Title
Novel machine learning and correlation network methods for genomic data

Author
Song, Lin

Publication Date
2013

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Novel machine learning and correlation network methods for genomic data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Human Genetics

by

Lin Song

2013

Abstract of the Dissertation

Novel machine learning and correlation network methods for genomic data

by

Lin Song

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2013

Professor Stefan Horvath, Chair

Correlation measures are often used to define co-expression networks among genes. As an alternative, mutual information (MI) is often used because it measures non-linear relationships. It is not clear how much MI adds beyond standard (robust) correlation measures or regression model based association measures. Further, it is important to assess which measures lead to biologically meaningful modules (clusters of genes). We provided a comprehensive comparison between mutual information and several correlation measures in 8 empirical data sets and in simulations. We confirmed close relationships between MI and correlation in all data sets, reflecting the fact that most gene pairs satisfy linear or monotonic relationships. The biweight midcorrelation, a robust form of correlation, outperformed MI in terms of elucidating gene pairwise relationships. Coupled with the topological overlap matrix transformation, it often led to modules superior to MI and maximal information coefficient (MIC) in terms of gene ontology enrichment. In addition, we proposed the use of polynomial or spline regression models as an alternative to MI for capturing non-linear relationships between quantitative variables. Overall, our results indicated that MI networks could be safely replaced by correlation networks for stationary co-expression data.

Sample classification, especially disease status prediction, is an important area of investigation for gene expression studies. Many machine learning methods, i.e. predictors, have been developed to tackle this problem. Ensemble predictors such as random forest are known to have superior accuracy but their black-box predictions are difficult to interpret. In contrast, a generalized linear model (GLM) coupled with forward feature selection is very interpretable but tends to overfit the data and leads to low predictive accuracy. We proposed a novel bootstrap aggregated (bagged) GLM predictor randomGLM (RGLM) that shares the advantages of a random forest predictor and those of a GLM predictor. RGLM incorporates several elements of randomness and instability, such as random subspace method, optional interaction terms and forward feature selection. The prediction performances of various predictors were evaluated on hundreds of genomic data sets, the UCI machine learning benchmark data and simulations. RGLM often outperformed alternative methods including random forests and penalized regression models (ridge regression, elastic net, lasso) in both binary and continuous outcome predictions. Further, RGLM provides variable importance measures that can be used to define a "thinned" ensemble predictor (involving few features) retaining excellent predictive accuracy.

RGLM has won the 2012 COPD Improver Challenge, in which we aimed to predict the chronic obstructive pulmonary disease (COPD) status based on gene expression data. We outlined how RGLM compared with random forest on the COPD data set, and discussed potential reasons for the superior performance of RGLM in this sub-challenge.

The dissertation of Lin Song is approved.

Matteo Pellegrini

Stanley F Nelson

Aldons J Lusis

Stefan Horvath, Committee Chair

University of California, Los Angeles

2013

This work is dedicated to my parents and my husband, in thanks to their unconditional love and support.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This research was supported by the following grants: 1R01DA030913-01, P50CA092131, R01NS058980, P30CA16042, and National Center for Advancing Translational Sciences UCLA CTSI Grant UL1TR000124.

Chapter 1 is a version of the following article: L. Song, P. Langfelder, S. Horvath: **Random generalized linear model: a highly accurate and interpretable ensemble predictor**, *BMC Bioinformatics* 2013, **14:**5.

Chapter 2 is a version of the following article: L. Song, P. Langfelder, S. Horvath: **Comparison of co-expression measures: mutual information, correlation, and model based indices**, *BMC Bioinformatics* 2012, **13:**328.

Chapter 3 is based on a co-authored paper submitted to Systems Biomedicine, which acknowledges the contributions of Stefan Horvath.

Vita

2008	B.S., Chemistry and Biology Tsinghua University Beijing, China
2008–2009	ACCESS Program Fellowship University of California, Los Angeles Los Angeles, California
2012	M.S., Biostatistics University of California, Los Angeles Los Angeles, California
2012	Winner of the COPD Improver Challenge Philip Morris International (PMI) and International Business Machines Corporation (IBM) Boston, Massachusetts
2013	University Fellowship University of California, Los Angeles Los Angeles, California
2008–2013	Graduate Student Researcher Department of Human Genetics University of California, Los Angeles Los Angeles, California

Publications and Presentations

L. Song, P. Langfelder, S. Horvath, Comparison of co-expression measures: mutual information, correlation, and model based indices, BMC Bioinformatics (2012), 13:328.

L. Song, P. Langfelder, S. Horvath, Random generalized linear model: a highly accurate and interpretable ensemble predictor, BMC Bioinformatics (2013), 14:5.

L. Song, S. Horvath, Predicting COPD status with a Random Generalized Linear Model, submitted to Systems Biomedicine.

ISMB conference network biology SIG. Long beach (CA). July 2012. Poster: Should mutual information, correlation, or model based co-expression measures be used for defining network modules?

Improver Symposium. Boston (MA). October 2012. Presentation: Random generalized linear model predictor for COPD diagnostics.

