Received: 18 March 2021 | Accepted: 16 June 2021

DOI: 10.1002/tpg2.20135

The Plant Genome

ORIGINAL RESEARCH

Machine learning approaches to identify core and dispensable genes in pangenomes

Alan E. Yocca[1,2] | **Patrick P. Edger**[2,3]

[1] Dep. of Plant Biology, Michigan State Univ., East Lansing, MI 48824, USA

[2] Dep. of Horticulture, Michigan State Univ., East Lansing, MI 48824, USA

[3] Genetics and Genome Sciences Program, Michigan State Univ., East Lansing, MI 48824, USA

Correspondence
Patrick P. Edger, Dep. of Horticulture, Michigan State Univ., East Lansing, MI, 48823, USA.
Email: edgerpat@msu.edu

Assigned to Associate Editor David Edwards.

Abstract

A gene in a given taxonomic group is either present in every individual (core) or absent in at least a single individual (dispensable). Previous pangenomic studies have identified certain functional differences between core and dispensable genes. However, identifying if a gene belongs to the core or dispensable portion of the genome requires the construction of a pangenome, which involves sequencing the genomes of many individuals. Here we aim to leverage the previously characterized core and dispensable gene content for two grass species [*Brachypodium distachyon* (L.) P. Beauv. and *Oryza sativa* L.] to construct a machine learning model capable of accurately classifying genes as core or dispensable using only a single annotated reference genome. Such a model may mitigate the need for pangenome construction, an expensive hurdle especially in orphan crops, which often lack the adequate genomic resources.

1 | INTRODUCTION

Reference genome assemblies contain information specific only to the individual of the species sequenced to create the assembly. They lack genomic regions present in other individuals of that species. Recently, the widespread adoption of pangenomics enabled characterization of the gene content diversity present in a species (Gao et al., 2019; Golicz et al., 2016; Gordon et al., 2017; Hübner et al., 2019; Hurgobin et al., 2018; Li et al., 2014; Lin et al., 2014; Montenegro et al., 2017; Ou et al., 2018; W. Wang et al., 2018; Yu et al., 2019; Zhou et al., 2017). The term *pangenome* was first coined in 2005, referring to collections of sequences across different strains of microorganisms (Tettelin et al., 2005). This early work built on the observation that genes often display presence–absence variation (PAV) across different strains. Genes present in every individual of a taxonomic group are called core genes, whereas genes absent in at least a single individual are called dispensable genes.

The generation of a pangenome allows us to determine if a particular gene in each reference assembly is either core or dispensable according to their presence or absence in individuals used to construct the pangenome. In addition, we know there are both qualitative and quantitative differences between core and dispensable genes. For example, in plants, core genes are often associated with essential metabolic processes, whereas dispensable genes are associated with adaptive functions (e.g., stress responses; Danilevicz et al., 2020). Previous work also demonstrated dispensable genes exhibit higher rates of polymorphism than core genes (Gordon et al., 2017; Hurgobin et al., 2018; Li et al., 2014; W. Wang et al., 2018). This framework is analogous to a binary classification problem, one potentially addressed by machine learning.

Abbreviations: AUC-ROC, area under the curve for the receiver operator curve; GC, guanine–cytosine base-pair; GNB, Gaussian naive Bayes; GO, Gene Ontology; Ka/Ks, ratio of nonsynonymous to synonymous substitutions; MCC, Matthews correlation coefficient; PAV, presence–absence variation; RF, random forest; SVC, support vector classifier.

We use the term *machine learning* to refer to the application of computer algorithms to classify observations based on previous information. Machine learning has increasingly more often been applied to genomics research (Golicz et al., 2020). For example, machine learning has been used to predict gene expression levels from genomic sequence data (Azodi et al., 2020). Machine learning has also been used in the biomedical field to diagnose disease (Kourou et al., 2015). A broad application of machine learning is executed in Deep Variant, a software tool that identifies variants based on short-read sequence alignments (Poplin et al., 2018).

Here we aim to apply machine learning algorithms to classify genes as core or dispensable in a new genome given nothing except a few simply determined characteristics of a gene. We first identify quantitative differences between core and dispensable genes in two different grass species, *Oryza sativa* L. and *Brachypodium distachyon* (L.) P. Beauv., for which high-quality pangenomes were developed (Gordon et al., 2017; W. Wang et al., 2018). Furthermore, a shared ancient polyploidization event and phylogenetic placement near many additional agronomically important species make these species befitting for our study. Then, we trained different machine learning models to differentiate between core and dispensable genes based on yet to be determined differences. Finally, we tested the feasibility of applying these trained models to species not used to train our models.

Core Ideas

- Previous pangenome studies identified differences between core and dispensable genes.
- We applied machine learning models to further differentiate between these gene classes.
- Machine learning models are capable at classifying genes as core or dispensable.

2 | METHODS

2.1 | Core and dispensable gene annotations

2.1.1 | *O. sativa*

We obtained a matrix of gene presence and absence from https://figshare.com/articles/dataset/Gene_presence_absence_variations_of_453_rice_accessions/5103769 taken from a recent publication (Wang et al., 2018). This study analyzed short read sequencing from <3,000 rice accessions, yet included gene PAV for 453 accessions (the authors selected these accessions for "sequencing depths >20× and mapping depths >15×").

