Title: Predicting fitness related traits using gene expression and machine learning

Authors: Georgia A. Henry[1]* & John R. Stinchcombe[1,2,3]

[1] Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

[2] Koffler Scientific Reserve at Joker's Hill, University of Toronto, King, Ontario, Canada

[3] Swedish Collegium for Advanced Study, Uppsala, Sweden

Corresponding author: georgia.henry@mail.utoronto.ca

Author contributions: G.A.H. and J.R.S contributed to the design, analysis and writing of the manuscript. G.A.H. performed the data collection and implemented analyses.

Conflict of Interest Statement: The authors declare no conflict of interest.

Data Availability: The data underlying this article will be made available through the NCBI Sequence Read Archive upon publication.

Abstract

Evolution by natural selection occurs at its most basic through the change in frequencies of alleles; connecting those genomic targets to phenotypic selection is an important goal for evolutionary biology in the genomics era. The relative abundance of gene products expressed in a tissue can be considered a phenotype intermediate to the genes and genomic regulatory elements themselves, and more traditionally measured macroscopic phenotypic traits such as flowering time, size, or growth. The high-dimensionality, low sample size nature of transcriptomic sequence data is a double-edged sword, however, as it provides abundant information but makes traditional statistics difficult. Machine learning has many features which handle high-dimensional data well and is thus useful in genetic sequence applications. Here we examined the association of fitness-components with gene expression data in *Ipomoea hederacea* (Ivyleaf Morning Glory) grown under field conditions. We combine the results of three different machine learning approaches and find evidence that expression of photosynthesis-related genes is likely under selection. We also find that genes related to stress and light response were overall important in predicting fitness. With this study we demonstrate the utility of machine learning models for smaller samples, and their potential application for understanding natural selection.

Introduction

Natural selection is a ubiquitous evolutionary force which acts on the phenotypes of individuals in a population, resulting in genetic changes if heritable variation exists for the traits under selection. Understanding the shape and strength of natural selection typically requires observers to choose specific traits to measure which may or may not be the subject of selection

(Lande and Arnold 1983). The type of traits studied has been found to influence the strength and temporal constancy of selection, with fecundity traits experiencing more consistent and strong selection than survivorship (Hoekstra et al. 2001; Kingsolver et al. 2001). Trait correlations and pleiotropy can additionally make detecting selection statistically difficult due to their potential to reduce power (Hersch and Phillips 2004). With high-throughput phenotyping, such as metabolomics or gene expression profiling, investigators often obtain a vast number of highly correlated traits, making traditional selection analysis difficult or impossible. Machine learning modelling and regularization techniques, which can alleviate issues of multicollinearity and overfitting due to an overabundance of explanatory variables, now make it possible for investigators to avoid having to decide which traits may be important to selection, and instead observe them from the data. Here, we use machine learning modelling to estimate natural selection on gene expression, using estimates of fitness and gene expression obtained from a field experiment with the Ivyleaf morning glory (*Ipomoea hederacea*).

Gene expression can be viewed as an intermediate between the genome and traditionally observable traits, and reflects both the underlying genetic makeup and the environment (Liao and Weng 2015; Josephs 2021). Quantifying gene expression through RNA sequencing of tissue serves as an unbiased sampling of phenotypic traits, unhindered by the observer's opinions on what might be an important trait (Rockman and Kruglyak 2006; Josephs 2021). RNA-sequence data is also highly multivariate, in that it captures a sample of all the genes being expressed in a tissue at the moment of sampling, which means that in addition to being unbiased it is also a broad representation of phenotypes. Although gene expression is expected to largely be under stabilizing selection (Rifkin et al. 2005; Gilad et al. 2006), there may additionally be directional selection on genes, especially when individuals are farther from the

fitness optima. Groen et al. (2020) measured selection on gene expression under drought and standard field conditions in domesticated rice (*Oryza sativa*). Using univariate approaches, they were able to estimate selection differentials for individual genes, finding generally weak selection. Selection was stronger under drought conditions, with earlier flowering time, modulated apparently by a single gene, allowing plants to "escape" the effects of drought (Groen et al. 2020). They were additionally able to identify significant selection on genes related to photosynthesis and growth, using a multivariate approach where they regressed fitness on principal component (PC) scores that summarized expression of many genes at once, although they were not able to distinguish direct and total selection on an individual gene's expression (Groen et al. 2020). Thus, measuring selection on gene expression may facilitate identifying genes or pathways which are important for fitness, with less anthropomorphic bias about what phenotypes of an organism are targeted by natural selection.

Genomic and transcriptomic sequencing data is, however, typically plagued by the "small *n*, large *p*" problem, where the number of "features" (genes, sequences, etc.) far exceeds the sample size. Machine learning tools are often well-suited to handling the "curse of dimensionality" due to intrinsic and extrinsic regularization, and are thus useful in the analysis of sequence data (Schrider and Kern 2018; James et al. 2021). Machine learning typically includes some type of loss function which serves to balance under- and over-fitting of the model, and as such observations with extreme values are not overly influential. The data is split into a training set (which is used to fit the model) and a testing set of data, which can then be used to evaluate the model performance, which can buffer against some overfitting by reducing the amount of statistical noise being fit to the model. Additionally, feature selection either through manual evaluation and pruning (as in Principal Component Regression), or through intrinsic model

behaviour (as in LASSO regression or Decision Trees), can alleviate problems due to too many explanatory variables. Regularized estimates are biased (i.e., no longer unbiased, least squares estimates) but often lead to more accurate overall prediction (see Morrissey 2014).