CHAPTER 1

Comparison of co-expression measures: mutual information, correlation, and model based indices

Introduction

Co-expression methods are widely used for analyzing gene expression data and other high dimensional "omics" data. Most co-expression measures fall into one of two categories: correlation coefficients or mutual information measures. MI measures have attractive information-theoretic interpretations and can be used to measure non-linear associations. Although MI is well defined for discrete or categorical variables, it is non-trivial to estimate the mutual information between quantitative variables, and corresponding permutation tests can be computationally intensive. In contrast, the correlation coefficient and other model based association measures are ideally suited for relating quantitative variables. Model based association measures have obvious statistical advantages including ease of calculation, straightforward statistical testing procedures, and the ability to include additional covariates into the analysis. Researchers trained in statistics often measure gene co-expression by the correlation coefficient. Computer scientists, trained in information theory, tend to use a mutual information (MI) based measure. Thus far, the majority of published articles use the correlation coefficient as co-expression measure [1–5] but hundreds of articles have used the mutual information (MI) measure [6–12].

Several articles have used simulations and real data to compare the two co-expression measures when clustering gene expression data. Allen et al. have found that correlation based network inference method WGCNA [5] and mutual information based method ARACNE [9] both perform well in constructing global network structure [13]; Steuer et al. show that mutual information and the Pearson correlation have an almost one-to-one correspondence when measuring gene pairwise relationships within their investigated data set, justifying the application of Pearson correlation as a measure of similarity for gene-expression measurements [14]. In simulations, no evidence could be found that mutual information performs bet-

ter than correlation for constructing co-expression networks [15]. However, MI continues to be used in recent publications. Some authors have argued that MI is more robust than Pearson correlation in terms of distinguishing various clustering solutions [10]. Given the debates, it remains an open question whether mutual information could be supplanted by standard model based association measures. We affirmatively answer this question by i) reviewing the close relationship between mutual information and likelihood ratio test statistic in the case of categorical variables, ii) finding a close relationship between mutual information and correlation in simulations and empirical studies, and iii) proposing polynomial and spline regression models as alternatives to mutual information for modeling non-linear relationships.

While previous comparisons involved the Pearson correlation, we provide a more comprehensive comparison that considers i) different types of correlation coefficients, e.g. the biweight midcorrelation (bicor), ii) different approaches for constructing MI based and correlation based networks, iii) different ways of transforming a network adjacency matrix (e.g. the topological overlap reviewed below [4, 16–18]), and iv) 8 diverse gene expression data from yeast, mouse and humans. Our unbiased comparison evaluates co-expression measures at the level of gene pair relationships and at the level of forming co-expression modules (clusters of genes).

This chapter presents the following results. First, probably the most comprehensive empirical comparison to date is used to evaluate which pairwise association measure leads to the biologically most meaningful network modules (clusters) when it comes to functional enrichment with GO ontologies. Second, polynomial regression and spline regression methods are evaluated when it comes to defining non-linear association measures between gene pairs. Third, simulation studies are used to validate a functional relationship (cor-MI function) between correlation

and mutual information in case that the two variables satisfy a linear relationship. Our comprehensive empirical studies illustrate that the cor-MI function can be used to approximate the relationship between mutual information and correlation in case of real data sets which indicates that in many situations the MI measure is not worth the trouble. Gene pairs where the two association measures disagree are investigated to determine whether technical artifacts lead to the incongruence.

A. Association measure and network adjacency

An association measure is used to estimate the relationships between two random variables. For example, correlation is a commonly used association measure. There are different types of correlations. While the Pearson correlation, which measures the extent of a linear relationship, is the most widely used correlation measure, the following two more robust correlation measures are often used. First, the Spearman correlation is based on ranks, and measures the extent of a monotonic relationship between x and y. Second, "bicor" (refer to Materials and Methods for definition and details) is a median based correlation measure, and is more robust than the Pearson correlation but often more powerful than the Spearman correlation [19, 20]. All correlation coefficients take on values between -1 and 1 where negative values indicate an inverse relationship. A correlation coefficient is an attractive association measure since i) it can be easily calculated, ii) it affords several asymptotic statistical tests (regression models, Fisher transformation) for calculating significance levels (p-values), and iii) the sign of correlation allows one to distinguish between positive and negative relationships. Other association measures, such as mutual information, will be introduced in the next sections.

Association measures can be transformed into network adjacencies. For n variables $v_1, \ldots, v_n$, an adjacency matrix $A = (A_{ij})$ is an $n \times n$ matrix quantifying the pairwise connection strength between variables. An (undirected) network

adjacency satisfies the following conditions:

$$\begin{aligned} 0 \leq A_{ij} &\leq 1, \\ A_{ij} &= A_{ji}, \\ A_{ii} &= 1. \end{aligned} \tag{1.1}$$

An association network is defined as a network whose nodes correspond to random variables and whose adjacency matrix is based on the association measure between pairs of variables [21]. Association networks describe the pair wise associations between variables (interpreted as nodes). For a given set of nodes, there is a one-to one relationship between the association network and the adjacency matrix. In order to build an association network for n variables $v = (v_1, \ldots, v_n)$, we start by defining an association measure $AssocMeasure(x, y)$ as a real valued function of two vectors x, y. We then apply this function on the set of $N = n^2$ variable pairs $\{Pair_1 = (v_1, v_1), Pair_2 = (v_1, v_2), \ldots, Pair_N = (v_N, v_N)\}$, resulting in an $n \times n$ dimensional matrix