The gene PAV matrix file is GenePAV.matrix.txt. It provides PAV information coded in binary (1 for presence, 0 for absence). Locus identifiers are provided according to the annotation downloaded below (e.g., Os01g0100100). Therefore, specific transcript information is unavailable. This potentially affects a few gene measures such as exon count and gene length. We take information for the longest listed transcript for each locus. We only consider loci with an available annotation in the IRGSP-1.0 rice annotation release. Therefore, we have PAV information for 35,633 genes with a locus identifier. Using this matrix, we defined core genes as those present in each of the 453 accessions in the matrix. Dispensable genes are those absent in at least a single accession.

2.1.2 | *B. distachyon*

Brachypodium distachyon core and dispensable gene information was downloaded from https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=BrachyPan (JGI login required). We found information from 54 brachypodium lines in accordance with Supplemental Table S1 from Gordon et al. (2017). We then created a PAV matrix for every locus in the reference genotype Bd21, again selecting the longest transcript. We define core genes as those present in all individuals and dispensable genes as those missing in at least a single accession. Importantly, given the information available, our core gene annotations may differ from those presented by Gordon et al. (2017). Their method focused on "Markov clustering in the GET_HOMOLOGUES-EST pipeline", while we simply infer ortholog presence from gene name maps provided between the reference genotype and each accession. The result is a shorter list of core genes in our study. Based on our results, we believe our list contains high confidence core gene annotations.

2.1.3 | Genome and annotation versions

We used the same genome and gene annotation versions as used in the pangenome studies from which we gathered core and dispensable annotations. For *O. sativa*, we collected the IGRSPv1.0 annotation (https://rapdb.dna.affrc.go.jp/download/irgsp1.html). For *B. distachyon*, we collected the Brachypodiumv2.1 annotation from JGI.

2.2 | Feature calculations

2.2.1 | Quantitative gene features

All gene features were gathered using the scripts 'annotate_core_genes_osat_nested.py' and 'annotate_core_genes_

bdis.py' for *O. sativa* and *B. distachyon*, respectively. We calculated the following features: gene length (transcription start site [TSS] to transcription end site [TES]), exon count, intron length, exon length, guanine-cytosine percentage, and the proportion of all possible dinucleotide pairs (AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, and CC).

2.2.2 | Duplication type

The duplication type for each gene was determined using the MCScanX duplicate_gene_classifier function. Genes were assigned to one of five classes: dispersed, proximal, singleton, tandem, or whole genome duplicate. Dispersed duplicates are those existing >20 genes apart from each other and not belonging to any other listed category. Proximal duplicates are paralogs located within 20 genes of each other. Singleton genes do not have a paralog. Tandem duplicates are labeled as paralogous pairs existing next to each other without any intervening genes. Whole genome duplicates are those labeled as anchor genes, in other words those which scaffold intragenomic collinear blocks (Wang et al., 2012). These anchor genes are hypothesized to have been duplicated by a polyploidization event. These five categories were one-hot encoded as separate features. They were listed as 1 if a gene was a member of the duplication class and 0 otherwise. We did not apply preprocessing to these columns.

2.2.3 | Substitution rate calculations

We calculated nonsynonymous and synonymous substitution rates in PAML using two different comparisons. First, we aligned orthologs between *B. distachyon* and *O. sativa*. Second, we aligned paralogs within each of the two species. PAML was run through a custom pipeline (https://github.com/Aeyocca/ka_ks_pipe). As not every gene model has a paralog and ortholog, there were missing values. We chose to code these missing values as an arbitrary number (5) the model could recognize as different from a potential true value.

2.3 | Model training and evaluation

Machine learning models were trained and evaluated using the scikit learn toolkit in the python programming language (Pedregosa et al., 2011). A few scripts were used to implement these functions:

Calculation of AUC-ROC for within and cross species predictions: 'osat_bdis_kfold_model_test_auc_21_02.py'

Creation of AUC-ROC curves for within-species 10-fold cross validation: 'osat_auc_roc_curves.py' and 'bdis_auc_roc_curves_array.py'

Calculation of accuracy for within and cross species predictions: 'osat_bdis_kfold_model_test_21_02.py'

Calculation of feature importance values for the Random Forest and Support Vector Machine models: 'osat_bdis_feat_imp.py'

3 | RESULTS

3.1 | Differences between core and dispensable genes

Previous pangenome studies have revealed that there are some functional differences between core and dispensable genes (Danilevicz et al., 2020). We investigated quantitative differences between gene models in reference genomes listed as core or dispensable according to two previous pangenomes. Wang et al. analyzed gene PAV across 453 *O. sativa* accessions (Wang et al., 2018). They found that roughly 58% of genes in the reference assembly were present in each of the 453 accessions. Gordon et al. (2017) generated full de novo assemblies for 54 *B. distachyon* accessions. They discovered that roughly 30% of genes in the reference genome were present in each accession. These two systems provide us with independent pangenome assemblies to train and test prediction models.

We observe quantitative differences for various features between the gene models of core and dispensable genes (Figure 1). Interestingly, we observe a bias for higher guanine–cytosine base-pair (GC) percentage in dispensable genes. As observed before, gene models in grass genomes have a bimodal GC percentage distribution (Clément et al., 2014; McKain et al., 2016). Other quantitative gene feature differences between core and dispensable gene models include gene length, exon count, intron count, exon length, and intron length (Figure 1; Supplemental Figure S1).