Population genomics has incorporated various deep learning methods for improved understanding of introgression and hybridization (Schrider et al. 2018; Small et al. 2020), selective sweeps (Xue et al. 2021), recombination (Adrion et al. 2020), and dispersal (Battey et al. 2020; Smith et al. 2022). Genome-wide association studies have also benefited from the introduction of various machine learning algorithms (Sun et al. 2021), using random forest (Lunetta et al. 2004), support vector machines (Mieth et al. 2016; Maciukiewicz et al. 2018), and unsupervised methods (Chang and Keinan 2014; Lu et al. 2015, 2016). Various statistical learning techniques have been applied broadly in biomedicine to associate gene expression with some outcome or disease risk (Kourou et al. 2015; O'Connell et al. 2016, 2017; Huang et al. 2018). Machine learning approaches using RNA-seq data have also been used for successfully predicting phenotypic diagnostic measures in biomedical studies such as interstitial pneumonia (Choi et al. 2018) and a suite of diagnostic and survival prediction tasks including diabetes and cancer (Smith et al. 2020). Generally, machine learning tools focus on optimizing prediction rather than estimation of parameter values within a model (Schrider and Kern 2018; James et al. 2021; Greener et al. 2022) and thus are valuable for understanding complex nonlinear relationships between variables. As such, machine learning serves as a good toolkit for understanding selection on gene expression in a field environment.

We used gene expression data from 96 *Ipomoea hederacea* individuals grown under natural field conditions to detect important fitness-related genes using various machine learning methods. We compare the results from a mix of unsupervised and supervised approaches. Unsupervised machine learning models are those in which the observations are not “labeled” with a response variable, and supervised models in contrast use observations which are “labeled” with a response variable. We first use principal component analysis on gene expression as an unsupervised dimensionality reduction method followed by linear regression of relative fitness on principal component scores. We then trained two different supervised classification models, a deep learning model and an ensemble method, to predict whether an individual set seed based on our gene expression data. In comparing disparate methods, we sought to find consistently important candidate genes or biological processes which are likely contributing to differences in fitness among individuals in the field. We investigated the predictive ability of various machine learning approaches and extracted the genes whose expression contributed most to the model fit (i.e. those associated most strongly with fitness differences). We then used Gene Ontology (GO) analyses to deepen our understanding of the types of genes which were significantly enriched in each model and the biological processes commonly found to be important among the models.

Methods

Study species and Natural History

We collected tissue from *Ipomoea hederacea* (Convolvulaceae), an annual vine commonly distributed throughout the eastern USA in agricultural fields, roadsides and other disturbed habitats. It is found as far north as New Jersey and southern Pennsylvania and ranges into

Mexico. It exhibits quantitative and Mendelian genetic clines (Bright and Rausher 2008; Simonsen and Stinchcombe 2010; Campitelli and Stinchcombe 2013a; Stock et al. 2014) and neutral genetic patterns consistent with metapopulation dynamics (Campitelli and Stinchcombe 2014). Germination typically occurs in early summer, and once plants begin to flower, they continue to do so until frost.

Field Experimental Design and Sampling

For this study we subsampled from a larger field experiment (Henry and Stinchcombe 2023) where seeds were generated by self-fertilization in a common greenhouse environment. We sampled 100 individuals from 56 populations (1-3 individuals per population, mean = 1.80), gathered from 10 states in *I. hederacea's* eastern North American distribution, spanning ~7° of latitude (33.017681° to 40.340767° N). On 20 July 2021, we germinated the scarified seeds into Pro-Mix BX mycorrhizae soil in 4" peat pots in a glasshouse at ambient temperature and light. Three days later, we transplanted pots into a recently ploughed old at the Koffler Scientific Reserve (www.ksr.utoronto.ca), with pots placed into the ground flush with field soil; we soaked the individuals with water, and provided no further or additional supplementation. The plants were left to grow naturally among emergent weeds and/or die naturally. A killing frost ended the experiment on 27 October, 2021. For further details on the field experimental design and environment, see Henry and Stinchcombe (2023).

Based on preliminary results (*see below*), we decided to sample soils at our field site for a panel of heavy metals. As a consequence of the time required for initial data analysis, soils were collected 1 year after the field experiment concluded, from the same field, which had been plowed on the same schedule as the year of our experiment. We collected soil from across five

transects, homogenizing the soil sampled within each transect, and sent the samples to the Agricultural and Food Laboratories at the University of Guelph for evaluation. Levels of arsenic, cobalt, chromium, cadmium, copper, mercury, nickel, lead, and zinc were quantified.

Tissue Collection and Extraction

We collected leaf tissue on 29 September 2021, 71 days after sowing. We sampled whole leaves approximately 2.5 cm in diameter and placed them into RNAse-free 2 mL microtubes with a 6mm sterile glass bead. We immediately submerged the samples in liquid nitrogen and transferred them into a cooler containing dry ice and excess liquid nitrogen until tissue collection was complete. We stored samples in a -80 °C freezer until we performed extractions in November 2021.We extracted mRNA using Qiagen Plant RNEasy Extraction kits and followed the standard protocol, using a TissueLyser to prepare samples. Genome Quebec provided sequencing using paired-end Illumina mRNA sequencing (NovaSeq6000 S4 PE100 Sequencing). Of the 100 samples collected, 97 were successfully sequenced, representing 55 populations. Of those 97 individuals, 58 had flowered by leaf tissue collection, 86 had flowered by the end of the experiment and 26 had set seed by the end of the experiment, approximately 1 month later (27 October 2021). We examined the samples for quality and contamination using FastQC, summarized with MultiQC. The mean Phred score across all reads and samples was above 35, and no primer contamination was detected. The mean number of reads per sample was 31.7 million (range: 18-45.4 million).

Data processing

To align the raw reads we performed a two-step STAR alignment (Dobin et al. 2013) using *Ipomoea nil* (Hoshino et al. 2016), which is a close relative of *I. hederacea,* as the reference genome. The genome indexes for *I. nil* were generated with STAR default parameters except for the 'genomeSAindexNbases' parameter, which was reduced from the default 14 to 13, given the size of the genome (Dobin et al. 2013). On average 92% of the reads were mapped successfully to the *I. nil* genome. We then used gffread (Pertea and Pertea 2020) to generate a transcript annotation file for use in Salmon (Patro et al. 2017), which we ran to quantify gene counts from the aligned samples. As we needed to compare across samples, we imported our quantified gene counts into R using *tximport* to compile the samples into a single data frame and transformed the read counts using the Trimmed-M Means transformation (Robinson and Oshlack 2010) using the *edgeR* package (Robinson et al. 2010; McCarthy et al. 2012; Chen et al. 2016). We then filtered out genes with very low expression where we would not expect to be able to detect differences between individuals that never flowered and those which flowered and successfully set seed; we did so using the *filterByExpr()* function in the edgeR package (Chen et al. 2016), which reduced the number of genes in our dataset to 2753.

