1 Application to Data

1.1 Methods

1.1.1 Feature Description

Identifying essential genes that is, genes without which the organism cannot survive in the yeast Saccharomyces cerevisiae (Seringhaus et al., 2006) and related species has been an area of investigation. Several features have been used to predict essential genes including both features of the DNA sequence such as GC content, and predicted features of the proteins translated from the genes such as hydrophobicity and subcellular localization. All of the features may be expected to be have predictive value for essentiality, but none are particularly strong predictors. By combining these measures, the hope is to predict essential genes with a much higher degree of accuracy than would be possible with any one measure alone.

To predict essentiality of the genes in yeast Saccharomyces cerevisiae, our analysis uses features associated with each gene from two sources: 1) fourteen sequence-derived features from the Seringhaus, et al., 2006 study, and 2) eight additional features from the Ensembl website (ref). We added new features to check how sensitive the results were to the feature set. Feature definitions are listed in Table reftab:definitions. Three of 22 features (vacuole, in how many of 5 proks blast, and intron) were removed from the analysis due to low content (less than 5% of non-zero values). At lower training sizes, their low content resulted in deterministic models because the randomization employed multiple seeds until the design matrix had no columns of zeroes. Genes which had values for this low-content features would have been selected more frequently than genes without information in these features.

1.1.2 Cross-Validation Strategy

In the unsupervised simulation, semi-supervised was compared against the unsupervised method described in section 2.4. We trained our methods from the set of 769 essential genes (positive labels) of the total 3500 genes. Additionally, we contrasted all 22 features against a subset of 14 sequenced-derived features as predictors of essentiality (see Section 1.1.1). Training set sizes were based on increments of 5 with minimum set sizes greater than the number of features to prevent rank deficiency in training sets. Iterations (n=30) were used to average the AUC or other metrics used to evaluate performance.

The cross-validation strategy for the supervised case incorporates an unbalanced strategy to the test set (Figure 1) along with a contamination rate. For an unbalanced design, test sets utilize the remaining genes not used in the training sets rather than a balanced strategy which matches training and testing set sizes. The unbalanced strategy was chosen because, in practice, an investigator would typically want to test all the remaining genes for essentiality rather than just a subset of genes. Another concept interwoven into the analysis is contamination. Contamination considers the dilemma from falsely assigned genes in the training sets. The semi-supervised and unsupervised methods do not consider negative labels in their computations and, thus, are unaffected by contamination. For supervised methods, positive labels in the training set are mixed with negative labeled genes for analysis when varying percentages of contamination are introduced.

In the supervised simulation, semi-supervised was compared against three supervised methods (LASSO, SVM, and Random Forest) at low training set sizes. AUC performance of these four methods was compared across training set sizes between 1% (n=35) and 10% (n=350) from all 3500 genes. Genes randomly chosen for the supervised training sets reflect the same ratio of positive and negative labels as seen in the full data set. Among the 3500 yeast genes, there are 769 essential genes resulting in a 21% ratio. Therefore, as an example, at 1% training size, 35 randomly chosen genes contained 7 positive labels (21% of 35) and 28 negative labels for supervised methods, while semi-supervised methods analyzed 35 positively labeled genes. In order to mimic contamination, negative labels were reassigned a positive label at rates of 0%, 20%, and 50%.

For all results, unique initial seeds were chosen based on the iteration number, training set size, and contamination (for supervised comparison only). Iterations were increased to 100 to better discriminate effects at low training sets. Once the cross-validation data was generated by a seed, the same data was used to compare each method.

1.1.3 Algorithms

All simulations were performed in R version 3.3.3. The semi-supervised and unsupervised analysis utilized functions from the lcmix package. The lcmix package developed and implemented in the previous paper by Dvorkin, Biehl, and Kechris A Graphical Model Method for Integrating Multiple Sources of Genome-scale Data and can be downloaded from http://r-forge.r-project.org/projects/lcmix/. LASSO was performed using the glmnet command in the glmnet package (ref Hastie and Qian). Using cv.glmnet, k-fold cross validation optimized the minimum lambda for the LASSO function. SVM analysis used the svm command under the e1071 package (ref David Meyer). Various runs using different criteria revealed a radial kernel density and C-classification optimized AUC performance. Random Forest was performed with the randomForest command under the randomForest package (ref Breiman, L.) All supervised predictions used the predict command in the stats package.

1.1.4 Performance

The AUC mean, variance, and CV (median absolute deviation/median) of the three supervised methods were contrasted against semi-supervised method. Because LASSO outperformed the other supervised models in AUC across all training set sizes and contamination rates, a closer evaluation of its performance was compared with semi-supervised method. In order to fairly contract LASSO performance to semi-supervised, the prediction scores were re-scaled to be between 0 and 1. Precision, recall, and f-measure further discriminated the two methods with four rescaled prediction score cutoffs including the median and prediction scores of 0.5, 0.8, and 0.95. The median cutoff is a relative measure based on the data while the other three cutoffs are absolute. The f-measure was calculated from the average precision and recall at each training set size from 1% to 5%.

1.2 Results

First, we describe each of the features and report the univariate performance in predicting essentiality for each one in Table 2. Based on the range of AUC, many of the features have low predictive value on their own but in the following comparisons we will explore their combined predictive power. We used cross-validation simulations to compare our hierarchical mixture model semi-supervised method with unsupervised and supervised methods. We hypothesize that semi-supervised method outperforms both unsupervised methods at any training set size and supervised methods at low training set sizes especially when positive labels are contaminated. Using the unsupervised comparison, we also explored the effect of different training sets or features.

