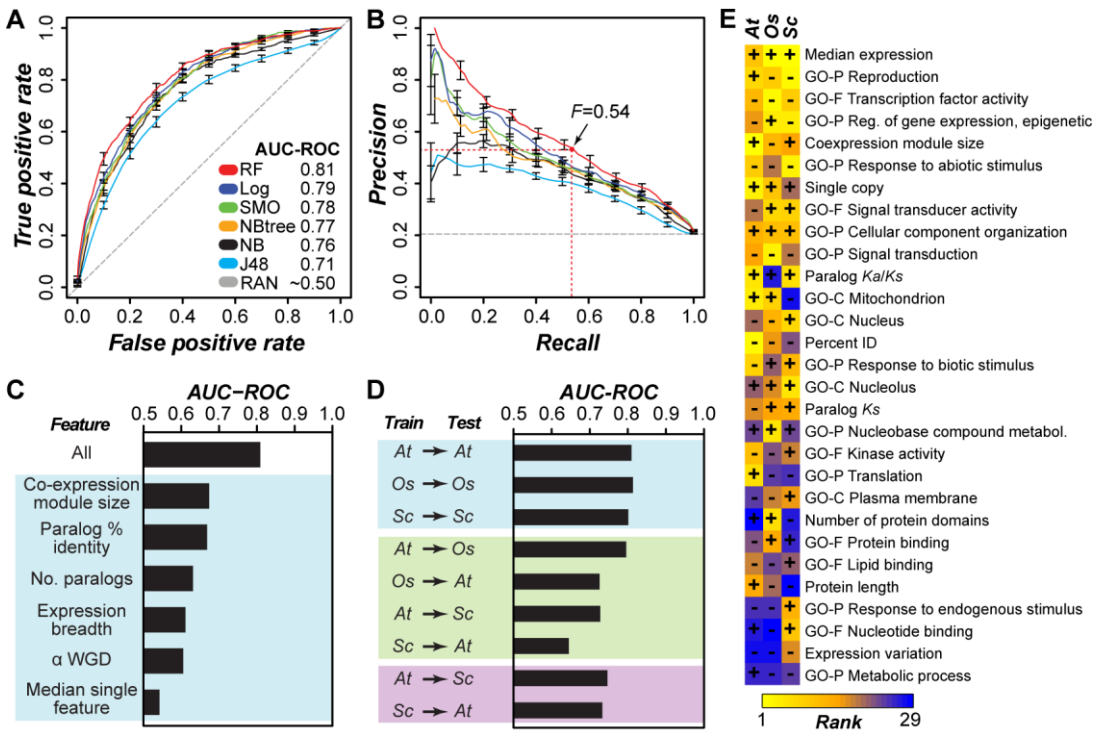


Figure 4.6 Machine learning performance of essential gene predictions. (A) Receiver Operating Characteristic (ROC) curves of the predictive models based on Random Forest (RF), Logistic regression (Log), SMO-SVM (SMO), Naïve Bayes tree (NBtree), Naïve Bayes (NB), and J48 decision tree (J48) using the best-performing parameter sets. Area Under the Curve (AUC)-ROC is indicated in the inset; an AUC-ROC value of ~0.5 is equivalent to random guessing while an AUC-ROC of 1 indicates perfect predictions. Diagonal dashed line: the expected performance of a model based on random guessing (RAN). Curves closer to the upper left corner of the chart represent a better predictive performance than curves that are closer to the diagonal dashed line. Error bars: standard error between 10 cross-validation runs. (B) Precision-recall curves for the models from (A). Precision: the proportion genes predicted as lethal that are actual lethal genes. Recall: the proportion of actual lethal genes predicted as lethal. Horizontal dashed line: the proportion of lethal genes in the dataset, which represents the expected precision

Figure 4.6 (cont'd)

based on random guessing. Error bars: standard error between cross-validation runs. (C) AUC-ROC values of the best-performing Random Forest machine learning classification using all features (All; median of ten cross-validation runs) in comparison to AUC-ROCs of the models based on each of the five most informative features (cyan background) and the median of all single feature predictions. (D) AUC-ROC values of within-species (cyan background) and cross-species (green and magenta background) predictions in *A. thaliana* (*At*), *O. sativa* (*Os*), and *S. cerevisiae* (*Sc*). Species on the left side of the arrows indicate the species from which data were used to train a prediction model. Independent datasets from species on the right of the arrows were used for testing. Predictions between A. *thaliana* and *S. cerevisiae* were performed both with a full feature set (green background) and with a subset of features in which the sign of SMO weights agree (magenta background). (E) Ranks and signs of SMO weights of 29 features available in *A. thaliana* (*At*), rice (*Os*), and yeast (*Sc*). A lower number rank of a feature weight indicates greater importance for within-species predictions. A more positive weight indicates better association with phenotype lethality and a more negative weight indicates better association with non-lethality.

reasonably well. By comparison, an earlier study based on co-expression clusters predicted and validated 6 novel essential genes out of a pool of 20 candidate genes (Mutwil et al., 2010). This represents a precision of 30%. However, this methodology applies only to essential gene-enriched clusters (357 genes, ~1.3% of *A. thaliana* annotated genes). As a result, any essential genes outside of these clusters cannot be predicted using this methodology and recall is expected to be very low. This highlights the need to consider a large suite of gene features for genome-wide predictions of essential genes.

We next used the best performing model to classify the rest of the 23,763 undocumented genes as potentially lethal or non-lethal when lost. This provided each gene with a "lethal-phenotype score", a value between 0 and 1 where higher values indicate higher confidence that a gene will display a lethal phenotype upon disruption. Notably, the highest lethal-phenotype score for an undocumented gene is 0.72 while the highest scoring lethal gene in our phenotype dataset is almost 0.90, indicating a distinction between the lethal genes in our training dataset and the rest of the genes in the *A. thaliana* genome and potential biases in our model. Applying the machine learning model and a lethal-phenotype score threshold resulting in the highest *F*-measure (harmonic mean of precision = 0.54 and recall = 0.54; arrow, **Fig. 4.6B**) in the training data, we identify 1,970 (8%) undocumented genes whose loss is expected to result in a lethal phenotype (Supplemental Table 4.1). Using this lethal-phenotype score threshold (0.31), we expect that 1,059 (1,970*precision) are correctly-predicted lethal genes and that there are 885 (1,059/recall-1,059) additional lethal genes that we fail to detect. Thus, we anticipate an additional 1,944 lethal genes in the undocumented gene set. Together with the 705 known lethal genes, 10% (~2,700) of *A. thaliana* protein-coding genes are expected to have lethal mutant phenotypes based on the lethal-phenotype score threshold of 0.31. As an additional validation

step, we collected an independent set of 60 *A. thaliana* phenotype genes based on a literature search (17 with a lethal phenotype; Supplemental Table 4.1) that are not included in our initial dataset of 3,443 lethal and non-lethal genes. The AUC-ROC of the best-performing Random Forest model is 0.83 for this independent set. Of the 17 genes with lethal phenotypes in this 60-gene dataset, 13 (77%) are correctly predicted as lethal and of the 43 non-lethal genes, 40 (93%) are correctly predicted as non-lethal.

To determine what features are among the most important to our predictions, we assessed the performance reduction resulting from the removal of each feature from prediction analysis. We found that no single feature is particularly critical for predictive performance by itself (all leave-one-out models have AUC-ROC ≥ 0.8 compared to 0.81 for the full model; Supplemental Table 4.4). These results are corroborated by the fact that machine learning predictions using all data types perform much better than predictions based on any single feature by itself (median AUC-ROC = 0.54, **Fig. 4.6C**; Supplemental Table 4.4). We also found that 46 features (80% of all features) are required to achieve an AUC-ROC of 0.80 (Supplemental Fig. 4.4), indicating that the contributions from most features are critical. This is consistent with our observations that, although the features we used are generally significantly distinct between lethal and non-lethal genes, in many cases the effect sizes are small (**Fig. 4.1-4.5**). In addition, many features we included here are likely dependent, although correlation between features is low (see Methods). In any case, our findings indicate that the predictive models for lethal-phenotype genes are robust and draw upon a wide variety of gene features to generate meaningful classifications of lethal and non-lethal genes.