It is common to correct estimates of gene expression for batch effects or effects due to the time of day of collection (Leek et al. 2012; see, e.g., Josephs et al. 2020). We elected not to do so for several related reasons. First, plants were planted in the field in randomized order, and as such tissue collection was in random order with respect to population of origin, genotype, and other known and unknown features of the samples. Second, our goal is to evaluate the relationship between gene expression, which may include time-of-day effects on the order of minutes to hours, and relative fitness, estimated from the cumulative number of seeds set by the end of the experiment, four weeks later; fitness itself reflects ~100 days of development, life history, and growth. Rather than focusing on individual genes or results from single models, we concentrate on broad patterns in the intersection of different types of models, each implemented with

permutation testing and cross-validation. Finally, preliminary analysis suggested that only one of 96 principal components of gene expression showed a significant rank correlation with order of collection; this PC was not predictive of relative fitness, and was not included in our analyses (*see below*).

Analytical Methods

Unsupervised modeling

We first performed principal component regression for dimensionality reduction without filtering out features. We standardized the data such that the transformed gene expression counts such that each gene had a standard deviation = 1 and mean = 0. We removed one outlier, which had set 86 seeds (mean seed set = 1.75, range = [0 – 24], excluding the outlier), to improve prediction. We calculated relative fitness for the remaining 96 individuals, by dividing the individual seed set by the mean seed set. Individuals that set no seeds were retained, having relative fitness of zero.

We performed a principal component analysis on the transformed and standardized data using *prcomp* from the *stats* package in R (R Core Team 2022). To determine which principal components (PCs) to include in the model we took a supervised regression approach, as lower PCs may still possess variation which may predict relative fitness well (Chong et al. 2018). We eliminated the last PC as the variance it described was approaching zero and was likely to introduce multicollinearity (Jolliffe 2002).

We first ran simple linear regression with relative fitness regressed on the PC scores of each individual and compiled the p-values of each principal component regression. We then sorted the PCs based on those p-values and used a 100-repeat 5-fold cross-validation regression with PCs of increasing p-values to determine the optimal number of PCs to include. We used the *train* function in the *caret* package to do so, saving the model parameters and performance measure of each iteration. We evaluated the error rates (both root mean squared and absolute error) and R-squared values of the models to determine the best model. Our model fitting procedure retained PCs based on their ability to predict relative fitness, and as such, PCs that explained relatively little variation in gene expression (the trailing principal components) but predicted fitness well were retained. We continued to add PCs until the R-squared, mean absolute error (MAE) and root mean squared error (RMSE) stopped improving.

To evaluate the importance of each gene we back-transformed the regression coefficients using the PCA rotations which resulted in the selection gradients (Chong et al. 2018), which account for selection on the expression of other genes. We then sorted the genes by their absolute coefficients to determine the 300 genes which demonstrated the strongest selection gradients, and thus were most influential in predicting relative fitness.

Supervised modeling

We performed supervised analyses in Python using scikit-learn (Pedregosa et al. 2011). We removed the outlier and split the data into a training set (60%, or 57 samples) and test set (40%,

or 39 samples). We again standardized the data to remove differences due to variance in expression among genes, using StandardScaler fit to the training data and then transforming both the training and test data sets based on the training data. For our supervised analyses we used classification algorithms, and as such classed our samples as having set seed (and thus having fitness > 0) or not (and having a fitness = 0).

For both models we used a grid search to tune the hyperparameters (such as learning rate or number of iterations) of our models, which runs models with each combination of selected model hyperparameter values (the “grid”) and compares the fit of each model. We started with coarser grids (a wider range of sparse hyperparameter options) to search a larger parameter space and tuning further with finer-scale grids. To manage overfitting in the model we used five-fold cross-validation during the grid search on the training data. Model score was calculated using the balanced accuracy score, as we had far fewer samples which set seed and balanced accuracy accounts for uneven class sizes. The balanced accuracy score is the average accuracy score of each group, that is, the fraction of true positive results out of the total positive cases and the fraction of the true negative results out of the total negative cases. In our case, the “positive” case is the “set seed” class, and the “negative” case is the “no seed set” class. We then tested the model on the 40% of the data withheld for testing and examined out-of-sample balanced accuracy and sensitivity of the less-common class (individuals that successfully set seed). Sensitivity is calculated as the true positive predicted results divided by the total positive cases.

We used a neural network, which is a simple type of deep learning algorithm, as our first classifier. We used a Multilayer Perceptron model, which has a layer of input nodes equal to the number of input variables, then has a predetermined number of "hidden" layers, each containing some set number of nodes, followed by the output layer and classifier. The nodes each represent some function, with input weights and a nonlinear "activation function" (e.g., a logistic or hyperbolic tangent function), the output of which connects to nodes in the next layer. Through "backpropagation", which seeks to minimize the loss function via gradient descent, the weights and activation functions of each node are tuned to optimize the model predictions (Hastie et al. 2009; James et al. 2021; Greener et al. 2022). The output layer then translates the results of the inner, hidden layers into the probability of the predicted outcomes, which is then classified based on those probabilities. Due to the nature of the algorithm, feature selection is not as straightforward as in simpler models. Through permutations of the data, we determined the most important features (i.e., genes), as the features which have the greatest reduction in the predictive metric of interest when randomized are those which are most important for prediction. We ran permutations of each gene 200 times and compared the balanced accuracy score to the permuted balanced accuracy score. The top 300 most important genes were then used for further analyses.

For our second classifier we used an ensemble method, which trains a large number of "weak learners" or models that are only marginally better than random, taken together to produce a "committee" model. We used a gradient tree boosting classifier, which is an ensemble method that uses some number of decision trees as weak learners, progressively adjusting weights to increase the importance of poorly classified observations from the previous tree, to build a final model weighted by the accuracy of each base model (Hastie et al. 2009). Unlike the neural

network, from this model we were able to directly extract the importance of genes, which is given as the Gini importance of the features in the model. We selected the top 278 features, as there were 197 ties in the importance of genes below this threshold.