1.2.1 Unsupervised Comparison

The complete essentiality data for Saccharomyces cerevisiae contains n=769 positive labeled genes [REF]. To explore whether our conclusions were sensitive to the choice of features, we first used only 14 sequence-derived features from REF and then a larger set of 22 features, which included the 14 sequenced-derived features and additional features collected from Ensembl (see Methods). Semi-supervised performs better than unsupervised for AUC regardless of predicting with 14 sequence-derived or all 22 features or training set size (Figure 2). The variance of the AUC for both methods increases as training size increases when training on all essential genes. This is expected as the test set is relatively larger and more constant with the smaller training sets. The eight additional features added from Ensembl Biomart generally improves AUC performance and decreases variance for both methods.

1.2.2 Supervised Comparison

Next, we compared the semi-supervised method with a supervised strategy using all essential genes for the training set and all 22 features. Supervised algorithms require both positive and negative labels. Therefore, we

picked a random set of the non-essential genes to be the negative labels (see Methods), but also included some contamination (some essential genes labeled as negatives in the training set) since in practice, the complete set of negative labels will not be known in many situations.

LASSO and semi-supervised outperforms the other two supervised methods - SVM and Random Forest (Figure 3). At low training sizes (¡ 2%, n = 70), semi-supervised method has a higher mean AUC than the three supervised methods. LASSO does not match the stability (lower variance) of semi-supervised until around 5% (n=175) training set size. However, for larger training set sizes, the AUC variance of semi-supervised increases while variance from LASSO slightly decreases. As contamination increases, all three supervised methods decrease in performance. At 50% contamination, semi-supervised method bests all methods across all training set sizes (up to 10%). The CV (median absolute deviance/median) for semi-supervised is lower than LASSO across all contamination levels and training set sizes up to 5% (Figure 4).

1.2.3 Semi-supervised versus LASSO Performance

To compare the best performing supervised method, LASSO, to semi-supervised method, prediction scores were rescaled to be between 0 and 1. At 1% training level, LASSO kernel densities of prediction scores exhibit an unimodal distribution while semi-supervised methods exhibit bi- or multi-modal behaviors (Figure 5). Uni-modal behavior makes it more difficult to find better separation of gene types (e.g., essential versus non-essential). As training level increases, LASSO kernel densities of prediction scores continue to exhibit unimodal distributions while semi-supervised methods maintain their multimodal behaviors.

Focusing on 0% contamination in Figure 6, the three absolute cutoffs (50%, 80%, and 95%) reveal a higher recall across all training set sizes for semi-supervised and the median cutoff shows semi-supervised outperforming LASSO up to 3% at which they become comparable. Also, up to 3%, semi-supervised outperforms LASSO in precision at the median cutoff. Precision generally increases as the absolute cutoff increases with LASSO besting semi-supervised as training set size increases. Contamination reduces all three performance measures (precision, recall, f-measure) for LASSO across training set sizes from 1% to 5% and all four cutoffs. Irrespective of contamination, the f-measure for semi-supervised outperforms LASSO for all training set sizes and cutoffs.

1.3 Potential Discussion Points

Due to the inherent capacity of supervised methods to utilize both positive and negative labels, they have natural advantages over semi-supervised methods which only handles positive labels. The contamination is a strategy to emulate a real-world scenario that a researcher may know a certain number of positive labels for genes in their experiment but are unsure if the remaining genes are truly negative.

Lasso may have an advantage over the other methods because it can reduce the effect of poorly predicting variables by collapsing their betas to 0. Semi-supervised method does not utilize negative labels, thus, alleviating the turbulence caused by contamination, when negative labels are unknown or tentative.

Supervised methods such as LASSO with unimodal distributions do not intrinsically show a clear optimal cutoff compared to the multi-modality of semi-supervised predicted probabilities. The multimodality of semi-supervised prediction scores provides more natural cutoffs than the unimodal distribution from LASSO.

With posterior probabilities ranging from near 0 and 1, the mid-range for both methods is near 0.5. Because of the heavily, skewed low probabilities in the unimodal distribution from LASSO (Figure 4), the median would divide the set somewhere on the backside of the slope, greater than the maximum but less than 0.5. The median and mid-range for semi- supervise tend to fall near a local minimum, a useful indicator for separating distributional behaviors and cannot be evaluated in unimodal distributions. The most dynamic difference between the two methods is the behavior of the recall performance. Recall measures how many positive labels were predicted out of the true number of positive labels. In the mid-range cutoff for LASSO, recall would naturally be lower than the median cutoff due to the small area in the right tail greater than 0.5. The expected increase in precision wasn't strong enough to outperform semi-supervise in the combined f-measure.

Say something about how the 1-5% training set sizes correspond to realistic values of positive labels #s? (for example the Drosophila study and pre-2002 #s?)

Supervised is constant. increasing variance, and primarily driven by negative labels at higher training sets.

Table 1: Feature Definitions.