$$S = (AssocMeasure(v_i, v_j)). \tag{1.2}$$

Then, one needs to specify how the association matrix S is transformed into an adjacency matrix. This involves three steps: 1) symmetrize S; 2) transform (and/or threshold) S to $[0, 1]$; 3) set diagonal values to 1. As for step 1, many methods can be used to symmetrize S if it is non-symmetric, such as the following three ways:

$$S_{ij}^{min} = min(S_{ij}, S_{ji}) \tag{1.3}$$

$$S_{ij}^{ave} = \frac{S_{ij} + S_{ji}}{2} \tag{1.4}$$

$$S_{ij}^{max} = max(S_{ij}, S_{ji}). \tag{1.5}$$

As for step 2, if $LowerBounds(S)$ and $UpperBounds(S)$ denote symmetric matrices of element-wise lower and upper bounds for S, then a simple transformation can be defined as:

$$A = \left(\frac{S - LowersBound(S)}{UpperBounds(S) - LowerBound(S)} \right)^{\beta}, \tag{1.6}$$

where the power β is constant and denotes a soft threshold. As an example, assume that the association measure is given by a correlation coefficient, i.e. $S = (cor(\boldsymbol{x}_i, \boldsymbol{x}_j))$. Since each correlation has the lower bound -1 and upper bound $+1$, Eq. 1.6 reduces to the case of a signed weighted correlation network given by [4, 22]:

$$A_{ij} = \left(\frac{1 + cor(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2} \right)^{\beta}. \tag{1.7}$$

Additional details of correlation based adjacencies (unweighted or weighted, unsigned or signed) are described in Materials and Methods.

Network adjacency based on co-expression measures

When dealing with gene expression data, x_i denotes the expression levels of the i-th gene (or probe) across multiple samples. In this article, we assume that the m components of x_i correspond to random independent samples. Co-expression measures can be used to define co-expression networks in which the nodes correspond to genes. The adjacencies A_{ij} encode the similarity between the expression profiles of genes i and j. In practice, transformations such as the topological overlap measure (TOM) [4, 16–18] are often used to turn an original network adjacency matrix into a new one. Details of TOM transformation are reviewed in Materials and Methods.

B. Mutual information networks based on categorical variables

Assume two random samples dx and dy of length m from corresponding discrete or categorical random variables DX and DY. Each entry of dx equals one of the following R levels $ldx_1, \ldots, ldx_R$. The mutual information (MI) is defined as:

$$MI(dx, dy) = \sum_{r=1}^{R_x} \sum_{c=1}^{R_y} p(ldx_r, ldy_c) \log \left(\frac{p(ldx_r, ldy_c)}{p(ldx_r)p(ldy_c)} \right) \tag{1.8}$$

where $p(ldx_r)$ is the frequency of level r of dx, and log is the natural logarithm. Note that the following simple relationship exists between the mutual information (Eq. 1.8) and the likelihood ratio test statistic:

$$MI(dx, dy) = \frac{LRT.statistic(dx, dy)}{2m} \tag{1.9}$$

This relationship has many applications. First, it can be used to prove that the mutual information takes on non-negative values. Second, it can be used to calculate an asymptotic p-value for the mutual information. Third, it points to a way for defining a mutual information measure that adjusts for additional conditioning variables $z_1, z_2, \ldots$ Specifically, one can use a multivariate *multinomial regression model* for regressing dy on dx and the conditioning variables. Up to a scaling factor of $2m$, the likelihood ratio test statistic can be interpreted as a (non-symmetric) measure of mutual information between dx and dy that adjusts for conditioning variables. More detailed discussion of mutual information can be found in [14, 23, 24].

As discussed below, numerous ways have been suggested for construct an adjacency matrix based on MI. Here we describe an approach that results in a weighted adjacency matrix. Consider n categorical variables $dx_1, dx_2, \ldots, dx_n$. Their mutual information matrix $MI(dx_i, dx_j)$ is a similarity matrix S whose entries are

bounded from below by 0. To arrive at an upper bound, we review the relationship between mutual information and entropy (the following equation is text book knowledge):

$$MI(dx, dy) = Entropy(dx) + Entropy(dy) - Entropy(dx, dy) \tag{1.10}$$

where $Entropy(dx)$ denotes the entropy of dx and $Entropy(dx, dy)$ denotes the joint entropy (refer to Materials and Methods). Using Eq. 1.10, one can prove that the mutual information has the following 3 upper bounds:

$$\begin{aligned} MI(dx, dy) &\leq min(Entropy(dx), Entropy(dy)), & (1.11)\\ MI(dx, dy) &\leq \frac{Entropy(dx) + Entropy(dy)}{2}, & (1.12)\\ MI(dx, dy) &\leq max(Entropy(dx), Entropy(dy)). & (1.13) \end{aligned}$$

Using Eq. 1.6 with $\beta = 1$, lower bounds of 0 and $UpperBounds_{ij} = (Entropy(dx_i) + Entropy(dx_j))/2$ (Eq. 1.12) results in the *symmetric uncertainty based mutual information adjacency matrix*:

$$A_{ij}^{MI,SymmetricUncertainty} = \frac{2MI(dx_i, dx_j)}{Entropy(dx_i) + Entropy(dx_j)}. \tag{1.14}$$

A transformation of $A^{MI,SymmetricUncertainty}$ leads to the *universal mutual information based adjacency matrix version 1* (denoted AUV1):

$$A_{ij}^{MI,UniversalVersion1} = \frac{A_{ij}^{MI,SymmetricUncertainty}}{2 - A_{ij}^{MI,SymmetricUncertainty}} \tag{1.15}$$

One can easily prove that $0 \leq A_{ij}^{MI,UniversalVersion1} \leq 1$. The term "universal" reflects the fact that the adjacency based dissimilarity $dissMI_{ij}^{UniveralVersion1} = 1 - A^{MI,UniversalVersion1}$ turns out to be a universal distance function [25]. Roughly

speaking, the universality of $dissMI_{ij}^{UniveralVersion1}$ implies that any other distance measure between dx_i and dx_j will be small if $dissMI_{ij}^{UniveralVersion1}$ is small. The term "distance" reflects the fact that $dissMI^{UniveralVersion1}$ satisfies the properties of a distance including the triangle inequality.