As dispensable genes are absent in at least one individual in a taxonomic group, we hypothesize core and dispensable genes may evolve differently. One signature of past selection is the ratio of nonsynonymous to synonymous substitutions (Ka/Ks ratio). Previous pangenome studies have reported a variety of approaches comparing nonsynonymous and synonymous substitution rates. Consistently, they find higher nonsynonymous substitution rates, as well as elevated Ka/Ks ratios for dispensable genes compared with core genes implying greater positive selection acts on dispensable genes

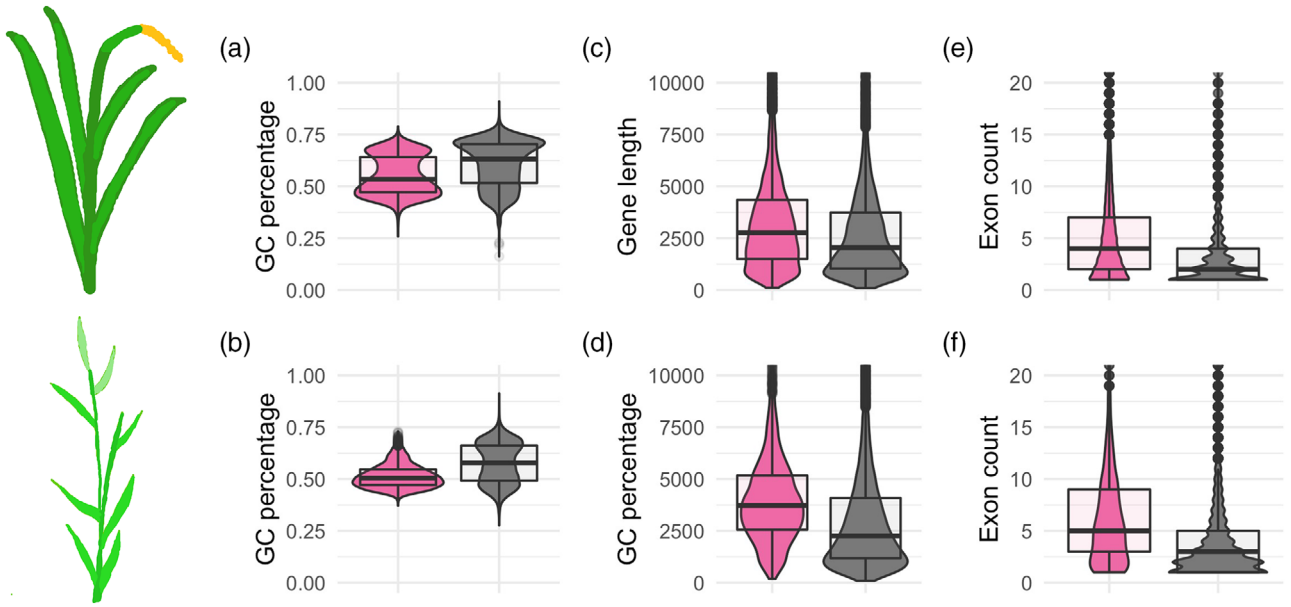

FIGURE 1 Quantitative differences in core (pink) and dispensable (gray) gene models for both *O. sativa* (a–c) and *B. distachyon* (d–f). Features are as follows: (a, d) guanine-cytosine percentage, (b, e) gene length measured from annotated transcription start site to transcription end site, and (c, f) number of exons

(Golicz et al., 2016; Gordon et al., 2017; Hurgobin et al., 2018; Li et al., 2014; Pinosio et al., 2016; W. Wang et al., 2018).

There are two different ways to calculate Ka/Ks values: (a) alignments between intragenomic paralogs or (b) alignments between orthologs across species. Both methods were applied here. We found 88.25% of *O. sativa* genes have a paralog, while 52.4% of genes have an ortholog with a *B. distachyon* gene. For *B. distachyon*, these values were 87.1 and 67.4% for paralogs and orthologs to a *O. sativa* gene, respectively. As reported in previous studies, Ka/Ks ratio distributions were higher for dispensable genes than for core genes (Supplemental Figure S1).

A consistent observation in pangenomic studies is that dispensable genes are enriched with functions associated with biotic and abiotic stress response. Therefore, we wanted to incorporate these underlying sequence differences in our models. We considered some sort of quantitative measure of Gene Ontology (GO) term similarity. However, given that the primary goal of our study is to develop a machine learning approach that may be suitable for orphan crops and lineages for which functional annotations are likely absent, we excluded GO terms for training our models. Please see other studies that have incorporated GO term differences into machine learning models (Cusack et al., 2020).

In an attempt to account for sequence differences between core and dispensable genes without the onus of missing data, we investigated the proportion of all possible dinucleotides as a feature. This measure has a value for each gene, is readily available, and may allow our machine learning models to learn differences in underlying sequence between core and dispensable gene functions. Indeed, we observe differences in dinucleotide proportions between core and dispensable genes, suggesting this information may contribute toward core and dispensable gene differentiation (Supplemental Figure S2).

3.2 | Are there differences between core and dispensable genes in relation to duplication type?