	Abbreviation	Description	Type	Family	K
Sequence Derived Features	cytoplasm	Predicted subcellular location: cytoplasm	binary	bernoulli	2
	er	Predicted subcellular location: er	binary	bernoulli	2
	mitochondria	Predicted subcellular location: mitochondria	binary	bernoulli	2
	nucleus	Predicted subcellular location: nucleus	binary	bernoulli	2
	vacuole	Predicted subcellular location: vacuole	binary	bernoulli	2
	other	Predicted subcellular location: other	binary	bernoulli	2
	tm helix	Number of predicted transmembrane helices	integer	neg bin	2
	cai	Codon adaptation index	[real]	gamma	2
	l aa	Length of putative protein in amino acid	integer	neg bin	3
	nc	Effective number of codons	(real)	normal	2
	gravy	Hydrophobicity	(real)	normal	2
	gc	% GC content	[real]	gamma	2
	close ratio	% codons one-third base pairs from stop codon	[real]	gamma	2
	rare aa ratio	% of rare aa in translated ORF	[real]	gamma	2
Additional Features	intxn partners	Number of interaction proteins	integer	neg bin	3
	6 yeast blast	Number of related genes in 6 species of yeast	integer	poisson	2
	blast yeast	Number of related genes in yeast BLAST	integer	neg bin	2
	dovexpr	Dov Expression	(real)	pearson	3
	chromosome	Chr number	integer	poisson	2
	chr position	Chr position as % of chromosome length	[real]	gamma	2
	5 proks blast	Number of related genes in 5 prokaryotes BLAST	integer	poisson	2
	intron	Contains an intron in DNA/RNA sequence	binary	bernoulli	2

Table 2: **Description of Features.** The sequence-derived features were compiled by Seringhaus (2001 ref). Additional features were assembled from the Gerstein labs (ref). Dov expression is the normalized difference between absolute mRNA expression levels (Jansen R (2002)).

References

Gerstein M Jansen R, Greenbaum D. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12(1):37–46, 2002.

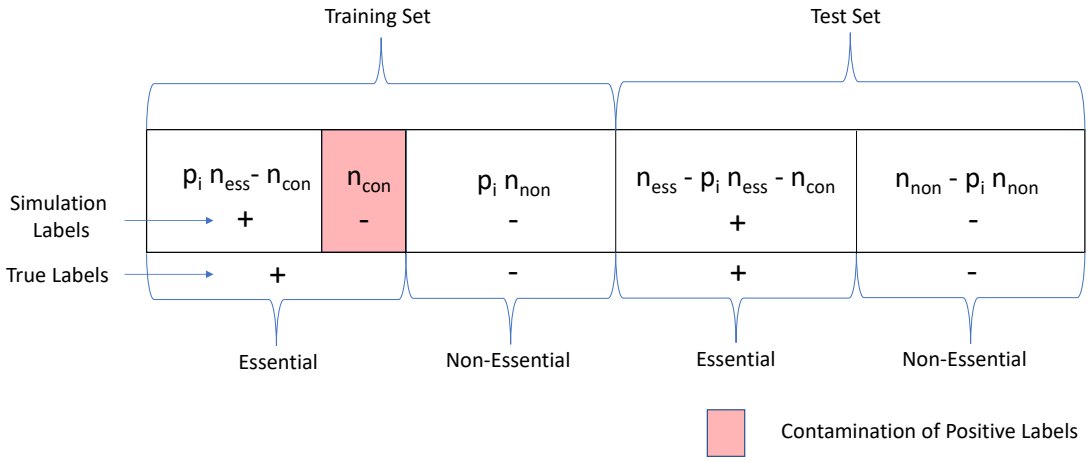
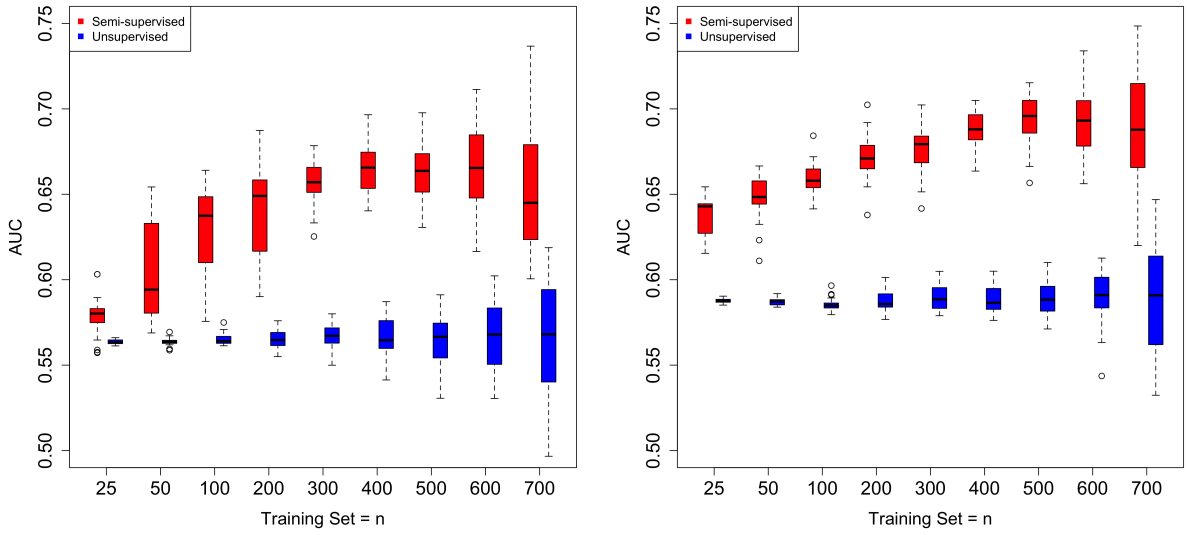