GO analysis

We then sought to characterize the types of genes which were important in predicting "success" (i.e., having set seed) in each of the models individually and which types were common among all models, using Gene Ontology (GO) biological terms. We also performed GO enrichment analysis on the subsets of genes from our analyses. To do so we used Blast2Go (Conesa et al. 2005), first to perform BLAST using blastx-fast of the genes to an *Arabidopsis thaliana* genome (Lamesch et al. 2012), then EMBL-EBI Interpro scan (Cantelli et al. 2022) using default options, followed by mapping the associated biological GO terms and annotation by Blast2Go. We subset the most important genes in each of the three models, as described above and for each subset we repeated the BLAST, mapping and annotation of GO terms. We constructed GO term mapping for each set and extracted the union of all three to evaluate common GO terms. We then used goseq (Young et al. 2010) correcting for read count, comparing the entire set of genes from our samples to use as a reference for GO enrichment analyses of each of the model sets, using Fisher's Exact Test with a p-value of 0.05 as a filter to generate a set of over and under expressed GO terms. Again, we took the union of the results from the three models to look at commonly enriched GO terms.

Results

Unsupervised modeling

The Principal Component Analysis (PCA) transformed the 2753 standardized gene expression counts from 96 individuals into 96 linear combinations of the standardized gene expression measures. The first 10 principal components (PCs) describe 50.62% of the variation, and 68 PCs are required to explain >90% of the variation (recall that PCs are in descending order of the variance in gene expression they explain). Our best model included a total of 61 PCs with individual p-values <0.60, which minimized the RMSE and MAE, which maximized the R-squared value (Figure S1); note that these 61 PCs are *not* the first 61. The *caret* package, which we used to perform our repeat cross-validation regressions, uses the correlation between the predicted and observed response values as the R-squared value (Kuhn 2008). The model R-squared was 0.549, with an RMSE of 1.719 and MAE of 1.381, and included 55 PCs (Table 1).

We then transformed the regression coefficients for the PCs used in the final model back to the standard-deviation scaled gene expression counts using the rotations from the PCA (Chong et al. 2018) to evaluate the strength and distribution of gene expression selection gradients (***β***). We also calculated the standard error for each gene, and evaluated whether the absolute value of the selection gradient was greater than 1.96 x SE. Many (~59%) of the absolute selection gradients were greater than 1.96 x SE, suggesting that most of our estimates are significantly greater than sampling error, within the multivariate space spanned by these 61 PC axes. The distribution of *β*'s was centered on zero, with an approximately normal distribution, which suggests that there is directional selection for both increased and decreased gene expression (Figure S2). There was slightly more skew in the distribution of positive selection gradients (Figure S3). Overall, the magnitude of *β*'s was very small: an order of magnitude smaller than

the significant β's estimated for macroscopic traits in the same experiment, as reported by Henry and Stinchcombe (2023; Figure S2B). The only β for macroscopic, traditional traits that was within the range of the gene expression selection gradients was for anther-stigma distance, which was not significant. Given that these selection gradients are on individual genes and not the combined outcome of many genes, as is the case for the quantitative traits studied by Henry and Stinchcombe (20223), we expected the directional selection gradients on each gene's expression individually to be small.

Supervised modeling

We tested a variety of hyperparameters through a grid search with five-fold cross validation for both of our supervised models. The Multilayer Perceptron (MLP) model that best fit the data was one with two hidden layers, the first having 50 nodes and the second having 10 nodes. We used a ReLU activation function for the nodes (Glorot et al. 2011). For weight optimization we used a stochastic gradient descent solver, which approximates the gradient of each parameter using each sample to find the minima of the loss function (gradient descent). The learning rate (the step-size used in weight updating) for the model was adaptive, such that if the score does not improve for two iterations of training the learning rate is reduced, and thus narrow minima can be found without sacrificing initial efficiency. Further model hyperparameter details are available in the supplementary materials (Table S1). The MLP model had a balanced accuracy score of 0.593, and sensitivity to the successful seed set class of 0.364 (Table 1, Figure S4). Balanced accuracy is the average frequency of assigning an observation to its correct class, as described above. The sensitivity is the frequency of correctly assigning the "positive" label, in our case "set

seed" , out of all the positive cases. To determine the importance of the genes on the balanced accuracy score of the model we ran permutation tests on each of the models. For each gene, the expression values were permuted across samples 200 times and the class was then predicted based on the permuted data. The permuted predictions were compared to the original prediction such that genes with less influence on the prediction when shuffled are less important to the model overall.

Table 1: Metrics of model fit for both supervised and unsupervised methods. Multilayer Perceptron and Gradient boosting metrics are calculated from the testing data only. PCA regression metrics are from repeat five-fold cross-validation, and R-squared is calculated as the correlation between the observed and predicted data.

Model	Balanced Accuracy	Sensitivity	RMSE	R-squared
Multilayer Perceptron	0.593	0.364	-	-
Gradient Boosting	0.737	0.545	-	-
PCA regression	-	-	1.719	0.549

Our second supervised machine learning model was a gradient boosting classifier, which fits weakly predictive decision trees, progressively adjusting the parameter weights to improve classification of misclassified observations. We again tuned the hyperparameters of the model using a grid search with five-fold cross validation of the data. The best fitting model used a binomial deviance loss function, and the decision split quality was determined by Friedman adjusted mean-squared error (Friedman 2001). The models were regularized by using stochastic gradient boosting, which randomly subsamples the test data each iteration to reduce

overfitting in the final ensemble model. We used 1000 estimators, a constant learning rate of 1.0, and subsampled 40% of the data each iteration. The final model had an out-of-sample balanced accuracy score of 0.737, with sensitivity to the successful seed set class of 0.545 (Table 1, Figure S5). We extracted the 278 most important features based on their Gini coefficients from the model.