Gene duplications have played a major role in shaping the gene content in eukaryotic genomes and have contributed to the evolution of novel traits (Ohno, 1970). There are multiple mechanisms of gene duplication that may exhibit differences between core and dispensable genes as reported previously in sesame (Yu et al., 2019). We used MCScanX (Wang et al., 2012), a toolkit for evolutionary analyses, to classify each gene in both *O. sativa* and *B. distachyon* (Figure 2) into different gene duplication classes with the intention to test whether core or dispensable genes are enriched for tandem or whole genome duplicates. McScanX provides a function called duplicate_gene_classifier that assigns genes to one of five classes: dispersed, proximal, singleton, tandem, or whole genome duplicate. Dispersed duplicates are those existing >20 genes apart from each other and not belonging to any other listed category. Proximal duplicates are paralogs located within 20 genes of each other. Singleton genes do not have a paralog. Tandem duplicates are labeled as paralogous pairs existing next to each other without any intervening genes. Whole genome duplicates are those that were

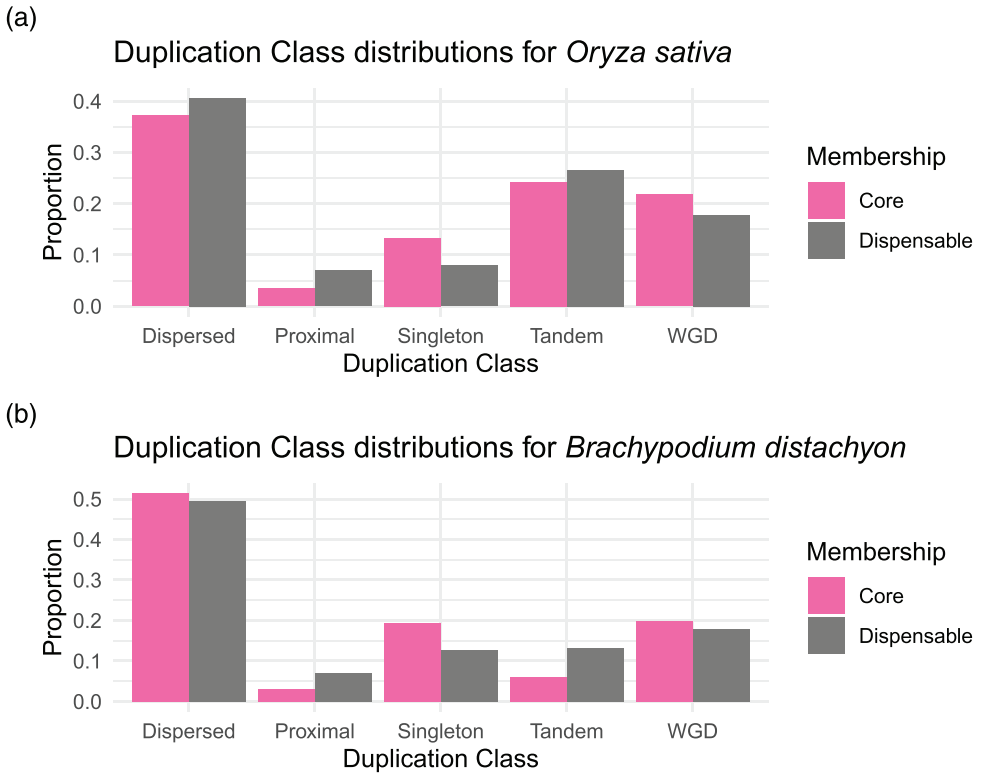

FIGURE 2 Proportion of retained duplicates by duplication class for core and dispensable genes. Panel a depicts differences for gene models in the *O. sativa* reference genome. Panel b depicts differences for gene models in the *B. distachyon* reference genome. WGD, whole genome duplication

derived from an ancient polyploid event (Wang et al., 2012). The genomes of *B. distachyon* and *O. sativa* share the remnants of the same three ancient polyploidization events: rho, sigma, and tau whole genome duplications (McKain et al., 2016).

We find nearly the same pattern for both *B. distachyon* and *O. sativa*. All differences between core and dispensable genes are significant (z-test p-value <.01). Dispensable genes contain a larger proportion of proximal and tandem duplicates, whereas core genes contain a larger proportion of whole-genome duplicates and single-copy genes. This pattern, combined with functional enrichment differences, is consistent with the gene balance hypothesis (M. Freeling, 2008). *O. sativa* and *B. distachyon* differ in which class contains a larger proportion of dispersed duplicates.

3.3 | Machine learning methods

We tested three separate machine learning methods: support-vector machine, Gaussian naive Bayes (GNB), and random forest (RF). These three approaches encompass different classification techniques. Use of all three techniques allow us to robustly investigate our potential to differentiate between core and dispensable genes.

Detailed descriptions of these methods can be found elsewhere (Vapnik, 1995; Hand & Yu, 2001; Breiman, 2001). The support-vector machine classifier shapes our data in multidimensional space. It then searches for a vector through that space which best separates our two classes, in our case core and dispensable genes. Therefore, when given a new gene to classify as core or dispensable, it plots that gene's values in the same multidimensional space and classifies the new gene according to the created vector of best separation. Gaussian naive Bayes takes each feature independently and assumes the feature values follow a Gaussian distribution whose midpoint represents the difference between the two classes. With this distribution for each feature, when given a new value, it can assign a probability of either class to that value. Summing these probabilities across all features, this classifier determines the new gene's class by the probability of belonging to either class given its feature values. The RF classifier creates a forest of decision trees. A decision tree is similar to a flowchart where different paths are taken at each node. Nodes in these trees represent values of a feature that will send a new gene along different classification paths depending on their values. The RF is a common method used in classification and is capable of learning high-order and nonlinear associations in the classification data.