Another adjacency matrix is based on the upper bound implied by inequality 1.13. We define the *universal mutual information based adjacency matrix version 2*, or AUV2, as follows:

$$A^{MI,UniversalVersion2} = \frac{MI(dx_i, dx_j)}{max(Entropy(dx_i), Entropy(dx_j))}. \tag{1.16}$$

The name reflects the fact that $dissMI^{UniveralVersion2} = 1 - A^{MI,UniversalVersion2}$ is also a universal distance measure [25]. While $A^{MI,UniversalVersion1}$ and $A^{MI,UniversalVersion2}$ are in general different, we find very high Spearman correlations ($r > 0.9$) between their vectorized versions.

Many alternative approaches exist for defining MI based networks, e.g. ARACNE [9], CLR [26], MRNET [27] and RELNET [6, 28] are described in Materials and Methods.

C. Mutual information networks based on discretized numeric variables

In its original inception, the mutual information measure was only defined for discrete or categorical variables, see e.g. [23]. It is challenging to extend the definition to *quantitative* variables. But, several strategies have been proposed in the literature [7,28,29]. In this article, we will only consider the following approach which is based on discretizing the numeric vector x by using the equal width discretization method. This method partitions the interval $[min(x), max(x)]$ into equal-width bins (sub-intervals). The vector $discretize(x)$ has the same length as

x but its l-th component reports the bin number in which x_l falls:

$$dx_l = discretize(x)_l = r \text{ if } x_l \in bin_r. \tag{1.17}$$

The number of bins, $no.bins$, is the only parameter of the equal-width discretization method.

In our subsequent studies, we calculate an MI-based adjacency matrix using the following three steps. First, numeric vectors of gene expression profiles are discretized according to the equal-width discretization method with the default number of bins given by $no.bins = \sqrt{m}$. Second, the mutual information $MI_{ij} = MI(discretize(x_i), discretize(x_j))$ is calculated between the discretized vectors based on Eq. 1.10 and the Miller Madow entropy estimation method (detailed in Materials and Methods). Third, the MI matrix is transformed into one of three possible MI-based adjacency matrices: $A^{MI,SymmetricUncertainty}$ (Eq. 1.14), $A^{MI,UniversalVersion1}$ (Eq. 1.15), $A^{MI,UniversalVersion2}$ (Eq. 1.16).

Results

A. An equation relating $MI(discretize(x), discretize(y))$ to $cor(x, y)$

As described previously, the mutual information $MI(discretize(x), discretize(y))$ between the discretized vectors can be used as an association measure. Note that $MI(discretize(x), discretize(y))$ is quite different from $cor(x, y)$ in the following aspects. First, the estimated mutual information depends on parameter choices, e.g. the number of bins used in the equal-width discretization step for defining $dx = discretize(x)$. Second, the mutual information aims to measure general dependence-relationships while the correlation only measures linear or monotonic relationships. Third, the equations for the two measures are very different. Given these differences, it is surprising that a simple approximate relationship holds between the two association measures if x, y are samples from a bivariate normal distribution and the equal-width discretization method is used with $no.bins = \sqrt{m}$. Under these assumptions, we will show that $A^{MI,UniversalVersion2}$ can be accurately approximated as follows:

$$\begin{aligned} A^{MI,UniversalVersion2}(dx, dy) &= \frac{MI(dx, dy)}{max(Entropy(dx), Entropy(dy))} \\ &\approx F^{cor-MI}(cor(x, y)), \end{aligned} \tag{1.18}$$

where the "cor-MI" function [21]

$$F^{cor-MI}(s) = \frac{\log(1 + \epsilon - s^2)}{\log(\epsilon)}(1 - \omega) + \omega \tag{1.19}$$

depends on the following two parameters

$$\begin{aligned} \omega &= 0.43m^{-0.30} \\ \epsilon &= \omega^{2.2}. \end{aligned} \tag{1.20}$$

In general, one can easily show that $F^{cor-MI}(s)$ is a monotonically increasing function that maps the unit interval [0,1] to [0,1] if the two parameters ω and ϵ satisfy the following relationship

$$0 < \epsilon \leq \omega < 1\,. \tag{1.21}$$

Eq. 1.19 was stated in terms of the Pearson correlation, but it also applies for bicor as can be seen from our simulation studies.