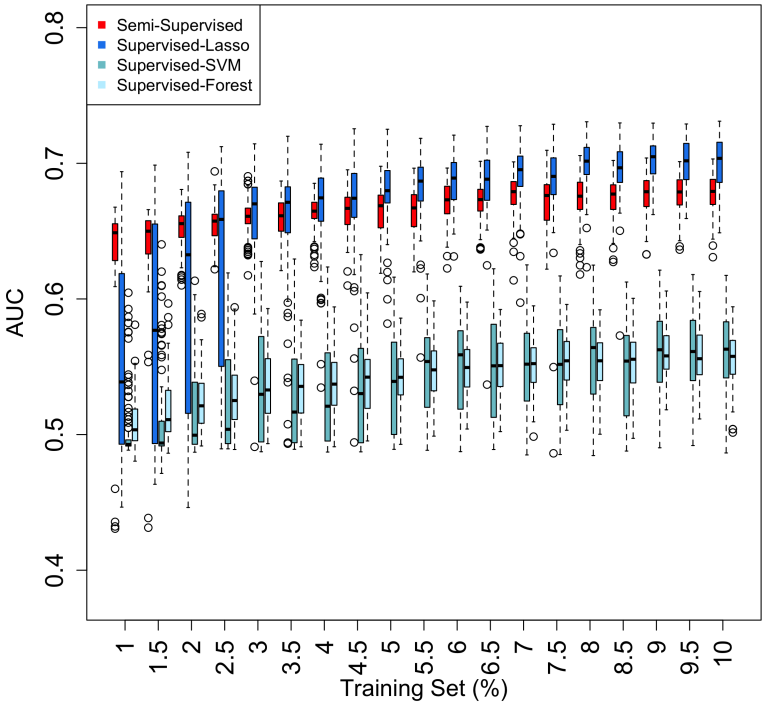
Figure 1: **Diagram of unbalanced design describing true label contamination in training sets for supervised methods.** p_i, n_{ess}, n_{con}, and n_{non} represent training set size, total number of essential genes, number of contaminated genes, total number of non-essential genes, respectively. n_{con} is determined by contamination percentage of training set essential genes. Training and test sets are indicated above while true label of gene essentiality is indicted below figure. Simulation label describes the label assignment for analysis. Shaded area indicates the contamination of training set essential genes where simulation and true labels differ.



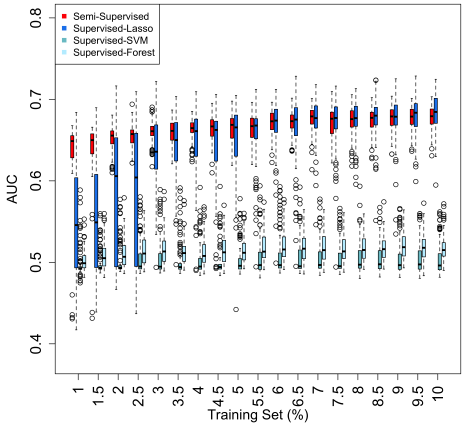
(a) Sequence derived features.

(b) All features.

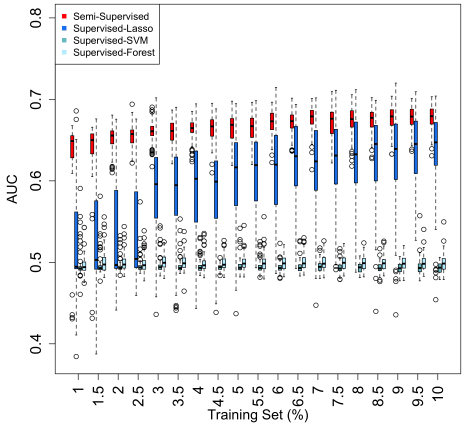
Figure 2: **Boxplots of AUC at various training sizes of 769 essential genes using sequence derived (14) or all (22) features as predictors.** Semi-supervised and unsupervised method are shown in red and blue, respectively.