GO Analysis

Once we obtained a subset of most important genes for each model, we performed BLAST searches on gene sets using an *Arabidopsis thaliana* genome (Lamesch et al. 2012) as a reference, and then mapped and annotated GO terms. We generated combined GO maps for our gene subsets, filtering out intermediates, for each model. We found 8 GO terms common across all models, in addition to the higher level "biological process" term (Table 2, Figure 1A). They involved various stimulus response, metabolic, and developmental terms. These common terms highlight the importance in the response to both plant enemies (defense response and defense response to other organisms), and the abiotic environment (response to light stimulus, and development terms). More broadly, non-identical but related GO terms common across the models were associated with defense and stress response (GO:0009414, GO:0033554, GO:0006970, GO:0009651), metabolism GO:0044237,GO:0044238, GO:0006807,GO:0019222, GO:0071704, GO:0006629, GO:0043170) and reproductive structure development (GO:0048608, GO:0099402, GO:0009653, GO:0022414) (see Supplemental Material, Table S2 – Table S4).

Table 2: Shared GO terms from GO mapping of gene subsets. Gene subsets were composed of the most important genes in each mode, which were included in a BLAST search and annotated with GO terms. Response terms dominate, along with developmental and metabolic terms. The GO term for light response is also shared.

GO ID	GO Name
GO:0009791	post-embryonic development
GO:0009987	cellular process
GO:0050896	response to stimulus
GO:0044238	primary metabolic process
GO:0098542	defense response to other organism
GO:0051716	cellular response to stimulus
GO:0009416	response to light stimulus
GO:0006952	defense response

We then used *goseq* to adjust for count bias in our gene set to determine which GO terms were significantly enriched in each model set, using p <0.10 as our threshold. No specific terms were significant among all three models, but all three sets did include similar terms including those related to photosynthesis, seed maturation and dormancy, stress, and heavy metals (see Supplemental Material, Table S5 – Table S7). The shared specific GO terms between the MLP model and PCR were “regulation of catalytic activity” and “intracellular protein transport”. Shared specific GO terms between the PCR model and GB model were both related to reproduction, “vegetative to reproductive phase transition of meristem” and “seed maturation”. Shared specific GO terms between the MLP model and GB model involved metals (“response to iron ion starvation”, “response to copper ion”), stress (“response to endoplasmic reticulum stress”), seed dormancy (“negative regulation of seed germination”), photosynthesis (“granum assembly”), and more general GO terms related to metabolism and development (“regulation of protein catabolic process”, “developmental process”).

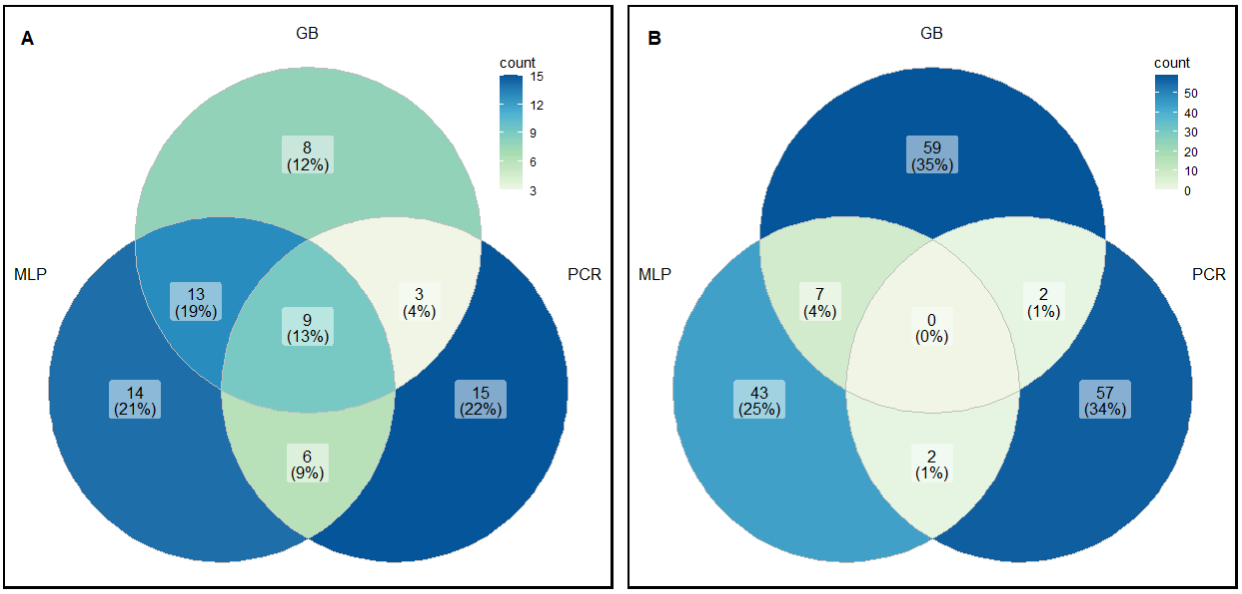

Figure 1: Venn diagrams of GO terms across models. GB - Gradient Boosting, MLP - Multilayer Perceptron, PCR - PCA regression. A) GO terms from mapping results of the important gene subsets from each model. Nine GO terms were shared among all models, one term was for the highest-level term “biological processes” and was removed from Table 2. B) GO terms from enrichment analyses of each model subset.

Shared Important Genes

The most important genes in each model were somewhat shared among models. The pairwise unions were ~10% of the totals of each model subset, ranging from 27 shared genes (from the PCA regression and Gradient Boosting classifier) to 32 shared genes (from PCA regression and Multilayer Perceptron classifier) (Figure S6). The union of all three gene subsets was only one gene, LOC109174332, a heavy-metal associated isoprenylated plant protein 7-like gene which is associated with transport or detoxification of heavy-metals in plants.

Heavy Metal Analysis of Soils

None of the metals in the screen (Table S8) were above government agricultural guidelines (Ontario Ministry of the Environment 2011), and were largely below the average for agricultural fields in the area (Frank et al. 1976), indicating that contamination of the soil was unlikely to be contributing to the importance of LOC109174332 in these analyses. While heavy-metal associated plant proteins (HIPP) have clearly been found to be associated with heavy-metal tolerance and detoxification, hence their name, they have also demonstrated responses to drought and cold tolerance (Barth et al. 2009; de Abreu-Neto et al. 2013; Zschiesche et al. 2015; Zheng et al. 2023), which interact with plant growth and development (Zschiesche et al. 2015; Guo et al. 2021).