Cross-species predictions of lethal-phenotype genes

Considering that some of the features we found to be correlated with gene lethality have been shown to be important in other species (Seringhaus et al., 2006; Yuan et al., 2012), this raises the question whether a prediction model trained with *A. thaliana* data (*A. thaliana* model) can be used to predict phenotype lethality across species boundaries. To test this, we first collected rice phenotype data for 92 genes (18 lethal, see Methods; Supplemental Table 4.1) and analogous genomics and functional annotation data (**Table 4.1**). Then a "rice model" was generated and applied to predict lethal genes within the rice test set using 2-fold cross validation. Surprisingly, this performed as well as within-*A. thaliana* predictions (AUC-ROC = 0.82; **Fig. 4.6D**), indicating that a significantly smaller gene set still allows lethal gene classification with comparable accuracy. We also tested if good predictions could be made in *A. thaliana* using a reduced gene set by randomly sampling 20 lethal and 80 non-lethal genes from our full dataset and making predictions with 2-fold cross-validation. This was repeated 100 times. The AUC-ROCs of these 100 models range from 0.55 to 0.88 with a median of 0.75. The median AUC-ROC indicates that few phenotype genes can be used to establish lethal gene prediction model with reasonable performance. The rather large variance in AUC-ROCs indicates that the genes included during model building can have a significant effect, particularly if the sample size is small. Next, to test if prediction across plant species is feasible, we trained prediction models using data from one species and predicted phenotype lethality for genes in test sets from another species (see Methods). Using the *A. thaliana* model, we can predict rice lethal genes with an AUC-ROC of 0.80 (**Fig. 4.6D**). A rice model is also capable of identifying *A. thaliana* lethal genes, although the performance is reduced (AUC-ROC = 0.72; **Fig. 4.6D**). This is potentially

because use of a model trained on a small gene set is ineffective in classifying a large number of genes in another species.

Given that cross-species phenotype prediction is feasible between *A. thaliana* and rice, which diverged over 200 million years ago, we sought to determine if lethal-phenotype genes were predictable across a significantly greater phylogenetic distance by predicting lethal genes in *S. cerevisiae*. We collected a *S. cerevisiae* phenotype dataset consisting of 6,075 genes (1,189 lethal, see Methods; Supplemental Table 4.1), 11 types of genomic data (**Table 4.1**) and assignments to 25 GO terms. Similar to earlier studies, the yeast model performed well in predicting yeast lethal genes (AUC-ROC = 0.82; **Fig. 4.6D**). Application of the *A. thaliana* model on yeast data performed reasonably well (AUC-ROC = 0.73), while an *S. cerevisiae* model on *A. thaliana* data performed worse (AUC-ROC = 0.65). The reduced performance in cross plant-fungal species predictions prompted us to investigate which features were meaningful for predictions in one species but not the other. The relative importance of features can be assessed according to a weight measure derived by the Support Vector Machine (SVM) classifier, which indicates the importance of a feature for predicting lethal (more positive weight) or non-lethal (more negative weight) genes. Between *A. thaliana* and *S. cerevisiae*, we found that 15 out of 36 features (42%) had opposing signs on their SVM weights (Supplemental Table 4.5), suggesting that, despite their importance for distinguishing lethal and non-lethal genes, these features have opposite contributions. For example, genes associated with the reproduction GO term tend to be phenotype-lethal in *A. thaliana*, but non-lethal in *S. cerevisiae*. When features with opposing correlations with lethality between the two species were removed, the performance improved in predicting *S. cerevisiae* lethal genes with the *A. thaliana* model (AUC-ROC from 0.73 to 0.75) and in predicting *A. thaliana* lethal genes with the *S. cerevisiae* model

(AUC-ROC from 0.65 to 0.73; **Fig. 4.6D**). While not as accurate as *A. thaliana*-*O. sativa* cross-species predictions, these results demonstrate that lethal phenotypes can be predicted between two species separated by 1.4 billion years of evolution. In addition, although many features of essential genes are similar between species, some features are predictive of lethal phenotypes in one species but of non-lethal phenotypes in another.

As lethal-phenotype genes tend to be well conserved, it may be expected that cross-species predictions would perform well. However, we should emphasize that the *A. thaliana*-rice cross-species predictions do not make use of any conservation-based features, and as a result, sequence conservation is unrelated to the performance of these cross-species predictions. For *A. thaliana*-yeast cross-species predictions, only one feature is related to gene conservation: presence as a core eukaryotic gene. While this is important for predictions (based on SVM feature weights; Supplemental Table 4.5), if cross-species predictions are performed using only the core eukaryotic gene feature, the AUC-ROC of predictions falls from 0.75 to 0.63 for *A. thaliana*-to-yeast cross-species predictions and from 0.73 to 0.55 for yeast-to-*A. thaliana* predictions. Further, if within-*A. thaliana* predictions are performed using only sequence conservation and evolutionary rate features, the AUC-ROC of essential gene predictions is 0.60 (compared to 0.81 with the full feature set). These results serve to further emphasize that neither protein conservation nor any single feature can sufficiently explain gene essentiality by itself, and that drawing upon a robust set of gene features provides a far more accurate prediction of essential genes.

To compare and contrast gene features important for essential gene prediction in all three species, we evaluated the importance of 29 features that are available in *A. thaliana*, rice, and yeast. Feature importance and relationship with phenotype lethality were assessed using the rank

and sign of SVM weights that are akin to the importance of a feature for predicting lethal (more positive weight) or non-lethal (more negative weight) genes (**Fig. 4.6E**). Features have generally similar importance for essential gene predictions in each species, although their relationship with lethality in each species is often not the same (i.e. opposing signs on SVM weights). We find five features that are relatively important for predictions and have the same sign in each species: median expression level, transcription factor activity, singleton status, cellular component organization, and signal transduction. These features likely represent characteristics shared by essential genes across kingdoms. Despite general similarity in feature importance, there are apparent species-specific features. For example, mitochondrial protein localization represents a feature important for predicting essential genes in plants but not in yeast. In addition, while response to endogenous stimulus and expression variation are relatively unimportant for predictions in plants, they are important for yeast predictions. Some species-specific features are not shared between more closely related taxa. For example, translation is important for lethal-phenotype predictions in *A. thaliana* but not in rice and yeast. For yeast, this may be due to a larger portion of translation-related genes being identified, including factors that are less central and essential to the process of protein synthesis. In rice, the smaller dataset of phenotype genes may not include many genes involved in the translation process, and therefore the term is not relevant to the predictions. Thus, we cannot rule out the possibility that some of the differences we found are due to differences in how functional categories are annotated across species. It will be more informative to examine lethal phenotype status on a gene-by-gene basis by asking whether and why orthologous genes are essential in one species but not the other.

Lastly, because few sequenced species have the extensive functional genomic resources found in *A. thaliana*, rice, and *S. cerevisiae*, we sought to determine if a model based only on

features that can be generated from a genome sequence (**Table 4.1**) can accurately predict lethal-phenotype genes. A machine learning model without input from expression and interactome features was generated for predicting *A. thaliana* lethal genes and performed with an AUC-ROC of 0.74. This result suggests that essential genes can be predicted with only sequence-based features. Interestingly, it has also been shown that sequence-based features are important in the identification of functional overlap between related genes (Chen et al., 2010). Our finding represents an important step that should prove useful in analyzing newly sequenced organisms that lack robust expression and interactome datasets.

CONCLUSION

We identified a set of genomic features that significantly correlate with genes that have lethal phenotypes when disrupted in *A. thaliana*. Similar to findings in yeast and mouse (Seringhaus et al., 2006; Yuan et al., 2012), these features can be used to predict genes with lethal phenotypes in plants. We also show that lethal-phenotype gene prediction models can be applied across species with reasonable performance. This provides strong evidence that the characteristics of essential genes can be defined based on genome sequence features and large-scale functional genomics data and, in some cases, are shared between species. We predict that a smaller percentage of *A. thaliana* genes are essential (10%) in comparison to *S. cerevisiae*, *M. musculus*, and *S. pombe* (18%, 19%, and 26%, respectively; Kim et al., 2010; Yuan et al., 2012). Considering the presence of multiple rounds of genome duplication in the past 100 million years in the *A. thaliana* lineage, the presence of duplicates is likely a major contributor to the difference. Nonetheless, we should emphasize that although individual characteristics can be used to distinguish between genes with lethal and non-lethal phenotypes, in many cases the effect sizes are rather small. Thus, despite the statistical significance, lethal-phenotype genes are more accurately predicted when many features are considered jointly.