B. Simulations where x and y represent samples from a bivariate normal distribution

Here we use simulation studies to illustrate that F^{cor-MI} (Eq. 1.19) can be used for predicting or approximating $A^{MI,UniversalVersion2}$ from the corresponding correlation coefficients (Eq. 1.19). Specifically, we simulate 2000 pairs of sample vectors x and y from a bivariate normal distribution. Each pair of vectors x and y is simulated to exhibit different pairwise correlations. Figure 1.1 shows the relationships of the MI-based adjacency measures with the (observed) Pearson correlation (cor) or biweight midcorrelation (bicor) when each of the vectors contains $m = 1000$ components but the relationship has been confirmed for m ranging from 20 to 10000. As can be seen from Figures 1.1 (A-B), the cor-MI function (Eq. 1.19) with parameters specified in Eq. 1.20 provides a highly accurate prediction of $A^{MI,UniversalVersion2}$ (Eq. 1.16) on the basis of $cor(x, y)$ and m. Since x and y are normally distributed, the Pearson correlation and bicor are practically indistinguishable (Figure 1.1 (C)). Thus, replacing cor by bicor leads to equally good predictions of $A^{MI,UniversalVersion2}$ (Figure 1.1 (D)). Figure 1.1 (E) shows that $A^{MI,UniversalVersion2}$ is practically indistinguishable from $A^{MI,SymmetricUncertainty}$. This suggests that cor-MI function can also be used to

predict $A^{MI,SymmetricUncertainty}$ on the basis of the correlation measure. Figure 1.1 (F) indicates that $A^{MI,UniversalVersion1}$ and $A^{MI,UniversalVersion2}$ are different from each other but satisfy a monotonically increasing relationship.

C. Empirical studies involving 8 gene expression data sets

Our simulation results show that both the robust biweight midcorrelation and the Pearson correlation can be used as input of F^{cor-MI} for predicting $A^{MI,UniversalVersion2}$ when the underlying variables satisfy pairwise bivariate normal relationships. However, it is not clear whether F^{cor-MI} can also be used to relate correlation and mutual information in real data applications. In this section, we report 8 empirical studies to study the relationship between MI and the robust correlation measure bicor. To focus the analysis on genes that are likely to reflect biological variation and to reduce computational burden, we selected the 3000 genes with highest variance across the microarray samples for each data set. Description of data sets can be found in Materials and Methods.

We first calculate bicor and $A^{MI,UniversalVersion2}$ for all gene pairs in each data set. The two co-expression measures show strong monotonic relationships in most data sets (Figure 1.2). Then, we predict $A^{MI,UniversalVersion2}$ from bicor based on F^{cor-MI} (Eq. 1.19). Our predictions are closely related to true $A^{MI,UniversalVersion2}$ values (Figure 1.3). These results indicate that most gene pairs satisfy linear relationships in real data applications. Among the 8 data sets, SAFHS shows the strongest association between bicor and $A^{MI,UniversalVersion2}$ (Spearman correlation 0.72) and also gives the most accurate $A^{MI,UniversalVersion2}$ prediction (Pearson correlation 0.92). A possible reason is that the large samples size ($m = 1084$) leads to more accurate estimation of mutual information, thus enhancing the association with bicor and the performance of the prediction function. In contrast, the small sample size ($m = 44$) of the yeast data set adversely affects the calcu-

lation of mutual information and hence the prediction performance of F^{cor-MI}. In summary, our examples indicate that for most gene pairs, $A^{MI,UniversalVersion2}$ (Eq. 1.16) is a monotonic function (cor-MI) of the absolute value of bicor. This finding likely reflects the fact that the vast majority of gene pairs satisfy straight line relationships. This approximation improves with increasing sample size m, possibly reflecting more accurate estimation of mutual information.

Although F^{cor-MI} reveals a close relationship between bicor and $A^{MI,UniversalVersion2}$ for most gene pairs, there are cases where the two association measures strongly disagree. In the following, we present scatter plots to visualize the relationships between pairs of genes where MI found a significant relationship while bicor did not and vice versa. To facilitate a comparison between bicor and MI, we standardized each association measure across pairs, which resulted in the Z scores denoted by $Z.MI_{ij} = (MI_{ij} - mean(MI))/\sqrt{(var(MI))}$ and $Z.bicor_{ij} = (bicor_{ij} - mean(bicor))/\sqrt{(var(bicor))}$. Next we selected gene pairs whose value of $Z.MI_{ij}$ was large but $Z.bicor_{ij}$ was low and vice versa. The resulting pairs correspond to the blue and red circles in Figures 1.2 and 1.3. To see what dependence patterns drives the discordant behavior of MI and bicor, we used scatter plots to visualize the relationship between the pairs of variables (Figure 1.4). Gene pairs in Figure 1.4 (A) have extreme $A^{MI,UniversalVersion2}$ but insignificant bicor values. Note that the resulting dependencies seem haphazard and may not reflect real biological dependencies. For example, the gene pair in the brain cancer data set exhibits no clear relationships as correctly implied by bicor, while the significant MI value is driven by an array outlier with extremely high expression for both genes. In the SAFHS data, the gene pair exhibits an unusual pattern that is more likely to be the result of batch effects rather than biological signals. The mouse liver data set displays a pairwise pattern that is neither commonly seen nor easily explained. The ND data set shows no obvious patterns at

all, making mutual information less trustworthy. On the contrary, gene pairs with significant value of Z.bicor but insignificant $Z.MI$ values show approximate linear relationships in all data sets (Figure 1.4 (B)). Thus, bicor captures gene pairwise relationships more accurately and sensitively than the mutual information based adjacency $A^{MI,UniversalVersion2}$.