(a) 0% Contamination

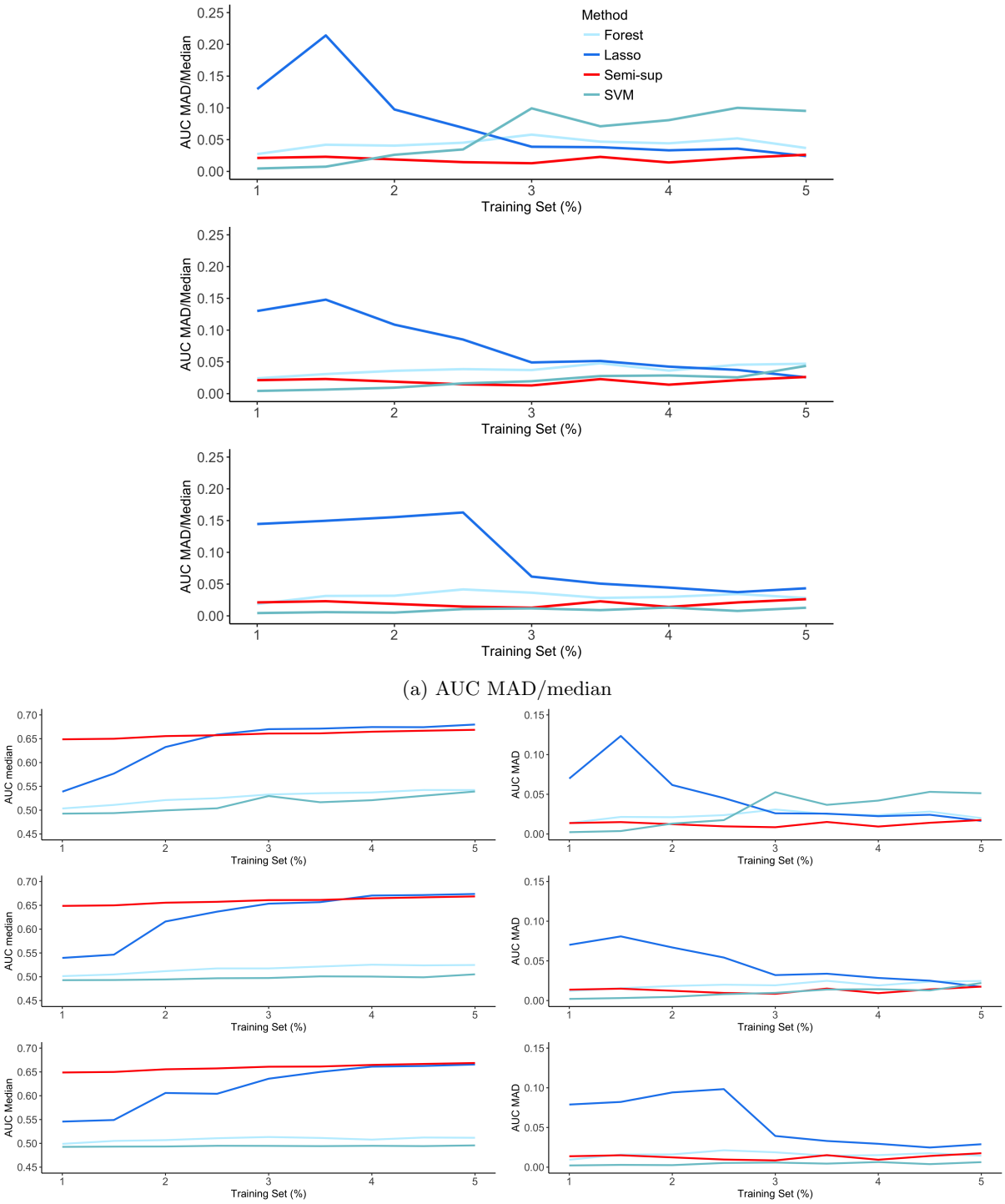


(b) 20% Contamination



(c) 50% Contamination

Figure 3: **AUC comparison between semi-supervised and supervised methods at various training sizes with all essential genes in yeast using 19 features as predictors.** 100 iterations were executed at training sets percentages (1, 1.5, 2, ... ,10) for all four methods and negative contamination levels (0% (a), 20% (b), and 50% (c)). Semi-supervised method is shown in red while the supervised methods (LASSO, SVM, and Random Forest) are shown in blue, aquamarine, and light blue, respectively.