Another consideration is that the cause-effect relationship between these features and phenotype lethality are not always obvious. For example, while lethal-phenotype genes tend to be single copy or have ancient duplicates, it is not known if stochastic gene loss simply results in essentiality for the remaining duplicate. Alternatively, there may be preferential loss of duplicates of essential genes, perhaps due to an inability to neo- or subfunctionalize many essential gene functions or because essential genes disproportionately function in dosage-

dependent processes. While our finding that lethal-phenotype genes tend to have similar duplicate retention and loss patterns across lineages is consistent with the preferential loss possibility, a more detailed analysis on this topic is warranted. Although the machine learning model performs well with high AUC-ROC, there also remains room for improvement in essential gene prediction. For our analysis, we restricted features to those in which we could provide *a priori* reasoning for association with phenotype lethality. Alternatively, a more data-driven approach that includes more genomic signatures without apparent relationships to phenotype lethality (e.g. histone marks, *cis*-regulatory complexity, or chromatin state) may allow the discovery of previously ignored factors. Another potential way to improve prediction is to focus on more narrowly-defined sets of essential genes. Because lethal phenotypes can result from the loss of a broad range of functions, we cannot necessarily expect all essential genes to possess the same sets of characteristics, as suggested by the significant association of lethal-phenotype genes with multiple characteristics but mostly with small effect sizes. As a result, it is reasonable to hypothesize that there exist distinct sets of essential genes where genes in each set share common characteristics. If this is the case, it will be intriguing to uncover the underlying reasons for the existence of such gene sets.

Taken together, our findings provide a detailed look at the factors predictive of gene phenotype lethality. Through a joint analysis of evolutionary (duplication, conservation) and functional (expression, *Ka/Ks*) characteristics of lethal-phenotype genes, this study advances our understanding of the evolution of essential genes. In addition, we provide genome-wide plant essential gene predictions and large-scale validation of cross-species lethal-phenotype predictions, building on earlier results focused on fungal or metazoan species and on smaller plant gene datasets. The predictive performance of our models highlights a promising avenue for

prioritizing candidate genes for large-scale phenotyping efforts in *A. thaliana*, particularly essential genes. The feasibility of cross-species predictions suggests that model plant phenotype data can be useful for the identification of essential genes in other plant species.

METHODS

Phenotype Data Sources

Descriptions of gene-based, loss-of-function mutant phenotypes in *Arabidopsis thaliana* were retrieved from three sources: (1) a published phenotype dataset (Lloyd and Meinke, 2012), (2) the Chloroplast 2010 Database (Ajjawi et al., 2010; Savage et al., 2013), and (3) the RIKEN Phenome database (Kuromori et al., 2006). Phenotype descriptions for genes in *Oryza sativa* (rice) were gathered from four sources: (1) a published phenotype dataset (Lloyd and Meinke, 2012), (2) literature search and manual curation (search terms: rice, lethal, mutant, phenotype, null, and knockout), (3) Oryzabase (Kurata and Yamazaki, 2006), and (4) Gramene (Monaco et al., 2014). *Saccharomyces cerevisiae* (yeast) phenotype annotations were obtained from the Saccharomyces Genome Database (http://www.yeastgenome.org; Cherry et al., 2012). If a gene had conflicting phenotype assignments from multiple sources, the lethal phenotype description was given priority. For yeast, phenotypes annotated to the "inviable" phenotype ontology term were considered lethal, while those annotated to the "viable" term were considered non-lethal. Only phenotypes associated with a null allele were included. An independent set of 60 *A. thaliana* phenotype genes were identified from recently-published literature by searching for articles in the PubMed database that included the keywords "Arabidopsis" and "lethal" and were published in 2012 or 2013. This independent dataset includes 24 genes for which a homozygous single-gene mutant was viable and included in at least the attempted construction of a double knockout mutant for the GABI-DUPLO project (Bolle et al., 2013).

Gene Ontology Functional Annotation

Gene Ontology (GO) gene annotations for *A. thaliana* and yeast were downloaded from the GO database (http://www.geneontology.org/), and version 7 rice annotations were downloaded from the Rice Genome Annotation Project (Kawahara et al., 2013) (http://rice.plantbiology.msu.edu). *A. thaliana* and yeast annotations were mapped to the plant slim ontology using the map2slim program in the GOperl package (http://search.cpan.org/~cmungall/go-perl/). For *A. thaliana*, only gene-GO terms associated with experimental or computational evidence codes were utilized, while those based only on curation and author statements were excluded. Of the 97 terms in the plant GO slim subset, three were excluded because they are the root terms (biological process, molecular function, cellular component), 59 were excluded because they are not significantly over- or underrepresented in lethal genes, one was excluded because it is associated with few *A. thaliana* phenotype genes (<1%), and five were excluded because they are highly overlapping in gene membership with another significantly enriched term (Pearson's Correlation Coefficient, PCC $\geq$ 0.50). Among overlapping terms, one representative term was chosen based on the lowest adjusted *p*-value from FETs, except in the case of pairwise overlap between the "response to stress" term and "response to biotic stimulus"/"response to abiotic stimulus," where "response to stress" was removed despite having a lower *p*-value to maintain the distinction in functional responses to biotic and abiotic environmental factors. Because 151 of 329 genes in the embryo development term are included in our phenotype dataset, it was excluded to prevent ascertainment bias. Plant GO slim terms plastid, embryo development, and pollination were excluded from analysis involving yeast data. GO enrichment analysis using the full list of terms beyond the plant slim subset was also performed in *A. thaliana* to determine enrichment of both highly expressed

(genes with the top 1/3 expression levels) and weakly expressed (genes with the bottom 1/3 expression levels) lethal genes (Supplemental Table 4.2). In all GO analyses, *p*-values were adjusted for multiple testing based on the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995).

Evolutionary Rate Calculations and Analysis of Duplicates and Pseudogenes

Paralogs in *A. thaliana*, *O. sativa*, and *S. cerevisiae* and homologs between *A. thaliana* and five different plant species (*Arabidopsis lyrata*, *Populus trichocarpa*, *Vitis vinifera*, *O. sativa*, and *Physcomitrella patens*) were identified with OrthoMCL (inflation parameter = 1.5). Protein sequences for *A. thaliana* were downloaded from The Arabidopsis Information Resource (Version 10; www.arabidopsis.org), sequences for *S. lycopersicum* were downloaded from the Sol Genomics Network (Version 2.4; www.solgenomics.net), sequences for *O. sativa* were downloaded from the Rice Genome Annotation Project (Version 7; rice.plantbiology.msu.edu), sequences for *S. cerevisiae* were downloaded from the Saccharomyces Genome Database (www.yeastgenome.org), and sequences for *A. lyrata*, *P. trichocarpa*, and *V. vinifera* were downloaded from Phytozome (Version 9; www.phytozome.net).

In **Fig. 4.1A**, the paralog copy number for each *A. thaliana* gene equaled the size of the OrthoMCL cluster the gene in question resided in. In **Fig. 4.1B**, to identify orthologs between *A. thaliana* and *O. sativa* and to assess duplicate retention and loss, a gene-species tree reconciliation approach was used. First, protein sequences of genes in each *A. thaliana*-*O. sativa* OrthoMCL cluster were aligned using MUSCLE (Edgar, 2004). Ten maximum likelihood trees for each aligned cluster were built using RAxML (Stamatakis, 2014) to identify the tree with the highest likelihood. The trees were midpoint rooted with retree in the PHYLIP package and parsed with Notung (Chen et al., 2000) to identify duplication and speciation nodes in the gene

trees. A group of genes sharing a speciation node in a gene tree were regarded as an orthologous group.

Rates of synonymous (*Ks*) and nonsynonymous (*Ka*) substitutions were calculated between homologous gene pairs using the yn00 package in PAML (Yang, 2007). Highly similar or dissimilar sequence pairs (*Ks* < 0.005 and *Ks* > 3, respectively) were excluded from further analyses. In cross-species *Ka/Ks* calculations, the median *Ka/Ks* value between each *A. thaliana* gene and genes from other species in the same OrthoMCL cluster was used as a representative value. In *A. thaliana*, *O. sativa*, and *S. cerevisiae*, a paralogous pair for *Ks* analysis only (e.g. **Fig. 4.2A**) was defined by identifying a gene and its top-scoring match in BLAST similarity searches (Altschul et al., 1990). Nucleotide diversity between 80 *A. thaliana* accessions was calculated according to an earlier study (Moghe et al., 2013). Genes with paralogs produced in the α or $\beta\gamma$ whole genome duplication events were identified by Bowers et al. (2003). Two genes were defined as a tandem duplicate pair if they have a BLASTP E-value < 1e-10 and are no more than 10 genes apart. Pseudogenes were identified through the pipeline described by Zou et al. (2009). Clusters of orthologous genes were downloaded from the National Center for Biotechnology Information (Tatusov et al., 2003). Core eukaryotic genes were defined as genes present in clusters that included at least one gene from each of the seven species in the analysis (*A. thaliana*, *C. elegans*, *D. melanogaster*, *Encephalitozoon cuniculi*, *Homo sapiens*, *S. cerevisiae*, and *S. pombe*).