Discussion

Connecting gene expression to fitness components can elucidate which genes are the targets of selection, but the high dimensionality, low sample size nature of gene expression data makes this goal computationally and statistically challenging. Machine learning (ML) techniques are well-suited to this sort of complex task, as feature selection, regularization to prevent overfitting (and thus poor prediction out of sample), and nonlinear functions are key aspects of many ML algorithms. In this study we successfully used gene expression in leaf tissue to predict the reproductive success of 96 *Ipomoea hederacea* individuals using statistical learning approaches. We find that after appropriate preprocessing of the data, a reasonable level of model accuracy can be attained using both supervised classification algorithms, with one model predicting correctly over 70% of the test cases. Similarly, unsupervised dimensionality reduction followed by regression, had a correlation of 0.55 between the predicted and observed fitness

values. Below, we discuss our results in light of the typical strength of selection gradients for gene expression, the insight of multiple analysis approaches, and GO categorization for understanding the relationship between expression and fitness components.

Distribution and strength of selection on gene expression

We found that gene expression is experiencing directional selection, and by using principal component analyses (an unsupervised machine learning approach) along with a repeat 5-fold cross validation approach to linear regression (a supervised machine learning approach) we can estimate selection gradients. By back-transforming the selection gradients for the principal component scores to return the selection gradients for each individual gene's expression, we can improve interpretability and further investigate the most relevant genes. We found that directional selection gradients for gene expression were overall symmetrically distributed around zero, and thus selection for both increasing and decreasing gene expression were approximately equal. The strength of selection was an order of magnitude lower than that of other traits measured in this field experiment (Henry and Stinchcombe 2023). Our finding was not unexpected, given low estimates of selection differentials in univariate analyses by Groen (2020) and by intuition — combined gene expression over time leads to higher level quantitative traits, and thus the individual expression components experience a fraction of the selection pressure compared to the total, end-point phenotype. We believe these results represent the first selection gradients for gene expression in the Lande-Arnold framework and illustrate the merits of PC regression for estimating selection on high-dimensional traits.

Leveraging information from multiple machine learning approaches

All our models had moderate predictive accuracy, which demonstrates the general applicability of machine learning in understanding selection on gene expression. The Multilayer Perceptron model (MLP) had the lowest balanced accuracy and sensitivity, despite its flexibility in processing complex, nonlinear relationships, which we might expect with gene expression data. It may have proved more accurate given a greater number of samples, but regardless, the added complexity in the model makes direct interpretations more difficult than the other models. We suggest considering whether more flexibility is truly required before undertaking a more complex model in lieu of a simpler one. The ensemble method of classification was able to predict with a moderately high degree of sensitivity given our limited data and had a balanced accuracy score of 0.737. These results were similar but not directly comparable to the regression model, which was a mix of unsupervised and supervised approaches, using cross-validation to determine the optimal number of principal components to include in the final regression. The noise inherent in tissue-level gene expression data paired with its high dimensionality and intrinsic modeling error can reduce confidence in the repeatability of analyses. However, comparing the results from disparate approaches, each optimized to reduce overfitting, improves the certainty in the overall importance of genes and biological processes which emerge repeatedly.

Implications from GO analysis

We found common GO terms through the union of the most influential genes in each of the models. These common terms suggest types of genes and gene networks whose expression is directly associated with fitness, and thus experiencing selection. The terms which were common across all models from the GO mapping results were largely related to seed development, stress and defense responses, as well metabolic processes and light response. Growth rate is

important for fitness, and nonlinear selection analysis has revealed that it has a slightly compensatory relationship with early flowering time (see Henry and Stinchcombe 2023), which is itself strongly selected for in *Ipomoea hederacea* grown in field environments (Simonsen and Stinchcombe 2010; Campitelli and Stinchcombe 2013b, Henry and Stinchcombe 2023). Metabolic and developmental processes are thus not unexpected terms to be relevant for predicting fitness related traits. Additionally stress and defense response genes were commonly important, indicating that the individuals better able to respond to biotic and abiotic stressors such as water deficiency and fungal pathogens may be better able to successfully reproduce. The single gene important across all three models, was a heavy-metal associated isoprenylated plant protein (LOC109174332). However, the analysis of heavy metals in the soils of our field site suggest that, if anything, heavy metal concentrations are lower than regional averages. GO enrichment analysis also suggests that genes related to iron ion starvation are more common than expected in the sets of influential genes from two of our models. Thus, it seems most likely that the significance of LOC109174332 lies in its role in mediating the stress responses and plant development in the field. These data suggest that the appropriate response to environmental challenges such as water deficiency and biotic interactions are key for the success of *Ipomoea hederacea* in the field.

Reproductive organ development GO terms were also found among the three models, which, given the observed strong selection on flowering time in this and other systems might be expected, and provides some level of assurance that the genes which were most important for prediction across our models are directly relevant for fitness. We found only one GO term significantly enriched in all three model subsets: photosynthesis. The field site for this experiment is 400 km north of the observed northern range limit and the significance of this

single GO term highlights the overwhelming importance of photosynthetic genes relevant for photoperiodic cues in development and flowering time (Shimizu et al. 2015). Our results here complement those of Groen et al. (2020), who also detected selection on photosynthesis in a field selection experiment on gene expression in rice under stressful environmental conditions.

Future applications and conclusions

Here we have demonstrated the applicability of machine learning in detecting gene functional groups and loci associated with fitness, despite a relatively small sample. The utilization of machine learning for associating genes with fitness related traits whilst including environmental influences has obvious applications for agricultural crop improvements. Finding genetic targets for increased seed or biomass production is a bedrock of modern plant breeding (Moose and Mumm 2008; International Wheat Genome Sequencing Consortium (IWGSC) 2018). Additionally, evaluating combinations of genetic features which have a concerted impact on fitness may be used for index selection, which may provide greater gains than improvement at single loci. More generally, similar applications of machine learning with gene expression could assist in identifying or validating associated higher-level phenotypes for trait-based studies or selection, especially for non-model organisms. We described the distribution of selection gradients on gene expression, but future evolutionary studies would benefit from comparing selection differentials to selection gradients in order to evaluate the relative strength of direct and indirect selection on genes.