In summary, bicor usually detects linear relationships between gene pairs accurately while mutual information is susceptible to outliers, and sometimes identifies pairs that exhibit patterns unlikely to be of biological origin or that exhibit no clear dependency at all. We note that MI results tend to be more meaningful when dealing with a large number of observations (say $m > 300$). Although we only consider 3000 genes with highest variances, our results are highly robust with respect to the number of genes (data not shown).

D. Gene ontology enrichment analysis of co-expression modules defined by different networks

Gene co-expression networks typically exhibit modular structure in the sense that genes can be grouped into modules (clusters) comprised of highly interconnected genes (i.e., within-module adjacencies are high). The network modules often have a biological interpretation in the sense that the modules are highly enriched in genes with a common functional annotation (gene ontology categories, cell type markers, etc) [3, 30, 31]. In this section, we assess association measures (and network construction methods) by the gene ontology (GO) enrichment of their resulting modules in the 8 empirical data sets.

In order to provide an unbiased comparison, we use the same clustering algorithm for module assignment for all networks. Toward this end, we use a module detection approach that has been used in hundreds of publications: modules are defined as branches of the hierarchical tree that results from using $1-Adjacency$ as

dissimilarity measure, average linkage, and the dynamic tree cutting method [32]. An example of the module detection approach is illustrated in Figure 1.5. To provide an unbiased evaluation of GO enrichment of each module, we used the *GOenrichmentAnalysis* R function to test enrichment with respect to all GO terms [33, 34] and retained the 5 most significant p-values for each module.

The 10 different adjacencies considered here are described in the last 2 columns of Table 1.1. We first compare modules based on $A^{MI,UniversalVersion2}$ with those resulting from 3 bicor based networks: unsigned adjacency (unsignedA, Eq. 1.32) , signed adjacency (signedA, Eq. 1.31) and Topological Overlap Matrix (TOM, Eq. 1.33) based on signed adjacency. GO enrichment p-values of modules in the 8 real data applications are summarized as barplots in Figure 1.6. Figure 1.6 indicates that, in terms of gene ontology enrichment, TOM is the best bicor based gene co-expression network construction method, and it is superior to $A^{MI,UniversalVersion2}$. Note that signed correlation network coupled with the topological overlap transformation exhibit the most significant GO enrichment p-values in all data sets, and the difference is statistically significant ($p < 0.05$) in 6 out of 8 comparisons. The effect of module size is discussed below. An obvious question is whether the performance of MI can be improved when using an alternative MI based network inference method. To address this, we compared the performance of the signed correlation network (with TOM) versus 4 commonly used mutual information: ARACNE, CLR, MRNET and RELNET (described in Materials and Methods). ARACNE allows one to choose a tolerance threshold ϵ ranging from 0 to 1. As ϵ increases, more edges of the ARACNE network will be preserved. We evaluated ARACNE ($\epsilon = 0$), ARACNE ($\epsilon = 0.2$) and ARACNE ($\epsilon = 0.5$) into our comparison. Similar to Figure 1.6, Figure 1.7 summarizes the GO enrichment p-values of modules in the 8 real data applications. TOM leads to the highest enrichment p-values in 5 cases, and the difference is statistically sig-

nificant in 4 of them. In two applications, ARACNE ($\epsilon = 0$) performs best, and MRNET performs best in one application. We need to point out that another mutual information based method, maximal information coefficient (MIC) [35], has been proposed recently. Although computational intensive, the MIC has clear theoretical advantages when it comes to capturing general dependence patterns. Figure 1.8 compares the performance of MIC with that of TOM when it comes to GO ontology enrichment. TOM clearly outperforms MIC to identify GO enriched modules in 6 out of 7 data sets which may suggest that MIC tends to overfit the data in these applications. SAFHS data set is not included because the computation of MIC was time-consuming on this large data set.

Overall, these unbiased comparisons show that signed correlation networks coupled with the topological overlap transformation outperform the commonly used mutual information based algorithms when it comes to GO enrichment of modules.

E. Polynomial and spline regression models as alternatives to mutual information

A widely noted advantage of mutual information is that it can detect general, possibly non-linear, dependence relationships. However, estimation of mutual information poses multiple challenges ranging from computational complexity to dependency on parameters and difficulties with small sample sizes. Standard polynomial and spline regression models can also detect non-linear relationships between variables. While perhaps less general than MI, relatively simple polynomial and spline regression models avoid many of the challenges of estimating MI while adequately modeling a broad range of non-linear relationships. In addition to being computationally simpler and faster, regression models also make available standard statistical tests and model fitting indices. Thus, in this section

we examine polynomial and spline regression as alternatives to MI for capturing non-linear relationships between gene expression profiles. We define association measures based on polynomial and spline regression models and study their performance.

Networks based on polynomial and spline regression models

Consider two random variables x and y and the following polynomial regression model of degree 3:

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{1.22}$$

The model fitting index $R^2(x, y)$ (described in Materials and Methods) can be used to evaluate the fit of the model. One can then reverse the roles of x and y to arrive at a model fitting index $R^2(y, x)$. In general, $R^2(x, y) \neq R^2(y, x)$.