BLAST similarity searches were performed between *A. thaliana* protein sequences and the protein sequences of 34 other plant species present in Phytozome v9, including 26 dicotyledonous, 6 monocotyledonous, and 2 other embryophyte species. Similarity searches were also performed between *A. thaliana* and 8 fungal species (*Aspergillus nidulans*, *Coprinopsis*

cinerea, *Cryptococcus neoformans*, *Fusarium oxysporum* f. sp. *lycopersici*, *Neurospora crassa*, *Puccinia graminis* f. sp. *tritici*, *S. cerevisiae*, and *S. pombe*) and 8 metazoan species (*C. elegans*, *Ciona savignyi*, *Danio rerio*, *D. melanogaster*, *Gallus gallus*, *H. sapiens*, *M. musculus*, and *Xenopus tropicalis*). Fungal and metazoan protein sequence annotations were retrieved from FungiDB (www.fungidb.org) and Ensembl (www.ensemblgenomes.org), respectively.

Expression Data Sources and Processing

The AtGenExpress development microarray data (Schmid et al., 2005) was downloaded from the Weigel lab (http://www.weigelworld.org/resources/microarray/AtGenExpress/). Samples involving data from mutant plants were removed and the median value of the replicates was used as a representative expression value for each gene. Pre-processed RNA-seq data from *O. sativa* were downloaded from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/expression.shtml). Data for testing differential gene expression were excluded from further analyses. For *S. cerevisiae*, a time-course cell-cycle expression dataset was used (Orlando et al., 2008). Median and maximum expression and variation of expression were calculated for each gene in the datasets. Expression variation was represented by median absolute deviation divided by the median as it is a measure that does not require a normality assumption. For *O. sativa* expression breadth was calculated by counting the number of tissues in which expression was greater than zero fragments per kilobase of exon per million reads mapped (FPKM). For AtGenExpress data, a series of thresholds ($\log_2$ intensity = 4~10) for calling whether a gene was expressed or not was tested. The $\log_2$ intensity threshold of 4 resulted in the lowest *p*-value (KST) from testing if the distributions of number of datasets a gene was expressed in were significantly different between lethal and non-lethal genes and was used in machine learning analysis. Expression correlations between a gene and putative paralogs

(defined as genes belonging to the same OrthoMCL cluster) were evaluated using PCC and the maximum PCC with a paralogous gene was reported in **Fig. 4.3A** and used in machine learning analysis.

Network Analysis

Co-expression modules in the AtGenExpress expression data were identified through K-means clustering with K = 5~2000. Clusters generated with K=2000 resulted in the lowest *p*-value from KSTs that tested whether the co-expression module size distributions for lethal and non-lethal genes were significantly different and were used in subsequent analysis. Pairwise expression correlations (PCC) between all genes for which expression data was available were calculated. A gene pair with a PCC of 0.86 (99^{th} percentile of all pairwise comparisons) was considered co-expressed. The AraNet gene network dataset (Lee et al., 2010) was downloaded from http://www.functionalnet.org/aranet/ and any gene pair with a log likelihood score ≥ 1 was considered to be functionally related for our analyses. *S. cerevisiae* gene network data generated by Costanzo et al. (2010) were retrieved from the *Saccharomyces* Genome Database. Protein-protein interaction data from the Arabidopsis Interactome Mapping Consortium (Arabidopsis Interactome Mapping Consortium, 2011) were retrieved from the supplemental data associated with the publication. Self-interactions and interactions involving a mitochondrial or plastid gene were excluded from analysis.

Machine learning predictions

Phenotype predictions were carried out using machine learning algorithms implemented in the Waikato Environment for Knowledge Analysis software (WEKA; Hall et al., 2009). The features we used are shown in **Table 4.1**, and a complete matrix of all genes and feature values is available in Supplemental Table 4.3. We first tested if the targeted features were correlated with

one another through pairwise Spearman rank correlation analysis. We found that 95% and 99% of feature pairs show a correlation of ≤0.22 and ≤0.40, respectively. This indicated that there was no extensive overlap and thus all features were used in subsequent analysis.

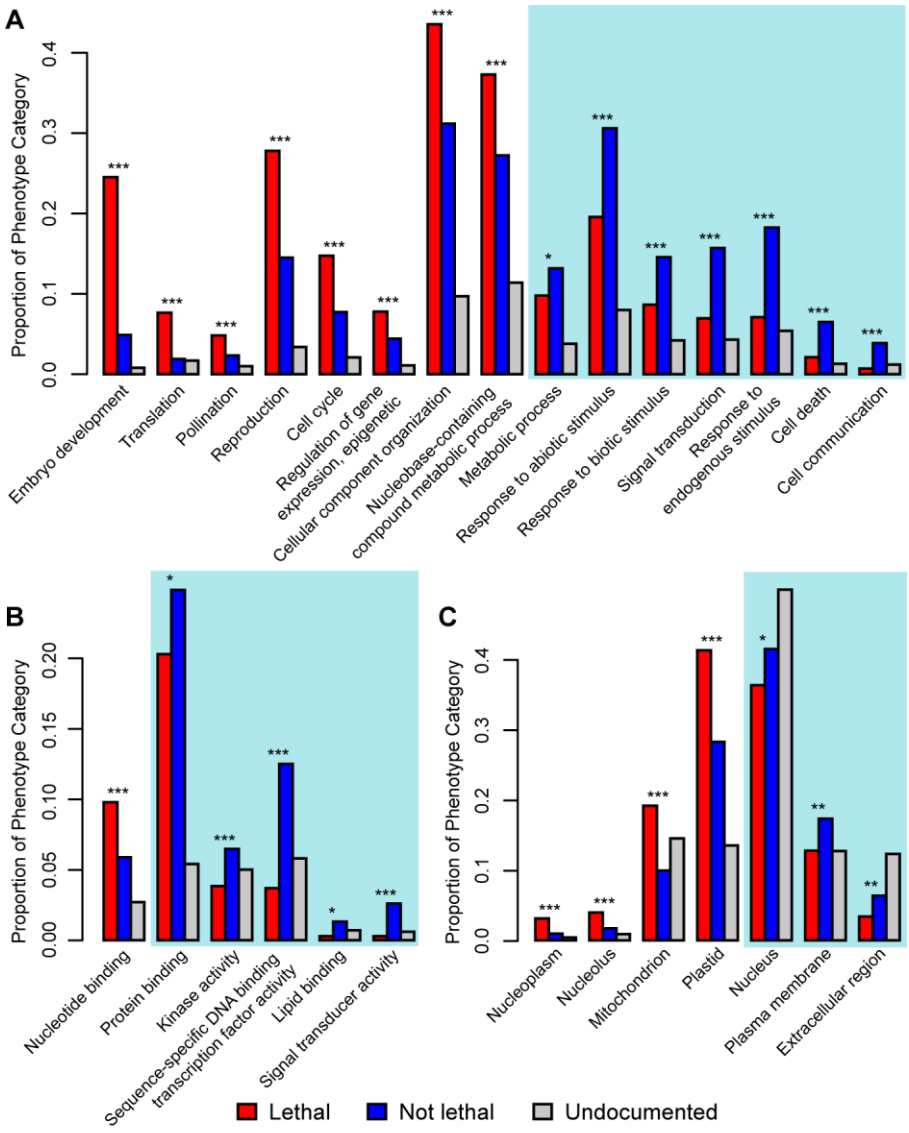
Six classifiers capable of handling binary, numeric, and missing data were tested: J48 decision tree, logistic regression, naïve Bayes, naïve Bayes tree, Random Forest, and sequential minimal optimization support vector machine (SMO-SVM). Ten-fold cross validation was performed for all machine learning runs, except for those involving rice phenotype data, where a low number of instances necessitated two-fold cross-validation. A grid search was implemented to identify best-performing parameters. Grid searches for each classifier included a parameter for the proportion of lethal-to-non-lethal instances to include in each round of predictions. For the Random Forest classifier, a model trained with a 1-to-1 ratio of lethal to non-lethal genes (AUC-ROC=0.8) performed similarly as models trained with a dataset containing other ratios of lethal to non-lethal genes (maximum AUC-ROC = 0.81). Additional parameters for the following classifiers were also examined: J48 decision tree, pruning confidence; logistic regression, ridge; SMO-SVM, complexity constant; random forest, number of random features to consider. The –M option was invoked in SMO-SVM runs, which provides a confidence score between 0 and 1 with predictions and was used as the "lethal-phenotype score." For random forest, 100 trees were built during the parameter search phase. All other parameters were default values. Best-performing parameter sets for each classifier were determined by AUC-ROC, which was calculated and visualized using the ROCR package (Sing et al., 2005). Models were built using best-performing parameter sets and randomly shuffled lethal and non-lethal gene labels. The AUC-ROC values from 100 iterations of gene label shuffling for all six classifiers ranged from 0.45 to 0.55 with a median of 0.5.