TABLE 1 Preliminary model assessment

Species	Model	Accuracy	AUC-ROC	MCC
Oryza sativa	SVM	0.695 ± 0.018	0.750 ± 0.024	0.359 ± 0.041
Oryza sativa	GNB	0.647 ± 0.020	0.713 ± 0.026	0.279 ± 0.044
Oryza sativa	RF	0.710 ± 0.017	0.774 ± 0.019	0.392 ± 0.035
Brachypodium distachyon	SVM	0.766 ± 0.010	0.792 ± 0.013	0.404 ± 0.025
Brachypodium distachyon	GNB	0.689 ± 0.011	0.785 ± 0.012	0.393 ± 0.017
Brachypodium distachyon	RF	0.799 ± 0.008	0.856 ± 0.009	0.496 ± 0.019

Note. Values are means and ranges are the standard deviation from 10 cross-fold validations. AUC-ROC, area under the curve for the receiver operator curve; GNB, Gaussian naive Bayes; MCC, Matthews correlation coefficient; RF, random forest; SVM, support vector machine.

3.4 | Model training and assessment

We could train our models using all of our data; however, testing on that same data may allow the models to effectively memorize certain genes feature values, a phenomenon called overtraining. These models will perform poorly on unseen data. To prevent overtraining, we only want to train our models on a subset of our data. That subset may not reflect all the patterns in our data. To ensure our model performance is not affected by a nonrepresentative subset of our data, we apply a *k*-fold cross validation approach. We split our data into *k* equally sized subsets ($k = 10$ in our case). For each subset, we train our models on all other subsets and test it on the remaining subset. Training and testing separately on all 10 subsets allows for a robust assessment of the model's performance.

There are several different metrics to test model performance. We focus on accuracy, area under the curve for the receiver operator curve (AUC-ROC), and Matthews correlation coefficient (MCC). Accuracy is simply the proportion of correctly classified genes in the testing subset. The AUC-ROC incorporates true positive classification rate and false positive classification rate. If the model is a random guesser, it will achieve an AUC-ROC score of 0.5 for binary classification problems. A perfect AUC-ROC score is 1 where genes are always correctly classified. This value allows us to better assess our true and false positive rates. MCC essentially measures the correlation between observed and predicted classes. Its value ranges from −1 to +1, with positive values indicating agreement between observation and prediction.

3.5 | La reveal magnifico

We trained and tested all three models for both species. The results are presented in Table 1 (Supplemental Table S1, Supplemental Figures S3–S6). Values correspond to averages across all *k*-folds. Overall, all models performed better than random expectations. This indicates we are able to learn differences between core and dispensable genes in different species and classify unseen genes as core or dispensable. Model performance is overall better in *B. distachyon* than in *O. sativa*. The RF model outperformed other models in terms of both accuracy and AUC-ROC.

3.6 | What features are most important for classifying core and dispensable genes?

Not every feature is likely to contribute equally to model performance. To determine which features are most important for differentiating between core and dispensable genes, we performed recursive feature elimination, one of several strategies of feature selection. In recursive feature elimination, we train our model using all features, and the least important feature is eliminated. After elimination, we retrain our model and perform the same operation. We measure model accuracy at each step. By viewing the accuracy as features are eliminated, we can select the combination of features that provides us with the greatest accuracy. These curves for both the support vector classifier (SVC) and RF models are shown in Supplemental Figure S4. We find using all features results in the highest model performance compared with excluding low performing features.

The RF and SVC models explicitly provide relative feature importance scores. These scores reflect how much relative weight each feature contributes to the final prediction. Relative feature importance scores are shown in Supplemental Figure S5. Overall, several measures related to GC percentage stand out as large contributors to final predictions. Comparing the GC percentage between core and dispensable genes in both *B. distachyon* and *O. sativa* reveal striking differences in this measure (Figure 1a, 1d). Therefore, we believe GC percentage is an important distinguishing character between core and dispensable genes in these grass species. However, we cannot determine the extent of causality in this relationship. We developed these models in two species within the same family, Poaceae. Perhaps the importance of GC percentage is tied to some other feature we do not directly measure. As pangenomic resources become available in a wider breadth of species, we will learn whether GC percentage is a

TABLE 2 Cross-species model performance (AUC-ROC and MCC) compared with intraspecific model performance

Training data	Testing data	Model	AUC-ROC	AUC-ROC self	MCC	MCC-self
Oryza sativa	*Brachypodium distachyon*	SVC	0.767	0.750 ± 0.019	0.402	0.360 ± 0.030
Oryza sativa	*Brachypodium distachyon*	GNB	0.758	0.712 ± 0.021	0.395	0.279 ± 0.038
Oryza sativa	*Brachypodium distachyon*	RF	0.750	0.775 ± 0.010	0.336	0.396 ± 0.014
Brachypodium distachyon	*Oryza sativa*	SVC	0.719	0.794 ± 0.012	0.317	0.459 ± 0.026
Brachypodium distachyon	*Oryza sativa*	GNB	0.701	0.785 ± 0.013	0.269	0.432 ± 0.024
Brachypodium distachyon	*Oryza sativa*	RF	0.687	0.857 ± 0.011	0.292	0.571 ± 0.023

Note. AUC-ROC, area under the curve for the receiver operator curve; GNB, Gaussian naive Bayes; MCC, Matthews correlation coefficient; RF, random forest; SVC, support vector classifier.

Poaceae-specific or a broader distinguishing feature between core and dispensable genes.

3.7 | Does the choice of reference genotype affect prediction quality?

We were interested if the choice of reference genotype affected prediction quality. Thankfully, annotations were available for multiple *B. distachyon* genotypes. To test if the choice of reference genotype affects our prediction models, we repeated our analyses using two additional *B. distachyon* genotypes: ABR2 and Tek-2. These accessions were randomly selected to represent all three structure groups reported by Gordon et al. (2017).