Machine learning is a wide and developing field, with many potential applications in improving understanding of genetic and transcriptomic sequencing data in evolutionary biology. We demonstrated that simple and complex models were able to generate deeper understanding of

the genes and biological processes likely important in adaptive evolution. Our supervised ensemble method and principal component regression both outperformed the more complex neural network, suggesting that even the complexity of gene expression data may still be better suited for more traditional machine learning algorithms. Despite noisy data, there is still enough signal of fitness differences that models were moderately successful and taken together we were able to improve our confidence in the union of the results. Overall, machine learning algorithms, and their combined results, are effective in understanding the selective importance of biological pathways and gene expression in a multivariate context, in understanding the strength and distribution of selection across the transcriptome and can be harnessed to find candidate genes and quantitative traits of interest for further study.

Acknowledgements

We gratefully acknowledge financial support from NSERC Canada Discovery Grants (JRS, GAH), Weis-Zimmerman Graduate Fellowships in field biology (GAH), and the Swedish Collegium for Advanced Study (JRS). Comments from Stephen Wright, Jacqueline Sztepanacz, Art Weis, Megan Bontrager, and Joel McGlothlin improved the manuscript. Computational support was provided by the Digital Research Alliance of Canada. We thank Kate Brown and Radana Molnarova for support in the field at the Koffler Scientific Reserve.

References

Adrion, J. R., J. G. Galloway, and A. D. Kern. 2020. Predicting the landscape of recombination using deep learning. Mol. Biol. Evol. 37:1790–1808.

Battey, C. J., P. L. Ralph, and A. D. Kern. 2020. Predicting geographic location from genetic variation with deep neural networks. Elife 9.

Bright, K. L., and M. D. Rausher. 2008. Natural Selection on a Leaf-Shape Polymorphism in the Ivyleaf Morning Glory (*Ipomoea hederacea*). Evolution 62:1978–1990.

Campitelli, B. E., and J. R. Stinchcombe. 2013a. Natural selection maintains a single-locus leaf shape cline in Ivyleaf morning glory, *Ipomoea hederacea*. Mol. Ecol. 22:552–564.

Campitelli, B. E., and J. R. Stinchcombe. 2014. Population dynamics and evolutionary history of the weedy vine *Ipomoea hederacea* in North America. G3 4:1407–1416.

Campitelli, B. E., and J. R. Stinchcombe. 2013b. Testing potential selective agents acting on leaf shape in *Ipomoea hederacea*: predictions based on an adaptive leaf shape cline. Ecol. Evol. 3:2409–2423.

Cantelli, G., A. Bateman, C. Brooksbank, A. I. Petrov, R. S. Malik-Sheriff, M. Ide-Smith, H. Hermjakob, P. Flicek, R. Apweiler, E. Birney, and J. McEntyre. 2022. The European Bioinformatics Institute (EMBL-EBI) in 2021. Nucleic Acids Res. 50:D11–D19.

Chang, D., and A. Keinan. 2014. Principal component analysis characterizes shared pathogenetics from genome-wide association studies. PLoS Comput. Biol. 10:e1003820.

Chen, Y., A. T. L. Lun, and G. K. Smyth. 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Res. 5:1438.

Choi, Y., T. T. Liu, D. G. Pankratz, T. V. Colby, N. M. Barth, D. A. Lynch, P. S. Walsh, G. Raghu, G. C. Kennedy, and J. Huang. 2018. Identification of usual interstitial pneumonia pattern using RNA-Seq and machine learning: challenges and solutions. BMC Genomics 19:101.

Chong, V. K., H. F. Fung, and J. R. Stinchcombe. 2018. A note on measuring natural selection on principal component scores. Evol Lett 2:272–280.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: a

universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29:1189–1232.

Gilad, Y., A. Oshlack, and S. A. Rifkin. 2006. Natural selection on gene expression. Trends Genet. 22:456–461.

Glorot, X., A. Bordes, and Y. Bengio. 2011. Deep sparse rectifier neural networks. Pp. 315–323 *in* Proc. 14th Int. Conf. Artif. Intell. Statis.

Greener, J. G., S. M. Kandathil, L. Moffat, and D. T. Jones. 2022. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 23:40–55.

Groen, S. C., I. Ćalić, Z. Joly-Lopez, A. E. Platts, J. Y. Choi, M. Natividad, K. Dorph, W. M. Mauck 3rd, B. Bracken, C. L. U. Cabral, A. Kumar, R. O. Torres, R. Satija, G. Vergara, A. Henry, S. J. Franks, and M. D. Purugganan. 2020. The strength and pattern of natural selection on gene expression in rice. Nature 578:572–576.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning. Springer New York.

Henry, G. A., and J. R. Stinchcombe. 2022. Strong selection is poorly aligned with genetic variation in *Ipomoea hederacea*: implications for divergence and constraint.

Hersch, E. I., and P. C. Phillips. 2004. Power and potential bias in field studies of natural selection. Evolution 58:479–485.

Hoekstra, H. E., J. M. Hoekstra, D. Berrigan, S. N. Vignieri, A. Hoang, C. E. Hill, P. Beerli, and J. G. Kingsolver. 2001. Strength and tempo of directional selection in the wild. Proc. Natl.

Acad. Sci. U. S. A. 98:9157–9160.

Hoshino, A., V. Jayakumar, E. Nitasaka, A. Toyoda, H. Noguchi, T. Itoh, T. Shin-I, Y. Minakuchi, Y. Koda, A. J. Nagano, M. Yasugi, M. N. Honjo, H. Kudoh, M. Seki, A. Kamiya, T. Shiraki, P. Carninci, E. Asamizu, H. Nishide, S. Tanaka, K.-I. Park, Y. Morita, K. Yokoyama, I. Uchiyama, Y. Tanaka, S. Tabata, K. Shinozaki, Y. Hayashizaki, Y. Kohara, Y. Suzuki, S. Sugano, A. Fujiyama, S. Iida, and Y. Sakakibara. 2016. Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. Nat. Commun. 7:13295.