To predict whether an undocumented gene was lethal, the lethal-phenotype score resulting in the highest *F*-measure [harmonic mean of precision (proportion of predictions correct) and recall (proportion of true positives predicted)] was used as the threshold to call potential lethal-phenotype genes. Features most important to the prediction analysis were evaluated by leave-one-out analysis, wherein features were excluded one at a time, and effects on performance in comparison to a full feature set were recorded. To evaluate how many features are required to have comparable performance as the full model, 57 models were built and evaluated with increasing numbers of features. The order in which features were included was based on SVM weight, where features with the highest absolute weight were added first. During cross-species predictions, numeric data were discretized into quantiles (for example, data points in the lowest quantile were set to 1, while data points in the highest quantile were set to 5), to ensure that data were present in similar ranges and distinctions within data drawn by the machine learning algorithms could be applied to data from another species.

ACKNOWLEDGEMENTS

I thank Sahra Uygun and David Hufnagel for providing illustrious data without which this project would have crumbled.

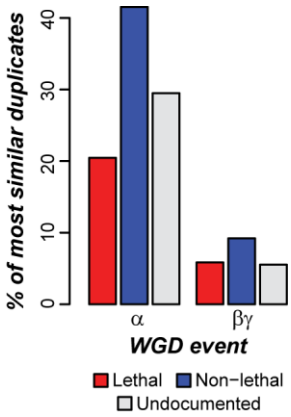
APPENDIX



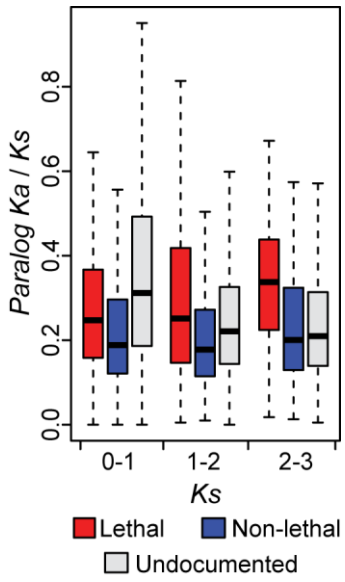
Supplemental Figure 4.1 Over- and under-representation of phenotype genes in Gene Ontology categories. Proportions of lethal, non-lethal, and undocumented genes present in (A) biological processes, (B) molecular functions, or (C) cellular components categories. Only categories that have significantly over- or under-represented numbers of lethal genes relative to non-lethal genes are shown, based on adjusted *p*-values of Fisher exact tests where *, **, and *** indicate α = 0.05, 0.01, and 0.001, respectively. Categories in which lethal genes are

Supplemental Figure 4.1 (cont'd)

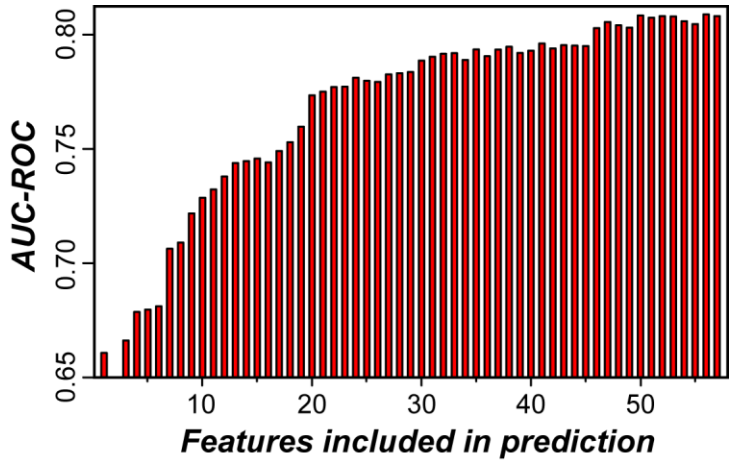
underrepresented are indicated by a cyan background.



Supplemental Figure 4.2 Proportions of most similar paralogs produced in whole-genome duplication events. Proportion of lethal, non-lethal, and undocumented genes with close paralogs (most similar by BLASTP e-value) generated in the α or βγ WGD events.



Supplemental Figure 4.3 Evolutionary rates between paralogs. Box plots of *Ka*/*Ks* values between paralogous pairs of lethal, non-lethal, or undocumented genes. Gene pairs are non-exclusive with regard to phenotype category. Distributions of ratios are shown as they relate to the rate of synonymous substitutions (*Ks*) between a gene pair, where low *Ks* values indicate recent duplicates and high *Ks* values indicate older duplicates.



Supplemental Figure 4.4 Performance of essential gene predictions with increasing numbers of features. AUC-ROCs of 57 Random Forest models using 1 to 57 features detailed in **Table 4.1**. Features were included in the order of highest-to-lowest absolute SVM weight.

CHAPTER 5: CONCLUSIONS

Assessing the functionality of genomic sequences

The research presented in Chapters 2 and 3 focused on defining functional regions in genomes, with an emphasis on intergenic transcribed regions (ITRs). This work was guided by three aims: 1) defining functional regions in genomes by generating function prediction models, 2) prediction of likely-functional ITRs through application of function prediction models, and 3) detailed evaluation of the evolutionary histories of ITRs. I find that functional genome sequences can be successfully defined by integrating genetic, biochemical, and evolutionary evidence, as machine learning models accurately distinguish between benchmark functional (phenotype and RNA genes) and non-functional (pseudogenes and random intergenic) sequences in both *A. thaliana* and rice. Expression data proved highly informative for predictions, as combining 24 transcriptional activity features in *A. thaliana* provided predictions with an AUC-ROC performance of 0.97, similar to results in human (Tsai et al., 2017). Nevertheless, transcript evidence alone was a poor predictor of functionality, with 80% of *A. thaliana* pseudogenes being expressed. This underscores that the presence of a transcript alone should not be used as evidence to indicate a sequence is functional.

Function prediction models were applied to ITRs and predict 2,754 and 4,427 rice and *A. thaliana* ITRs as likely functional, respectively. These are likely highly enriched in functional sequences and should be considered strong candidates for future experimental studies. However, only ~40% of ITRs in each species are predicted as functional, indicating that ITRs may primarily represent non-functional sequences. In addition, I anticipated that if ITRs generally represented non-functional sequence, the proportion of intergenic space covered by ITR transcript fragments would increase with genome size. When this hypothesis was tested, I observed that coverage of intergenic regions with transcript evidence has a positive and

significant correlation with genome size across 15 flowering plant species, while coverage of genic expression had no correlation. Further, evaluation of the evolutionary histories of ITRs finds that these sequences are predominantly species-specific that generally lack long-term duplicate retention. In addition, expression is rarely present in ITR duplicates or in the rare cases when an ITR ortholog can be identified, indicating that the expression state of ITRs is transient and unstable. Overall, these three lines of evidence suggest most ITRs are not under selection, and thus I conclude that intergenic expression is primarily the result of transcriptional noise. Because of this, I recommend that the null hypothesis for the functionality of intergenic transcripts is that they represent the product of noisy expression, which can be overturned with compelling experimental evidence for the functionality of a sequence.

Despite the overall success of function prediction models, interpretation of two sets of predictions remains challenging. First, function prediction models classified 18% of annotated protein-coding genes in *A. thaliana* and 15% of transcribed regions overlapping annotated exons in rice as non-functional. However, it is unclear whether these predictions should be interpreted as false negatives among truly functional sequences or true negatives among annotated genes that are decaying and *en route* to pseudogene status. Second, ITRs in *A. thaliana* and rice that were predicted as functional tend to be close to annotated genes. Given the proximity to genic regions, the influence of the open and active chromatin regions surrounding genes may represent a confounding factor. Additionally, ITRs near genes could be unannotated extensions of genes. Advances in read lengths produced during sequencing could yield evidence to determine whether and how often ITRs are extensions of nearby genes. Analysis of cap analysis gene expression (CAGE) datasets could also be informative. Last, experimental validation via loss-of-function

analysis would be particularly useful to assess annotated genes predicted as non-functional and gene-proximal ITRs predicted as functional, but is currently lacking.