Results are shown in Supplemental Tables S2 and S3. Overall, models trained on the reference genotype performed best across all metrics in cross-species comparisons. However, models trained on the two nonreference genotypes produced excellent metrics, often only a few percentage points lower than the reference-trained models. Therefore, we conclude the choice of reference genotype will not significantly hinder core and dispensable gene predictions. There may exist genotypes that better reflect broad core and dispensable gene patterns, but our analysis suggests these patterns hold across genotypes within a single species.

3.8 | How does a model trained on one species perform on the other?

We trained our models on one species and tested it on the other species. In this instance, it is important to consider the proportion of core and dispensable genes in each reference genome. A model trained on a species with 70% core genes will anticipate 70% of the testing data to be core as well. In addition, if 70% of genes in a reference genome are core, a model can obtain an accuracy of 70% simply by predicting every case to be core. This emphasizes the importance of measures other than accuracy alone to evaluate models such as MCC and AUC-ROC.

The proportion of core genes in a reference genome is variable across lineages (Golicz et al., 2020). To account for these differences, we (a) test balanced training and testing data, as well as (b) measure AUC-ROC and MCC rather than accuracy. To balance our datasets, we subset the majority class to the size of the minority class. For example, in *B. distachyon*, we find ~30% ($n = 9{,}308$) of genes are core out of all used to train our models ($n = 31{,}679$). We balance this data by subsetting 9,308 core genes and train our models on only 18,616 genes rather than the 31,679 total gene models.

As expected, model performance is worse cross-species than within species for *B. distachyon* (Table 2; Supplemental Table S1). However, SVC and GNB models trained on *O. sativa* and tested on *B. distachyon* on average performed better than the same model tested in *O. sativa*. The trend held for most comparisons using data from different *B. distachyon* genotypes as well. This suggests the quantitative differences between core and dispensable genes in *B. distachyon* are greater than those in *O. sativa*. The differences are at least distinct enough for these models to leverage when differentiating between core and dispensable genes.

There are two possible reasons for decreased model performance cross species. First, there may be lineage-specific differences between core and dispensable genes. Indeed, we visualize these differences by plotting gene model distributions on the same axes comparing *O. sativa* and *B. distachyon* (Supplemental Figure S1). Second, though we attempted to correct for it, our models could be overtrained on our data.

4 | DISCUSSION

In this study, we tested the efficacy of training machine learning models to differentiate between core and dispensable genes in a single genome based on various gene features. We first sought to characterize differences between core and dispensable genes in two species (*B. distachyon* and *O. sativa*) with available pangenomes. Determining the origin of dispensable genes is beyond the scope of this study. However, our observations of shorter dispensable genes with fewer exons suggest a fraction of them may have arisen de novo. This observation is consistent with the hypothesis that short sequences are more likely to gain genic functions from previously noncoding DNA than longer sequences, and reports of de novo gene origin in yeast and *Drosophila* (Carvunis et al., 2012; Siepel, 2009) (Figure 1). However, gene duplications are likely still the predominant source of new genes.

We observed differences in the evolutionary origin and evolutionary signatures between core and dispensable genes. In addition to differences in length and exon count, we investigated Ka/Ks, as well as duplication type differences between core and dispensable genes. We calculated nonsynonymous and synonymous substitution rates in PAML (Yang, 2007) using two different approaches. First, we aligned orthologs between *B. distachyon* and *O. sativa*. Second, we aligned paralogs within each of the two species. The differences in distributions between core and dispensable genes for paralogs and orthologs yielded similar results for both species. As shown previously, elevated rates of Ka/Ks observed in dispensable genes relative to core genes imply higher rates of positive selection on these genes and possibly the evolution of novel gene functions (matching the hypothesis outlined by Susumu Ohno; (Ohno, 1970).

The duplication history of each gene in both genomes was examined. Previous studies suggested that core and dispensable genes are enriched with different classes of gene duplicates (Yu et al., 2019). The observed gene duplication differences are consistent with hypotheses of dosage sensitive genes, as outlined by the gene balance hypothesis (Birchler & Veitia, 2007, 2012). If core genes encode for more essential cellular functions, which are known to be enriched with highly dosage-sensitive genes (Freeling, 2009), they would contain a higher proportion of retained duplicates from ancient whole genome duplications. Dosage sensitive genes must retain duplicates from ancient polyploid events to maintain proper stoichiometry in macromolecular complexes and gene networks (Birchler et al., 2001). This skewed pattern for retained whole genome duplicates was observed for core genes in both the rice and *Brachypodium* genomes. Similarly, previous studies have suggested that certain single-copy genes encode for essential functions, including organellar-nuclear interactions (Edger & Pires, 2009), and must remain in single copy due to gene dosage constraints (De Smet et al., 2013; Tasdighian et al., 2017). Our analyses also show that core genes are enriched with a greater number of single copy genes.

Dispensable genes, on the other hand, are enriched with more adaptive functions. Adaptive genes tend to be more poorly connected and, thus, are heavily skewed toward being more dosage insensitive (Rizzon et al., 2006). Dosage insensitive genes are known to be enriched with tandem duplicated genes (M. Freeling, 2008; Birchler & Veitia, 2007). Similarly, the dispensable gene content of the pangenome, as shown in this study, is enriched with tandem duplicates. Thus, tandem duplication appears to be the prominent mechanism giving rise to new dispensable genes. In summary, core genes contain a higher proportion of retained duplicates from whole genome duplications and single copy genes, while dispensable genes contain a higher proportion of retained tandem duplicates.