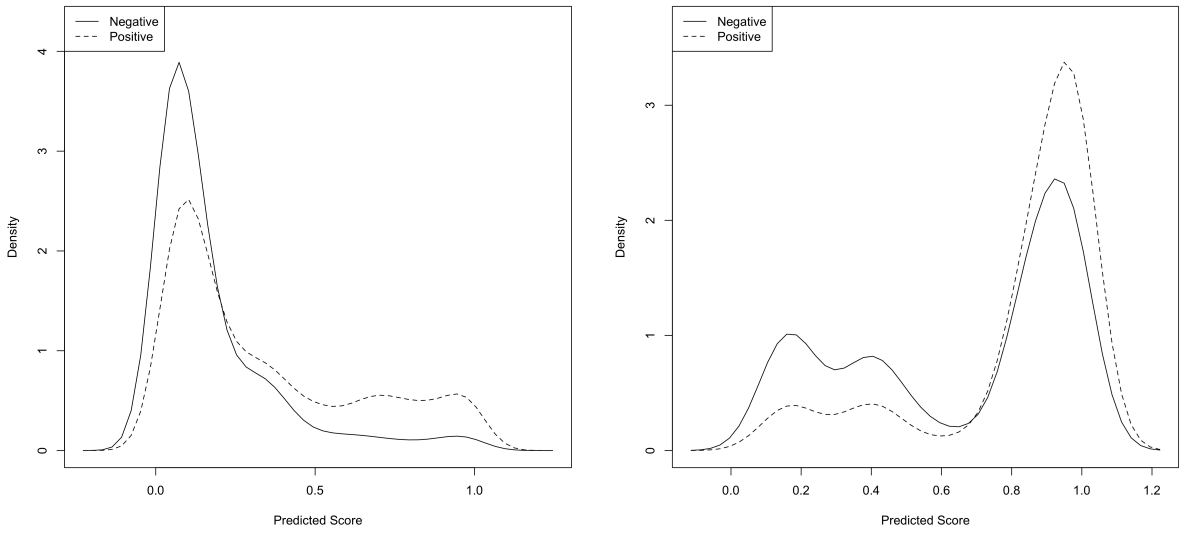


(a) AUC MAD/median

(b) AUC Median

(c) AUC Median Absolute Deviation

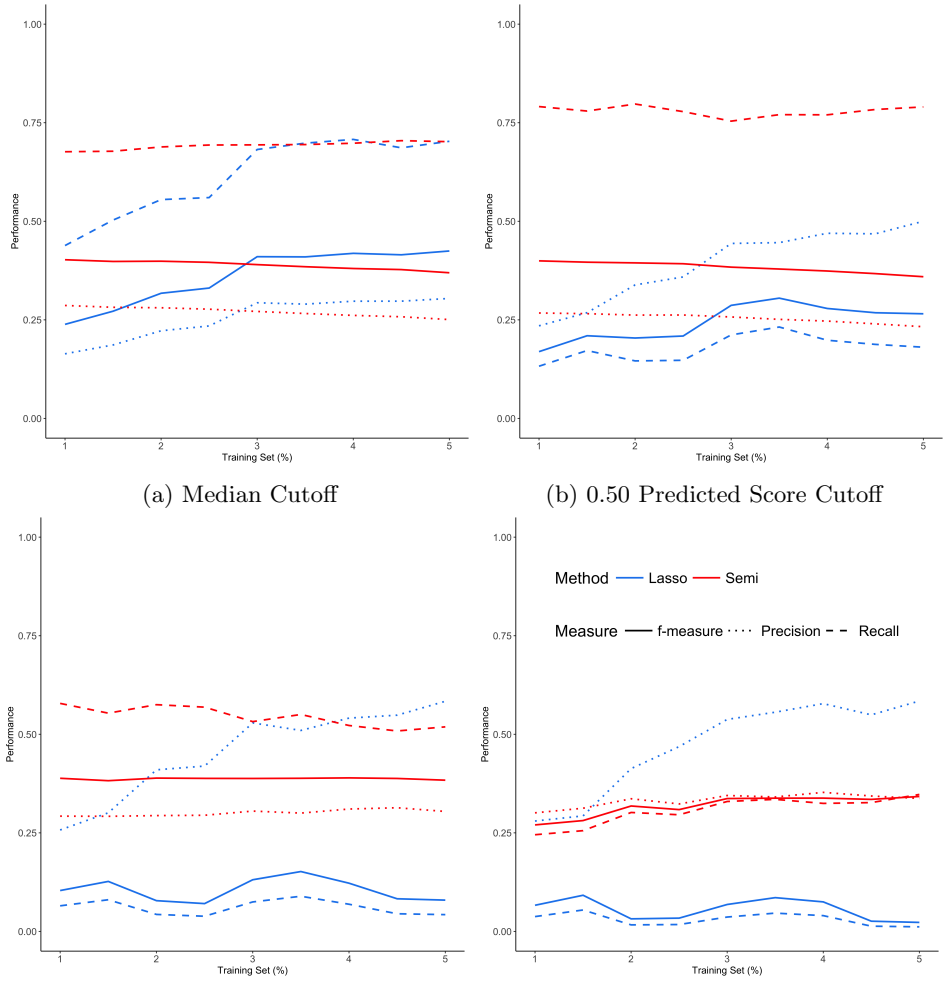
Figure 4: **Evaluation of summary statistics comparing semi-supervised and supervised methods.** 100 iterations were executed at training sets (1%, 1.5%, 2%,...,5%) for all four methods and negative contamination levels (0%, 20%, and 50%). Semi-supervised method is shown in red while the supervised methods (LASSO, SVM, and Random Forest) are shown in blue, aquamarine, and light blue, respectively. The AUC mean (a), variance (b), and CV (c) contrasts the four methods at each training set size across three contamination rates. CV is calculated as the Median Absolute Deviation (MAD) / Median



(a) LASSO

(b) Semi-Supervised

Figure 5: **Density plot of predicted scores juxtaposing semi-supervised and LASSO methods.** LASSO (a) and Semi-Supervised (b) methods display kernel densities for true positive (dashed) and negative (solid) labels at the 1% training set level and 0% contamination rate.