Pseudogenes functional classifications from prediction models that included benchmark RNA genes resulted in high false positive rates (FPRs; ~30% in *A. thaliana*, ~40% in rice). A small subset of pseudogenes likely represent truly functional sequences (Poliseno et al., 2010; Karreth et al., 2015), however the majority of pseudogenes are anticipated to be neutrally-evolving, non-functional sequences (Li et al., 1981; Svensson et al., 2006). Two issues likely contribute to high FPRs: 1) lack of high-quality benchmark RNA genes and 2) limited feature sets. For the first issue, there are only rare examples of RNA genes with documented loss-of-function phenotypes in plants. Instead, I turned to community-based curation and utilized "high-confidence" miRNA annotations. Without strong evidence provided by phenotype data, it is possible that these miRNAs contain a substantial proportion of false positive gene annotations, contributing to false positive predictions in pseudogenes. For the second issue, biochemical features utilized were associated with transcription, as DNA methylation, chromatin accessibility, and histone marks act as transcriptional regulators. Thus, if transcription characteristics are similar between sequence classes, functional predictions will be challenging. The identification of feature sets unrelated to transcription could provide information to distinguish between RNA genes and pseudogenes. One additional consideration is that the processing of existing features could be expanded. For example, identification of DNA methylation and histone mark profiles across the length of sequences could prove more informative than basic intensity or coverage measures that have been utilized thus far.

It is unlikely that prediction models in their current state are able to distinguish between the functionality of a DNA sequence and the functionality of transcript. For example, *trans*-

acting regulatory enhancers have been associated with transcript evidence (referred to as eRNAs) (Lam et al., 2014). Despite a regulatory role for the enhancer sequence, it is unknown whether eRNA transcripts are functional. Instead, the open chromatin state of enhancers may increase the likelihood of noisy transcription occurring within the region. Further, the physical act of transcription may be important for maintaining the chromatin state of a sequence (Gerstein et al., 2007). In both cases, a DNA sequence may be functional while a transcript derived from the sequence is not. This highlights the limitations of prediction models that have been established to classify likely-genic sequences. Future work to adapt prediction models toward identifying regulatory and other classes of potentially-functional genome sequences would represent an important extension.

Given the successful prediction of functional sequences in two distinct plant groups (dicotyledonous and monocotyledonous plants) and metazoans (Tsai et al., 2017), it is likely that the data integration framework described in Chapters 2 and 3 will be applicable in most biological systems. An intriguing next question is whether function prediction models can be applied across species. Currently, I have helped to generate datasets of similar function-related features in *A. thaliana*, rice, and human. Further, there is extensive transcriptional activity, epigenetic, and evolutionary data are available for numerous model systems (Gerstein et al., 2010; Roy et al., 2010; Cherry et al., 2012; Mouse ENCODE Consortium et al., 2012). Using the wealth of available data, prediction models can be generated and validated using analogous feature sets and benchmark sequences in multiple systems. It would then be straightforward to apply function prediction models across species (see Chapter 4 for an example using lethal phenotype gene predictions). The potential for cross-species predictions might be most exciting for models generated with only transcriptional activity features, which provided highly accurate

predictions. Successful transcription-only cross-species predictions would open up any species with a sequenced genome and moderate amount of transcriptome data as a feasible target for prediction models generated in model systems.

Predictions of mutant phenotypes

The research described in Chapter 4 aims to assess the role of genes (i.e. functional genome sequences) by predicting of loss-of-function mutant phenotypes. Phenotype predictions were focused on essential genes, i.e. those with lethal mutant phenotypes, as this set of genes has been a target of historical, focused study (Meinke et al., 2008). Based on this analysis, I found that essential genes shared characteristics that distinguished them from genes with non-lethal mutant phenotypes. Essential genes tend to be single-copy or derived from older duplication events, highly and broadly expressed, strongly conserved, and highly connected in gene networks. Machine learning-based integration of 57 total features provided accurate predictions of genes with lethal and non-lethal mutant phenotypes (AUC-ROC =0.81). Thus, machine learning approaches represent useful methods to prioritize genes for large-scale phenotyping analysis. Perhaps most intriguing, essential gene prediction models could be successfully applied across species, indicating that essential genes in different species exhibit similar characteristics. Together with successful predictions of essential genes based solely on sequence-based features (i.e. without expression or gene network data), cross-species predictions highlight the potential for a machine learning approach to facilitate translation of phenotype data in model systems to non-model systems, particularly for essential genes.

Essential gene predictions, while reasonably accurate, exhibit substantial false positive and false negative rates. One issue is likely that essential genes represent a heterogeneous group of genes. Genes with gametophyte-, embryo-, and seedling-lethal phenotypes, for example, may

exhibit distinct characteristics. Predictions may also be improved by employing additional features, particularly biochemical features such as DNA methylation and histone mark patterns, which were largely excluded. In addition, essential genes tend to be connected with one another within gene networks (Lee et al., 2010), but any features assessing putative interaction with known essential genes were not implemented. One last issue with the predictions in Chapter 4 is that it is not clear whether essential gene predictions are predicting lethality specifically or phenotype severity generally. To test this, it would be informative to compare the lethal phenotype prediction scores of genes that exhibit obvious visible phenotypes to those that are associated with subtle biochemical or cellular phenotype.

Building on the success of essential gene predictions, the next step will be to determine what other phenotypes can be accurately predicted. A foundational issue for such a task is identifying features that can effectively distinguish sets of functionally-related genes with specificity. Lethality represents a broad phenotype category and therefore properties such as median expression level, duplication patterns, and evolutionary histories were informative. However, similar features would likely provide minimal information for more specific phenotype categories. Instead, gene interactions may prove critical for such predictions. Analysis based on unsupervised clustering has established that genes functioning in similar pathways frequently exhibit co-expression (Eisen et al., 1998; Spellman et al., 1998; Lee et al., 2004a; Uygun et al., 2016). A next step would be to overlay supervised classification techniques on expression clustering to determine the accuracy and limitations in which genes with related functions can be classified through co-expression. One additional possibility could be to adapt function prediction models described in Chapters 2 and 3 toward predicting tissue- or cell type-specific functionality,

which could be feasible if appropriate functional genomics datasets are available. However, such an approach is untested to date.

Plants feature an abundance of duplication events. As a result, the presence of functionally-redundant paralogs frequently masks the consequences of gene disruption and hampers phenotyping efforts. For example, 10% of *A. thaliana* genes were predicted as essential, compared to at least 18% of genes in *Saccharomyces cerevisiae*, *Mus musculus*, and *Schizosaccharomyces pombe*. To help counteract pervasive duplication, efforts are underway to catalog phenotypes resulting from the disruption of gene pairs (Bolle et al., 2013). Computational methods could provide additional information needed to streamline phenotyping of genes that exhibit functional overlap. Initial machine learning-based predictions of gene pairs that are likely functionally redundant have been established (Chen et al., 2010). Extensions of these models that incorporate novel features, particularly biochemical signatures across duplicates, and semi-quantitative redundancy classes (e.g. fully-redundant vs. partially-redundant) could be provided. Overlaying predictions of redundancy on top of phenotype predictions could prove to be a highly effective approach.

Concluding remarks

Overall, the research described in this dissertation highlights successful computational approaches to merge heterogeneous sets of data and produce accurate biological classifications. These classifications serve not only to prioritize candidates for future experimental characterization, but provide meaningful biological insights into the characteristics that define functional genome sequences and gene essentiality. Such computational approaches work in harmony with experimental approaches by directing the production of high quality data. In turn, these data allow for synergistic refinement and expansion of computational approaches. With

increasingly large stores of data produced through the use of modern sequencing and biotechnological techniques, computational approaches will serve a key role in the future of the biological study.