Applying three separate machine learning models revealed similar results. We are able to differentiate between core and dispensable genes better than random, yet with imperfect accuracy. Our three models displayed different performances likely due to differences in the distributions of the data. For example, we believe the GNB model performed the worst because it fits each feature to a normal distribution when the distribution of each feature is not normally distributed. In addition, it gives equal weight to each feature in the decision-making process. Our other two models demonstrate each feature contributes a unique amount to the final prediction resulting in worse performance for the GNB model. The RF model outperformed the support vector machine model. This observation is consistent with other applications of machine learning that demonstrate RF often outperforms other models (except Cusack et al., 2020). We recommend using multiple models on applications of classifying genes as core or dispensable in the future.

Model performance performed better within than across species (with a few exceptions noted above). Previous applications of machine learning across species yielded similar results (Lee et al., 2011; Chen et al., 2018; Kelley, 2020; Mejía-Guerra & Buckler, 2019). For example, Meng et al. (2021) trained machine learning models to identify cold-responsive genes across a few different grass species. Consistent with our results, models performed best when trained and tested in the same species. Notably, they mentioned shared phenotypes are better indicators of cross-species model performance than ancestry. Therefore, perhaps our cross species model performances would be improved by testing in not only phylogenetically closer taxa, but also those exhibiting similar phenotypic and perhaps pangenome characteristics.

Annotation quality likely also factors into cross-species model performance. For example, poor quality annotations may inaccurately list transcription start and end sites as well as exon boundaries. In addition to annotation quality, inconsistent annotation methods between *B. distachyon* and *O. sativa* may also affect cross-species model performances due

to potential systematic biases in different annotation methods. This would reduce the efficacy of our models, as they are trained to recognize patterns not consistently represented between annotations. Though this was not explicitly tested here, we advocate high annotation quality as a key component to developing and testing these models to new lineages.

A potential application of these models is to classify genes as core or dispensable in a new species without the costly construction of a pangenome. Although our models perform better than random guessing, our accuracy rates are insufficient to substitute for pangenome construction for many downstream applications. However, if ~70% accuracy is all that is required, perhaps in the case of developing a genotyping array that consists of largely core genes for guiding breeding efforts, this strategy may likely suffice. We recommend training a model on a species as closely related to the study species as possible. Therefore, we advocate for a community-wide effort for pangenome construction of strategically phylogenetically placed taxa. Broad pangenome development will further increase our understanding of not only what combination of features differentiate core and dispensable genes, but also on various topics ranging from better understanding the evolutionary dynamics of gene families to genotype-phenotype associations.

ACKNOWLEDGMENTS

This work was supported by Michigan State University AgBioResearch, National Science Foundation (DEB #1737898), and USDA (AFRI #2019-51181-30015). We thank members of the Edger Lab and two meticulous reviewers for helpful feedback. Scripts used in this study are publicly available on github https://github.com/Aeyocca/Core_Prediction.

AUTHOR CONTRIBUTIONS

Alan E. Yocca: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Visualization; Writing-original draft; Writing-review & editing. Patrick P. Edger: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Funding acquisition; Visualization; Writing-original draft; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Alan E. Yocca https://orcid.org/0000-0002-0974-364X

REFERENCES

Azodi, C. B., Lloyd, J. P., & S-H, S. (2020). The cis-regulatory codes of response to combined heat and drought stress in *Arabidopsis thaliana*. *NAR Genomics and Bioinfomatics*, *2*(3), lqaa049. https://doi.org/10.1101/2020.02.28.969261

Birchler, J. A., Bhadra, U., Bhadra, M. P., & Auger, D. L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: Implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Developmental Biology*, *234*(2), 275–288. https://doi.org/10.1006/dbio.2001.0262

Birchler, J. A., & Veitia, R. A. (2007). The gene balance hypothesis: From classical genetics to modern genomics. *The Plant Cell*, *19*(2), 395–402. https://doi.org/10.1105/tpc.106.049338

Birchler, J. A., & Veitia, R. A. (2012). Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), 14746–14753.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–374. https://doi.org/10.1038/nature11184

Chen, L., Fish, A. E., & Capra, J. A. (2018). Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLOS Computational Biology*, *14*(10), e1006484. https://doi.org/10.1371/journal.pcbi.1006484

Clément, Y., Fustier, M.-A., Nabholz, B., & Glémin, S. (2014). The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biology and Evolution*, *7*(1), 336–348. https://doi.org/10.1093/gbe/evu278

Cusack, S. A., Wang, P., Moore, B. M., Meng, F., Conner, J. K., Krysan, P. J., Lehti-Shiu, M. D., & Shiu, S.-H. (2020). Genome-wide predictions of genetic redundancy in Arabidopsis thaliana. *bioRxiv*. https://doi.org/10.1101/2020.08.13.250225.

Danilevicz, M. F., Tay Fernandez, C. G., Marsh, J. I., Bayer, P. E., & Edwards, D. (2020). Plant pangenomics: Approaches, applications and advancements. *Current Opinion in Plant Biology*, *54*, 18–25. https://doi.org/10.1016/j.pbi.2019.12.005

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., & Van De Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2898–2903. https://doi.org/10.1073/pnas.1300127110

Edger, P. P., & Pires, J. C. (2009). Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, *17*(5), 699–717. https://doi.org/10.1007/s10577-009-9055-9

Freeling, M. (2008). The evolutionary position of subfunctionalization, downgraded. *Genome Dynamics*, *4*, 25–40. https://doi.org/10.1159/000126004

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology*, *60*, 433–453. https://doi.org/10.1146/annurev.arplant.043008.092122

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., Van Der Knaap, E., Huang, S., Klee, H. J., … Fei, Z. (2019).