Huang, C., E. A. Clayton, L. V. Matyunina, L. D. McDonald, B. B. Benigno, F. Vannberg, and J. F. McDonald. 2018. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. Sci. Rep. 8:16444.

International Wheat Genome Sequencing Consortium (IWGSC). 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2021. Introduction to Statistical Learning: With Applications in R. Springer.

Jolliffe, I. T. 2002. Principal Component Analysis. Springer New York.

Josephs, E. B., Y. W. Lee, C. W. Wood, D. J. Schoen, S. I. Wright, and J. R. Stinchcombe. 2020. The Evolutionary Forces Shaping Cis- and Trans-Regulation of Gene Expression within a Population of Outcrossing Plants. Mol. Biol. Evol. 37:2386–2393.

Josephs, E. B. 2021. Gene expression links genotype and phenotype during rapid adaptation.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245–261.

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13:8–17.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28:1–26.

Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala. 2012. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40:D1202–10.

Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28:882–883.

Liao, B.-Y., and M.-P. Weng. 2015. Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice. Proc. Natl. Acad. Sci. U. S. A. 112:4707–4712.

Lunetta, K. L., L. B. Hayward, J. Segal, and P. Van Eerdewegh. 2004. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. 5:32.

Lu, Q., Y. Hu, J. Sun, Y. Cheng, K.-H. Cheung, and H. Zhao. 2015. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci. Rep. 5:10576.

Lu, Q., R. L. Powles, Q. Wang, B. J. He, and H. Zhao. 2016. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in Genome Wide Association Studies. PLoS Genet. 12:e1005947.

Maciukiewicz, M., V. S. Marshe, A.-C. Hauschild, J. A. Foster, S. Rotzinger, J. L. Kennedy, S. H. Kennedy, D. J. Müller, and J. Geraci. 2018. GWAS-based machine learning approach to

predict duloxetine response in major depressive disorder. J. Psychiatr. Res. 99:62–68.

McCarthy, D. J., Y. Chen, and G. K. Smyth. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40:4288–4297.

Mieth, B., M. Kloft, J. A. Rodríguez, S. Sonnenburg, R. Vobruba, C. Morcillo-Suárez, X. Farré, U. M. Marigorta, E. Fehr, T. Dickhaus, G. Blanchard, D. Schunk, A. Navarro, and K.-R. Müller. 2016. Combining multiple hypothesis testing with machine learning increases the statistical power of Genome-wide Association Studies. Sci. Rep. 6:36671.

Moose, S. P., and R. H. Mumm. 2008. Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol. 147:969–977.

Morrissey, M. B. 2014. In search of the best methods for multivariate selection analysis. Methods in Ecology and Evolution 5:1095−1109.

O'Connell, G. C., P. D. Chantler, and T. L. Barr. 2017. Stroke-associated pattern of gene expression previously identified by machine-learning is diagnostically robust in an independent patient population. Genom Data 14:47–52.

O'Connell, G. C., A. B. Petrone, M. B. Treadway, C. S. Tennant, N. Lucke-Wold, P. D. Chantler, and T. L. Barr. 2016. Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke. NPJ Genom Med 1:16038.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14:417–419.

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

Pertea, G., and M. Pertea. 2020. GFF Utilities: GffRead and GffCompare. F1000Res. 9.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rifkin, S. A., D. Houle, J. Kim, and K. P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438:220–223.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Robinson, M. D., and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11:R25.

Rockman, M. V., and L. Kruglyak. 2006. Genetics of global gene expression. Nat. Rev. Genet. 7:862–872.

Schrider, D. R., J. Ayroles, D. R. Matute, and A. D. Kern. 2018. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. PLoS Genet. 14:e1007341.

Schrider, D. R., and A. D. Kern. 2018. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 34:301–312.

Shimizu, H., K. Katayama, T. Koto, K. Torii, T. Araki, and M. Endo. 2015. Decentralized circadian clocks process thermal and photoperiodic cues in specific tissues. Nat Plants 1:15163.

Simonsen, A. K., and J. R. Stinchcombe. 2010. Quantifying Evolutionary Genetic Constraints in the Ivyleaf Morning Glory, *Ipomoea hederacea*. Int. J. Plant Sci. 171:972–986.

Small, S. T., F. Labbé, N. F. Lobo, L. L. Koekemoer, C. H. Sikaala, D. E. Neafsey, M. W. Hahn, M. C. Fontaine, and N. J. Besansky. 2020. Radiation with reticulation marks the origin of a major malaria vector. Proc. Natl. Acad. Sci. U. S. A. 117:31583–31590.

Smith, A. M., J. R. Walsh, J. Long, C. B. Davis, P. Henstock, M. R. Hodge, M. Maciejewski, X. J. Mu, S. Ra, S. Zhao, D. Ziemek, and C. K. Fisher. 2020. Standard machine learning approaches outperform deep representation learning on phenotype prediction from

transcriptomics data. BMC Bioinformatics 21:119.

Smith, C. C. R., S. Tittes, P. L. Ralph, and A. D. Kern. 2023. Dispersal inference from population genetic variation using a convolutional neural network. Genetics 224.

Stock, A. J., B. E. Campitelli, and J. R. Stinchcombe. 2014. Quantitative genetic variance and multivariate clines in the Ivyleaf morning glory, *Ipomoea hederacea*. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369:20130259.

Sun, S., B. Dong, and Q. Zou. 2021. Revisiting genome-wide association studies from statistical modelling to machine learning. Brief. Bioinform. 22.

Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11:R14.

Xue, A. T., D. R. Schrider, A. D. Kern, and Ag1000g Consortium. 2021. Discovery of ongoing selective sweeps within *Anopheles* mosquito populations using deep learning. Mol. Biol. Evol. 38:1168–1183.