Now consider a set of n variables $x_1, \ldots, x_n$. One can then calculate pairwise model fitting indices $R^2_{ij} = R^2(x_i, x_j)$ which can be interpreted as the elements of an $n \times n$ association matrix (R^2_{ij}). This matrix is in general non-symmetric and takes on values in $[0, 1]$, with diagonal values equal to 1. A large value indicates a close relationship between variables x_i and x_j. To define an adjacency matrix, we symmetrize (R^2_{ij}) through Eq. 1.3, 1.4 or 1.5.

Spline regression models are also known as local polynomial regression models [36]. Local refers to the fact that these models amount to fitting models on subintervals of the range of x. The boundaries of subintervals are referred to as knots. In analogy to polynomial models, we build natural cubic spline model for all pairs of x_i, x_j. We use the following rule of thumb for the number of knots: if $m > 100$ use 5 knots, if $m < 30$ use 3 knots, otherwise use 4 knots. We then calculate model fitting indices and create corresponding network adjacencies. (Details of spline model construction can be found in Materials and Methods.)

Compared to spline regression, polynomial regression models have a potential shortcoming: the model fit can be adversely affected by outlying observations. A single outlying observation (x_u, y_u) can "bend" the fitting curve into the wrong direction, i.e. adversely affect the estimates of the β coefficients. Spline regression alleviates this problem by fitting model on sub-intervals of the range of x.

Figure 1.9 (A-B) illustrates the use of regression models for measuring non-linear relationships. In simulation, polynomial and cubic spline regression can correctly capture non-linear trends.

Relationship between regression and MI based networks

Previously, we discussed the relationship between correlation and mutual information based adjacencies in simulations where x and y represent samples from a bivariate normal distribution. Here, we consider the performance of polynomial and spline association measure in the same scenario (Figure 1.10). With all x, y pairs following linear relationships, both regression models reduce to simple linear models, and perform almost identically to correlation based measures (panel A and C). We find that the cor-MI function introduced previously also allows us to relate spline and polynomial regression based networks to the MI based network (panel B and D), e.g. $AUV2_{ij} \approx F^{cor-MI}(\sqrt{max(R^2(x_i, x_j), R^2(x_j, x_i))})$. Note that different symmetrization methods (Eq. 1.3) applied R^2 result in similar adjacencies in our applications (data not shown), thus it's valid to use any of them.

In addition, our empirical data show that regression models and mutual information adjacency $A^{MI,UniversalVersion2}$ are highly correlated, and the relationship is stronger than that between bicor and $A^{MI,UniversalVersion2}$ (Figure 1.9 (C-F)). This indicates that $A^{MI,UniversalVersion2}$ and regression models discover some common gene pairwise non-linear relations that can not be identified by correlations.

Simulations for module identification in data with non-linear relationships

Our empirical studies show that most gene pairs satisfy linear relationships, which implies that correlation based network methods perform well in practice. But one can of course simulate data where non-linear association measures (such as MI, spline R^2) outperform correlation measures when it comes to module detection. To illustrate this point, we simulated data with non-linear gene-gene relationships. Here we simulated 200 genes in 3 network modules across 200 samples. Two of the simulated modules, labeled for convenience by the colors turquoise and blue, contain linear and non-linear (quadratic) gene-gene relationships (Figure 1.5). We then use several different network inference methods to construct networks and define modules. To evaluate how well each network inference method recovers the simulated modules, we use the Rand index between the inferred and simulated module assignment. In this case, non-linear association measures, i.e. AUV2, polynomial and spline regression, identify modules more accurately than correlation based measures (Figure 1.5). In networks based on correlations, the simulated turquoise and blue modules are clearly divided into two separate ones, indicating that they miss the non-linear relationships within these two modules. In contrast, regression models capture non-linear gene pairwise relations and correctly assign these genes into the same modules. To study the effect of the number of observations, we repeated the analysis for m ranging from 10 to 500. Figure 1.11 shows that non-linear association measures, especially regression models, outperform correlation based measures as data sample size increases. Note that polynomial and spline regression based co-expression measures perform as well as MI based networks in this situation. Overall, our results validate the usage of polynomial and spline regression models as alternatives to mutual information for detecting non-linear relationships.

F. Overview of network methods and alternatives

A thorough review of network methods is beyond our scope and we point the reader to the many many review articles [37–40]. But Table 1.1 describes not only the methods used in this article but also alternative approaches. Table 1.1 also describes the kind of biological insights that can be gained from these network methods. As a rule, association networks (based on correlation or MI) are ill suited for causal analysis. While association networks such as WGCNA or ARACNE have been been successfully used for gene regulatory networks (GRNs) [13], a host of alternatives are available. For example, the DREAM (Dialogue for Reverse Engineering Assessments and Methods) project has repeatedly tackled this problem [41–43]. A limitation of our study is that we are focusing on undirected (as opposed to directed, causal models). Structural equation models, Bayesian networks, and other probabilistic graphical models are widely used for studying causal relationships. Many authors have proposed to use Bayesian networks for analyzing gene expression data [44–47] and for generating causal networks from observational data [48] or genetic data [49, 50].

While it is beyond our scope to evaluate network inference methods for time series data (reviewed in [51]), we briefly mention several approaches. A (probabilistic) Boolean network [52] is a special case of a discrete state space model that characterizes a system using dichotomized data. A Bayesian network is a graph-based model of joint multivariate probability distributions that captures properties of conditional independence between variables [45]. Such models are attractive for their ability to describe complex stochastic processes and for modeling causal relationships. Several articles describe the relationship between Boolean networks and dynamic Bayesian networks when it comes to models of gene regulatory relationships [47, 53]. Finally, we mention that correlation network methodology can be adapted to model time series data, e.g. many authors have proposed to use a

time-lagged correlation measure for inferring gene regulatory networks [54].