The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*, 1044–1051. https://doi.org/10.1038/s41588-019-0410-2

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., Mccombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*, 13390. https://doi.org/10.1038/ncomms13390

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics comes of age: From bacteria to plant and animal applications. *Trends in Genetics*, *36*(2). https://doi.org/10.1016/j.tig.2019.11.006

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., Stritt, C., Roulin, A. C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T. E., Amasino, R., … Vogel, J. P. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, *8*(1), 132–145. https://doi.org/10.1038/s41467-017-02292-8

Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not so stupid after all? *International Statistical Review*, *69*(3), 385–398. https://doi.org/10.1111/j.1751-5823.2001.tb00465.x

Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., Lee, J. S., Baute, G. J., Owens, G. L., Grassa, C. J., Ebert, D. P., Ostevik, K. L., Moyers, B. T., Yakimowski, S., Masalia, R. R., Gao, L., Ćalić, I., Bowers, J. E., Kane, N. C., … Rieseberg, L. H. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants*, *5*(1), 54–62. https://doi.org/10.1038/s41477-018-0329-0

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-.K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., Parkin, I. A. P., Pires, J. C., Chalhoub, B., King, G. J., Snowdon, R., Batley, J., & Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the *Amphidiploid brassica napus*. *Plant Biotechnology Journal*, *16*(7), 1265–1274. https://doi.org/10.1111/pbi.12867

Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. *PLOS Computation Biology*. Advance online publication. https://doi.org/10.1101/660563

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Lee, D., Karchin, R., & Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research*, *21*(12), 2167–2180. https://doi.org/10.1101/gr.121905.111

Lin, K.e, Zhang, N., Severing, E. I., Nijveen, H., Cheng, F., Visser, R. G.f, Wang, X., De Ridder, D., & Bonnema, G. (2014). Beyond genomic variation: Comparison and functional annotation of three *Brassica rapa* genomes: A turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics*, *15*, 250. https://doi.org/10.1186/1471-2164-15-250

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y. I., Zheng, L., Zhang, S.-S., Zuo, Q., Shi, X.-H., Li, Y.-F., Zhang, W.-K. E., Hu, Y., Kong, G., Hong, H.-L., Tan, B., … Qiu, L.i.-J. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*(10), 1045–1052. https://doi.org/10.1038/nbt.2979

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., dePamphilis, C. W., Givnish, T. J., Pires, J. C., Stevenson, D. W. M., & Leebens-Mack, J. H. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution*, *8*(4), 1150–1164.

Mejía-Guerra, M. K., & Buckler, E. S. (2019). A k-mer grammar analysis to uncover maize regulatory architecture. *BMC Plant Biology*, *19*(1), 103. https://doi.org/10.1186/s12870-019-1693-2

Meng, X., Liang, Z., Dai, X., Zhang, Y., Mahboub, S., Ngu, D. W., Roston, R. L., & Schnable, J. C. (2021). Predicting transcriptional responses to cold stress across plant species. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(10), e2026330118. https://doi.org/10.1073/pnas.2026330118

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-.K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*(5), 1007–1013. https://doi.org/10.1111/tpj.13515

Ohno, S. (1970). *Evolution by gene duplication*. Springer Nature. https://doi.org/10.1007/978-3-642-86659-3

Ou, L., Li, D., Lv, J., Chen, W., Zhang, Z., Li, X., Yang, B., Zhou, S., Yang, S., Li, W., Gao, H., Zeng, Q., Yu, H., Ouyang, B.o, Li, F., Liu, F., Zheng, J., Liu, Y., Wang, J., … Zou, X. (2018). Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *The New Phytologist*, *220*(2), 360–363. https://doi.org/10.1111/nph.15413

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., & Morgante, M. (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution*, *33*(10), 2706–2719. https://doi.org/10.1093/molbev/msw161

Poplin, R., Chang, P. I.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., Mclean, C. Y., & Depristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Rizzon, C., Ponger, L., & Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLOS Computational Biology*, *2*(9), e115. https://doi.org/10.1371/journal.pcbi.0020115

Siepel, A. (2009). Darwinian alchemy: Human genes from noncoding DNA. *Genome Research*, *19*, 1693–1695. https://doi.org/10.1101/gr.098376.109

Tasdighian, S., Van Bel, M., Li, Z., Van De Peer, Y., Carretero-Paulet, L., & Maere, S. (2017). Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *The Plant Cell*, *29*(11), 2766–2785. https://doi.org/10.1105/tpc.17.00313

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit Y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome.' *Proceedings of the National Academy of Sciences of*

the United States of America, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer. https://doi.org/10.1007/978-1-4757-2440-0

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., … Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, *557*(7703), 43–49. https://doi.org/10.1038/s41586-018-0063-9

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7), e49. https://doi.org/10.1093/nar/gkr1293

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D., & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal*, *17*(5), 881–892. https://doi.org/10.1111/pbi.13022

Zhou, P., Silverstein, K. A. T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A. D., Steele, K. P., Stupar, R. M., Miller, J. R., Tiffin, P., Mudge, J., & Young, N. D. (2017). Exploring structural variation and gene family architecture with *de novo* assemblies of 15 *Medicago* genomes. *BMC Genomics*, *18*(1), 261. https://doi.org/10.1186/s12864-017-3654-1

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Yocca AE, Edger PP. Machine learning approaches to identify core and dispensable genes in pangenomes. *Plant Genome*, 2022;*15:*e20135. https://doi.org/10.1002/tpg2.20135