REFERENCES

REFERENCES

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature **467**: 1061–73

Acencio ML, Lemke N (2009) Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics **10**: 290

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**: 529–40

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

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

Arabidopsis Interactome Mapping Consortium (2011) Evidence for network evolution in an *Arabidopsis* interactome map. Science **333**: 601–7

Bailey J a, Gu Z, Clark R a, Reinert K, Samonte R V, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science **297**: 1003–1007

van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "dark matter" transcripts are associated with known genes. PLoS Biol. doi: 10.1371/journal.pbio.1000371

van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2011) Response to "The reality of pervasive transcription." PLoS Biol **9**: 7–10

Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J (2011) Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets. Plant Cell **23**: 3101–16

Bazykin GA, Kochetov A V. (2011) Alternative translation start sites are conserved in eukaryotic genomes. Nucleic Acids Res **39**: 567–577

Beadle GW, Tatum EL (1941) Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci **79**: 499–505

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci USA **107**: 18724–18728

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc **57**: 289–300

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 K V, 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

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF a, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res **14**: 708–15

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

Bolle C, Huep G, Kleinbölting N, Haberer G, Mayer K, Leister D, Weisshaar B (2013) GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*. Plant J **75**: 157–71

Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol **14**: 292–9

Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Paro R, Perrimon N (2004) Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. Science **303**: 832–5

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**: 433–438

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

Bulger M, Groudine M (2010) Enhancers: The abundance and function of regulatory sequences beyond promoters. Dev Biol **339**: 250–257

Cai J, Zhao R, Jiang H, Wang W (2008) De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics **179**: 487–496

Campbell MS, Holt C, Moore B, Yandell M (2014a) Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinforma. doi: 10.1002/0471250953.bi0411s48

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun

R, Jiao D, Lawrence CJ, et al (2014b) MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. Plant Physiol **164**: 513–524

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18**: 188–96

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

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al (2012) Proto-genes and de novo gene birth. Nature **487**: 370–374

Chen F, D'Auria JC, Tholl D, Ross JR, Gershenzon J, Noel JP, Pichersky E (2003) An Arabidopsis thaliana gene for methylsalicylate biosynthesis, identified by a biochemical genomics approach, has a role in defense. Plant J **36**: 577–588

Chen H-W, Bandyopadhyay S, Shasha DE, Birnbaum KD (2010) Predicting genome-wide redundancy using machine learning. BMC Evol Biol **10**: 357

Chen K, Durand D, Farach-colton M (2000) Notung: a program for dating gene duplications and optimizing gene family trees. Bioinformatics **7**: 429–447

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al (2012) *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucleic Acids Res **40**: D700-5

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, et al (2010) The genetic landscape of a cell. Science **327**: 425–431

Crick FHC (1963) On the Genetic Code. Science **139**: 461–464

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res **16**: 738–49

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

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, Jiang N, Robin Buell C (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J **71**: 492–502

Davidson RM, Hansey CN, Gowda M, Childs KL, Lin H, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Jiang N, et al (2011) Utility of RNA Sequencing for Analysis of

Maize Reproductive Transcriptomes. Plant Genome J **4**: 191

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA **110**: 5294–300

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

Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, et al (2010) Genotype to phenotype: a complex problem. Science **328**: 469

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA **102**: 14338–43

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA **96**: 4482–4487

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

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

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95**: 14863–14868

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

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al (2016) The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res **44**: D279–D285

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature **391**: 806–811

Firon A, Villalba F, Beffa R, Enfert C (2003) Identification of essential genes in the human fungal pathogen *Aspergillus fumigatus* by transposon mutagenesis. Eukaryot Cell **2**: 247–256

Force a, Lynch M, Pickett FB, Amores a, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–45

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

Furey T, Cristianini N, Duffy N (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16**: 906–914

Gaertner FH, Cole KW (1977) A cluster-gene: Evidence for one gene, one polypeptide, five enzymes. Biochem Biophys Res Commun **75**: 259–264

Ganko EW, Meyers BC, Vision TJ (2007) Divergence in expression between duplicated genes in arabidopsis. Mol Biol Evol **24**: 2298–2309

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res **17**: 669–681

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science **330**: 1775–87

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Iii CAH, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA **103**: 425–430

Golling G, Amsterdam A, Sun Z, Antonelli M, Maldonado E, Chen W, Burgess S, Haldi M, Artzt K, Farrington S, et al (2002) Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. Nat Genet **31**: 135–40

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40**: D1178-86

Graur D (2017) An Upper Limit on the Functional Fraction of the Human Genome. Genome Biol Evol **9**: 1880–1885

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall R a, 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**: 578–90

Griffiths PE, Stotz K (2006) Genes in the postgenomic era. Theor Med Bioeth **27**: 499–521

Gu Z, Nicolae D, Lu HHS, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet **18**: 609–613

Gu Z, Steinmetz L, Gu X, Scharfe C, Davis R, Li W (2003) Role of duplicate genes in genetic robustness against null mutations. Nature **42**: 63–66

Guil S, Esteller M (2012) Cis-acting noncoding RNAs: friends and foes. Nat Struct Mol Biol **19**: 1068–1075

Gulko B, Gronau I, Hubisz MJ, Siepel A (2014) Probabilities of Fitness Consequences for

Point Mutations Across the Human Genome.

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

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn **46**: 389–422

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc **8**: 1494–1512

Hall M, Frank E, Holmes G (2009) The WEKA data mining software: an update. ACM SIGKDD **11**: 10–18

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 USA **110**: 2395–400

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol **148**: 993–1003

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

Harris RS (2007) Improved Pairwise Alignment of Genomic DNA. Pennsylvania State Univ. Thesis

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

Hilario M, Kalousis A, Müller M, Pellegrini C (2003) Machine learning approaches to lung cancer prediction from mass spectra. Proteomics **3**: 1716–9

Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, et al (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci **104**: 6478–6483

Hittinger CT, Carroll SB (2007) Gene duplication and the adaptive evolution of a classic

genetic switch. Nature **449**: 677–681

Ho TK (1995) Random decision forests. Proc. 3rd Int. Conf. Doc. Anal. Recognit. pp 278–282

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

Hubisz MJ, Pollard KS, Siepel A (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform **12**: 41–51

Hughes AL (1994) The Evolution of Functionally Novel Proteins after Gene Duplication. Proc R Soc B, Biol Sci **256**: 119–124

Hupalo D, Kern AD (2013) Conservation and functional element discovery in 20 angiosperm plant genomes. Mol Biol Evol **30**: 1729–44

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

Jacob F (1977) Evolution and tinkering. Science **196**: 1161

Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature **411**: 41–2

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

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE (2001) Positive selection of a gene family during the emergence of humans and African apes. Nature **413**: 514–519

Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J (2009) tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. Nucleic Acids Res **37**: 159–162

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res **20**: 1313–1326

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, et al (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature **421**: 231–7

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

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice **6**: 4

Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S (2013) Function of alternative splicing. Gene **514**: 1–30

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 USA **111**: 6131–8

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

Kim DU, Hayles J, Kim D, Wood V, Park HO, Won M, Yoo HS, Duhig T, Nam M, Palmer G, et al (2010) Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. Nat Biotechnol **28**: 617–623

King RD, Sternberg MJ (1990) Machine learning approach for the prediction of protein secondary structure. J Mol Biol **216**: 441–57

Knowles DG, Mclysaght A, Knowles DG, Mclysaght A (2009) Recent *de novo* origin of human protein-coding genes. Genome Res **19**: 1–9

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al (2003) Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA **100**: 4678–83

Kochetov A V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. BioEssays **30**: 683–691

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

Kurata N, Yamazaki Y (2006) Oryzabase. An integrated biological and genome information database for rice. Bioinformatics **140**: 12–17

Kuromori T, Takahashi S, Kondou Y, Shinozaki K, Matsui M (2009) Phenome analysis in plant species using loss-of-function and gain-of-function mutants. Plant Cell Physiol **50**: 1215–31

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**: 640–51

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

Lam MTY, Li W, Rosenfeld MG, Glass CK (2014) Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci **39**: 170–182

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res **40**: D1202-10

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

Law M, Childs KL, Campbell MS, Stein JC, Olson AJ, Holt C, Panchy N, Lei J, Jiao D, Andorf CM, et al (2015) Automated Update, Revision, and Quality Control of the Maize Genome Annotations Using MAKER-P Improves the B73 RefGen_v3 Gene Models and Identifies New Genes. Plant Physiol **167**: 25–39

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004a) Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Res **14**: 1085–1094

Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. Nat Biotechnol **28**: 149–56

Lee I, Date S V, Adai AT, Marcotte EM (2004b) A probabilistic functional network of yeast genes. Science **306**: 1555–8

Lehti-Shiu MD, Uygun S, Moghe GD, Panchy N, Fang L, Hufnagel DE, Jasicki HL, Feig M, Shiu S-H (2015) Molecular Evidence for Functional Divergence and Decay of a Transcription Factor Derived from Whole-Genome Duplication in *Arabidopsis thaliana*. Plant Physiol **168**: 1717–1734

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ (2006) Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci **103**: 9935–9939

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

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 W, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature

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

Liu Q, Liu H, Wen J, Peterson PM (2014) Infrageneric phylogeny and temporal divergence of Sorghum (Andropogoneae, Poaceae) based on low-copy nuclear and plastid sequences. PLoS One. doi: 10.1371/journal.pone.0104933