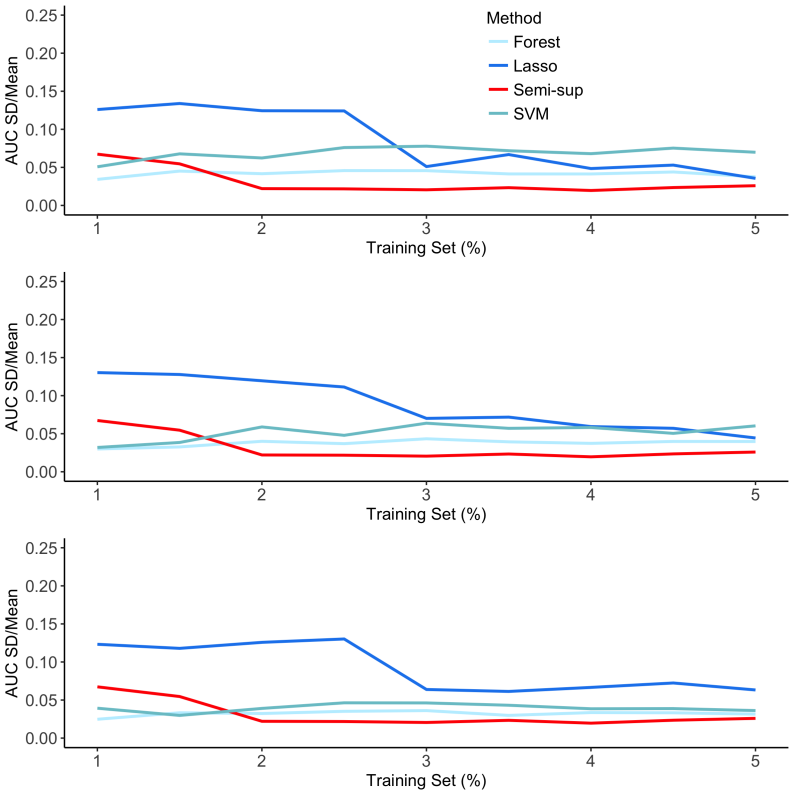
(a) Median Cutoff

(b) 0.50 Predicted Score Cutoff

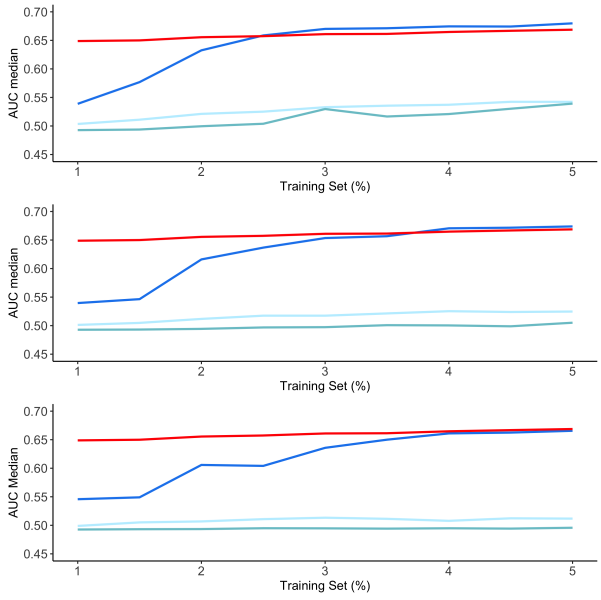
(c) 0.80 Predicted Score Cutoff

(d) 0.95 Predicted Score Cutoff

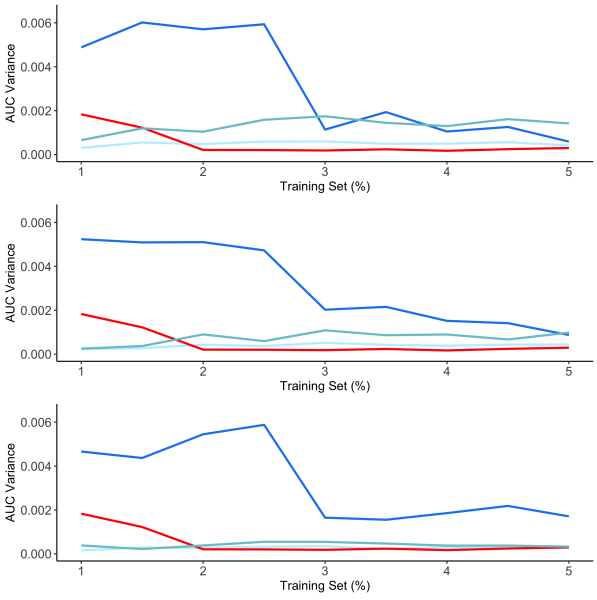
Figure 6: **Performance of Semi-supervised and LASSO methods with predicted score cutoffs at median, 0.50, 0.80, and 0.95 at 0% contamination.** Semi-supervised method is shown in red while LASSO is shown in blue. Precision, recall, and f-measures are represented by dotted, dashed, and solid lines, respectively. The median is a relative cutoff while the other cutoffs represent absolute cutoffs with re-scaled predicted probabilities.