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

Lloyd JP, Tsai ZT, Sowers RP, Panchy NL (2017) Defining the functional significance of intergenic transcribed regions based on heterogeneous features of phenotype genes and pseudogenes. bioRxiv 1–51

Lynch M, Conery JS (2000) The Evolutionary Fate and Consequences of Duplicate Genes. Science **290**: 1151–1155

Lynch M, Conery JS (2003) The Origins of Genome Complexity. Science **302**: 1401–1404

Lynch M, Force a (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–73

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci **102**: 5454–5459

Makalowska I, Lin CF, Makalowski W (2005) Overlapping genes in vertebrate genomes. Comput Biol Chem **29**: 1–12

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

Marais DL Des, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature **454**: 762–765

Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM (1999) A Biochemical Genomics Approach for Identifying Genes by the Activity of Their Products. Science **286**: 1153–1155

Massa AN, Wanjugi H, Deal KR, O'Brien K, You FM, Maiti R, Chan AP, Gu YQ, Luo MC, Anderson OD, et al (2011) Gene space dynamics during the evolution of aegilops tauschii, brachypodium distachyon, Oryza sativa, and sorghum bicolor genomes. Mol Biol Evol **28**: 2537–2547

Meinke D, Muralla R, Sweeney C, Dickerman A (2008) Identifying essential genes in *Arabidopsis thaliana*. Trends Plant Sci **13**: 483–91

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

Meyerowitz EM (1989) Arabidopsis, a useful weed. Cell **56**: 263–269

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

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S-H (2014) Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. Plant Cell **26**: 1925–1937

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**: 210–24

Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, et al (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Res **42**: D1193-9

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, et al (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol **13**: 418

Musso G, Tasan M, Mosimann C, Beaver JE, Plovie E, Carr L a, Chua HN, Dunham J, Zuberi K, Rodriguez H, et al (2014) Novel cardiovascular gene functions revealed via systematic phenotype prediction in zebrafish. Development **141**: 224–35

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol **152**: 29–43

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

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al (2015) Rfam 12.0: Updates to the RNA families database. Nucleic Acids Res **43**: D130–D137

Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29**: 2933–2935

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**: 1340–3

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

Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, Gkoutos G V, Harper L, He M, Hoehndorf R, et al (2015) An ontology approach to comparative phenomics in plants. Plant Methods **11**: 10

Ohno S, Wolf U, Atkin NB (1968) Evolution From Fish To Mammals By Gene Duplication. Hereditas **59**: 169–187

Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JES, Iversen ES, Hartemink AJ, Haase SB (2008) Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature **453**: 944–947

Panchy N, Lehti-Shiu MD, Shiu S-H (2016) Evolution of gene duplication in plants. Plant Physiol **171**: pp.00523.2016

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

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101**: 9903–8

Pearson H (2006) Genetics: what is a gene? Nature **441**: 398–401

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res **12**: 2825–2830

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al (2012) The GENCODE pseudogene resource. Genome Biol **13**: R51

Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. Nature **379**: 131–137

Pesole G (2008) What is a gene? An updated operational definition. Gene **417**: 1–4

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**: 1033–8

Ponting CP (2017) Biological function in the twilight zone of sequence conservation. BMC Biol **15**: 71

Ponting CP, Belgard TG (2010) Transcribed dark matter: meaning or myth? Hum Mol Genet **19**: R162-8

Portin P (2015) The Development of Genetics in the Light of Thomas Kuhn's Theory of Scientific Revolutions. Recent Adv DNA gene Seq **9**: 14–25

Portin P, Wilkins A (2017) The Evolving Definition of the Term "Gene." Genetics **205**: 1353–1364

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, et al (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol **19**: 45–50

Raff R (1996) The Shape of Life. University of Chicago Press, Chicago, IL

Rands CM, Meader S, Ponting CP, Lunter G (2014) 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLoS Genet **10**: e1004525

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol **2**: 989–1000

Roy S, Ernst J, Kharchenko P V, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow C a, Ma L, Lin MF, et al (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science **330**: 1787–97

Rutter MT, Cross K V, Van Woert P a (2012) Birth, death and subfunctionalization in the Arabidopsis genome. Trends Plant Sci **17**: 204–12

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

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet **37**: 501–6

Schreiber SL, Bernstein BE (2002) Signaling Network Model of Chromatin. Cell **111**: 771–778

Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M (2006) Predicting essential genes in fungal genomes. Genome Res **16**: 1126–35

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

Shipp M a, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, et al (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med **8**: 68–74

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15**: 1034–50

Silva JM, Marran K, Parker JS, Silva J, Golding M, Schlabach MR, Elledge SJ, Hannon GJ, Chang K (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. Science **319**: 617–20

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

Skendzic EM, Columbus JT, Cerros-Tlatilpa R (2007) Phylogenetics of Chloridoideae (Gramineae): A preliminary study based on nuclear ribosomal internal transcribed spacer and chloroplast trnL-F sequences. Aliso A J Syst Evol Bot **23**: 530–544

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA **110**: 2898–903

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**: 336–48

Spellman PT, Sherlock G, Zhang MQ, Vishwanath R, Anders K, Eisen MB, Brown PO, Futcher B (1998) Comprehensive Identification of Cell Cycle – regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Mol Biol Cell **9**: 3273–3297

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312–3

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

Stormo G, Schneider T (1982) Use of the "Perceptron" algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids … **10**: 2997–3012

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

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

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: Large-scale phytochemistry in the functional genomics era. Phytochemistry **62**: 817–836

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

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36**: 1009–1014

Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of sorghum and maize genome progenitors. Genome Res **14**: 1916–1923

Takuno S, Gaut BS (2012) Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. Mol Biol Evol **29**: 219–27

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

Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal

early polyploidy in the monocot lineage. Proc Natl Acad Sci USA **107**: 472–477

Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S (2007) Machine learning and its applications to biology. PLoS Comput Biol **3**: e116

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4**: 41

Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Genet **12**: 692–702

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Mar Albà M (2009) Origin of primate orphan genes: A comparative genomics approach. Mol Biol Evol **26**: 603–612

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, 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.

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

Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D, et al (2004) Identification of genes required for embryo development in *Arabidopsis*. Plant Physiol **135**: 1206–20

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

Upstill-Goddard R, Eccles D, Fliege J, Collins A (2013) Machine learning approaches for the discovery of gene-gene interactions in disease data. Brief Bioinform **14**: 251–60

Uygun S, Peng C, Lehti-Shiu MD, Last RL, Shiu SH (2016) Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. PLoS Comput Biol **12**: 1–27

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

Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I, Makałowski W,

Makałowska I (2004) Mammalian Overlapping Genes : The Comparative Perspective Mammalian Overlapping Genes : The Comparative Perspective Identification of Overlapping Genes. Genome Res 280–286

Wang X, Tang H, Paterson AH (2011) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. Plant Cell **23**: 27–37

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res **40**: 1–14

Wang Y, Tetko I V, Hall M a, Frank E, Facius A, Mayer KFX, Mewes HW (2005) Gene selection from microarray data for cancer classification--a machine learning approach. Comput Biol Chem **29**: 37–46

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

Winzeler E, Shoemaker D, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke J, Bussey H, et al (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science **285**: 901–6

Wu Y, Zhang W, Jiang J (2014) Genome-Wide Nucleosome Positioning Is Orchestrated by Genomic Regions Associated with DNase I Hypersensitivity in Rice. PLoS Genet. doi: 10.1371/journal.pgen.1004378

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**: 1193–203

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**: 1586–91

Ye Q-H, Qin L-X, Forgues M, He P, Kim JW, Peng AC, Simon R, Li Y, Robles AI, Chen Y, et al (2003) Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med **9**: 416–23

You Z-H, Yin Z, Han K, Huang D-S, Zhou X (2010) A semi-supervised learning approach to

predict synthetic genetic interactions by combining functional and topological properties of functional gene network. BMC Bioinformatics **11**: 343

Yuan Y, Xu Y, Xu J, Ball RL, Liang H (2012) Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. Bioinformatics **28**: 1246–52

Zhang J (2003) Evolution by gene duplication: An update. Trends Ecol Evol **18**: 292–298

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA **95**: 3708–3713

Zhang T, Jiang M, Chen L, Niu B, Cai Y (2013) Prediction of gene phenotypes based on GO and KEGG pathway enrichment scores. Biomed Res Int **2013**: 870795

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 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

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W (2008) On the origin of new genes in Drosophila. Genome Res **18**: 1446–1455

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

Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu S-H (2011) Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci USA **108**: 14992–7