(a) AUC SD/mean

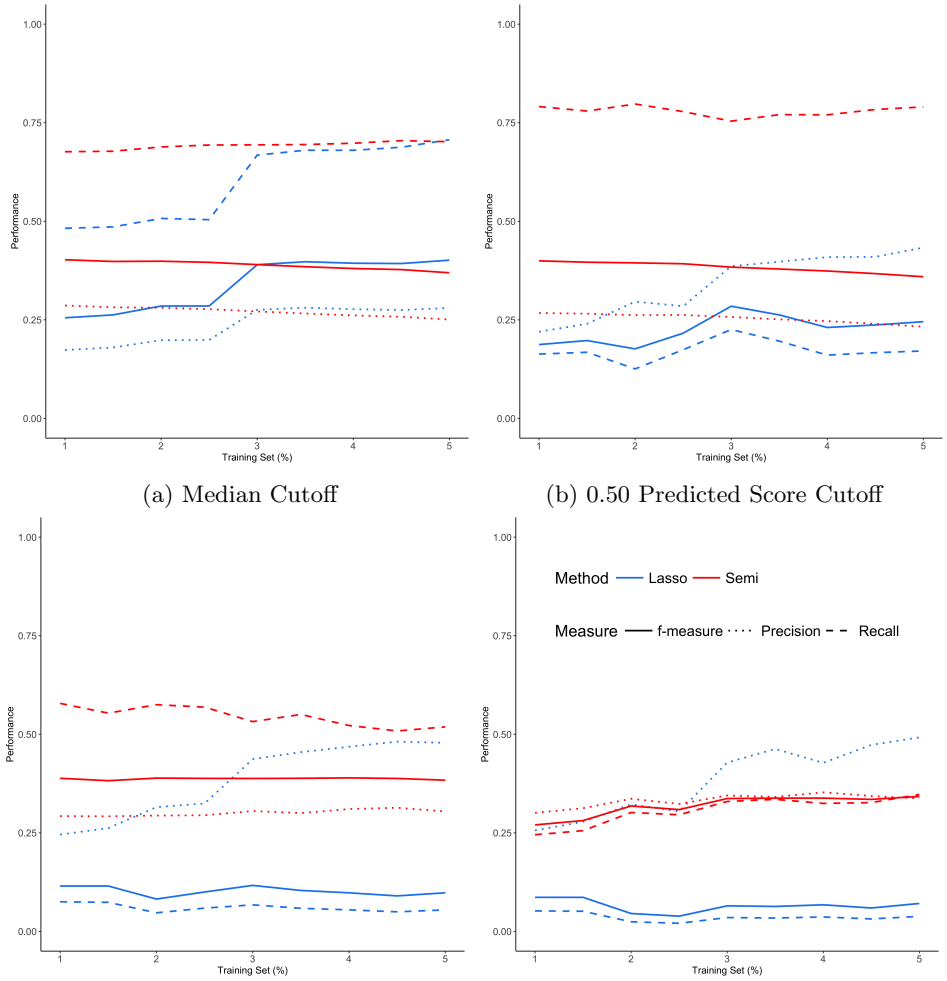


(b) AUC Mean



(c) AUC Variance

Supplementary Figure 1: **Evaluation of summary statistics comparing semi-supervised and supervised methods.** 100 iterations were executed at training sets (1%, 1.5%, 2%,...,5%) for all four methods and negative contamination levels (0%, 20%, and 50%). Semi-supervised method is shown in red while the supervised methods (LASSO, SVM, and Random Forest) are shown in blue, aquamarine, and light blue, respectively. The AUC mean (a), variance (b), and CV (c) contrasts the four methods at each training set size across three contamination rates. CV is calculated as the Median Absolute Deviation (MAD) / Median



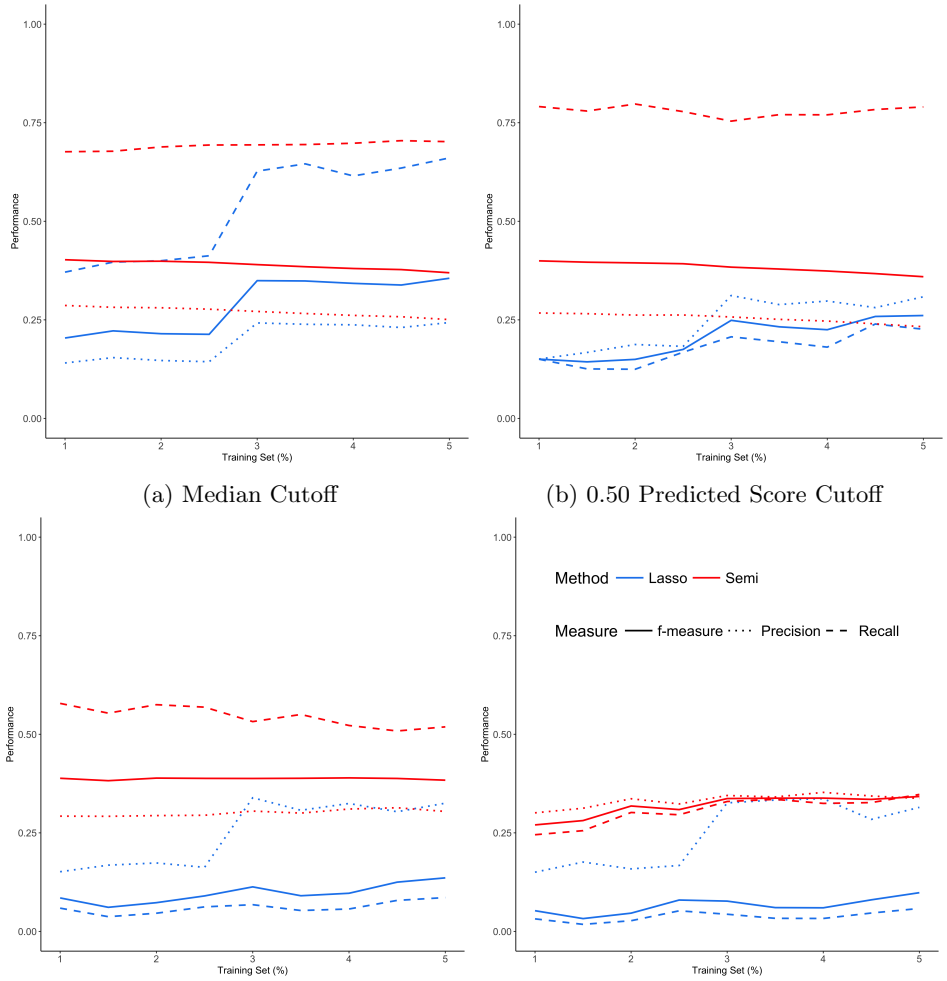
(a) Median Cutoff

(b) 0.50 Predicted Score Cutoff

(c) 0.80 Predicted Score Cutoff

(d) 0.95 Predicted Score Cutoff

Supplementary Figure 2: **Performance of Semi-supervised and LASSO methods with predicted score cutoffs at median, 0.50, 0.80, and 0.95 at 20% contamination.** Semi-supervised method is shown in red while LASSO is shown in blue. Precision, recall, and f-measures are represented by dotted, dashed, and solid lines, respectively. The median is a relative cutoff while the other cutoffs represent absolute cutoffs with re-scaled predicted probabilities.



(a) Median Cutoff

(b) 0.50 Predicted Score Cutoff

(c) 0.80 Predicted Score Cutoff

(d) 0.95 Predicted Score Cutoff

Supplementary Figure 3: **Performance of Semi-supervised and LASSO methods with predicted score cutoffs at median, 0.50, 0.80, and 0.95 at 50% contamination.** Semi-supervised method is shown in red while LASSO is shown in blue. Precision, recall, and f-measures are represented by dotted, dashed, and solid lines, respectively. The median is a relative cutoff while the other cutoffs represent absolute cutoffs with re-scaled predicted probabilities.