A large part of GRN research focuses on the accurate assessment of individual network edges, e.g. [55–58] so many of these methods are not designed as data reduction methods. In contrast, correlation network methods, such as WGCNA, are highly effective at reducing high dimensional genomic data since modules can be represented by their first singular vector (i.e. module eigengene) [21, 59].

Discussion

This article presents the following theoretical and methodological results: i) it reviews the relationship between the MI and a likelihood ratio test statistic in case of two categorical variables, ii) it presents a novel empirical formula for relating correlation to MI when the two variables satisfy a linear relationship, and iii) it describes how to use polynomial and spline regression models for defining pairwise co-expression measures that can detect non-linear relationships.

Mutual information has several appealing information theoretic properties. A widely recognized advantage of mutual information over correlation is that it allows one to detect non-linear relationships. This can be attractive in particular when dealing with time series data [60]. But mutual information is not unique in being able to detect non-linear relationships. Standard regression models such as polynomial and spline models can also capture non-linear relationships. An advantage of these models is that well established likelihood based statistical estimation and testing procedures are available. Regression models allow one to calculate model fitting indices that can be used to define network adjacencies as well as flag possible outlying observations by analyzing residuals.

For categorical variables, mutual information is (asymptotically) equivalent to other widely used statistical association measures such as the likelihood ratio statistic or the Pearson chi-square test. In this case, all of these measures (including MI) are arguably optimal association measures. Interpreting MI as a likelihood ratio test statistic facilitates a straightforward approach for adjusting the association measure for additional covariates.

We and others [14] have found close relationships between mutual information and correlation based co-expression networks. Our comprehensive empirical studies show that mutual information is often highly related to the absolute value of

the correlation coefficient. We observe that when robust correlation and mutual information disagree, the robust correlation findings appear to be more plausible statistically and biologically. We found that network modules defined using robust correlation exhibit on average higher enrichment in GO categories than modules defined using mutual information. Since our empirical studies involved expression data measured on a variety of platforms and normalized in different ways, we expect that our findings are broadly applicable.

The correlation coefficient is an attractive alternative to the MI for the following reasons. First, the correlation can be accurately estimated with relatively few observations and it does not require the estimation of the (joint) frequency distribution. Estimating the joint density needed for calculating MI typically requires larger sample sizes. Second, the correlation does not depend on hidden parameter choices. In contrast, MI estimation methods involve (hidden) parameter choices, e.g. the number of bins when a discretization method is being used. Third, the correlation allows one to quickly calculate p-values and false discovery rates since asymptotic tests are available. In contrast, it is computationally challenging to calculate a permutation test p-value for the mutual information between two discretized vectors. Fourth, the sign of the correlation allows one to distinguish positive from negative relationships. Signed correlation networks have been found useful in biological applications [22] and our results show that the resulting modules tend to be more significantly enriched with GO terms that those of networks that ignore the sign information. Fifth, modules comprised of highly correlated vectors can be effectively summarized by the module eigennode (the first principal component of scaled vectors). Sixth, the correlation allows for a straightforward angular interpretation, which facilitates a geometric interpretation of network methods and concepts [59]. For example, intramodular connectivity can be interpreted as module eigennode based connectivity.

Our empirical studies show that a signed weighted correlation network transformed via the topological overlap matrix transformation often leads to the most significant functional enrichment of modules. The recently developed maximal information coefficient [35] has clear theoretical advantages when it comes to measuring general dependence patterns between variables but our results show that the biweight midcorrelation coupled with the topological overlap measure outperforms the MIC when it comes to the GO ontology enrichment of resulting coexpression modules.

While defining mutual information for categorical variables is relatively straightforward, no consensus seems to exist in the literature on how to define mutual information for continuous variables. A major limitations of our study is that we only studied MI measures based on discretized continuous variables. For example, the cor-MI function for relating correlation to MI only applies when an equal width discretization method is used with $no.bins = \sqrt{m}$.

A second limitation concerns our gene ontology analysis of modules identified in networks based on various association measures in which we found that the correlation based topological overlap measure (TOM) leads to co-expression modules that are more highly enriched with GO terms than those of alternative approaches. A potential problem with our approach is that the enrichment p-values often strongly depend on (increase with) module sizes, and TOM tends to lead to larger modules. To address this concern, in Figure 1.12 we show the enrichment p-values as a function of module size for modules identified by TOM and by AUV2. It turns out that in most studies, the enrichment of modules defined by TOM is better than that of comparably sized modules defined by AUV2.

A third limitation concerns our use of the bicor correlation measure as opposed to alternatives (e.g. Pearson or Spearman correlation). In our study we find that all 3 correlation measures lead to very similar findings (Figure 1.13).

Materials and Methods

A. Empirical gene expression data sets description

Brain cancer data set. This data set was composed of 55 microarray samples of glioblastoma (brain cancer) patients. Gene expression profiling were performed with Affymetrix high-density oligonucleotide microarrays. A detailed description can be found in [61].

SAFHS data set. This data set [62] was derived from blood lymphocytes of randomly ascertained participants enrolled independent of phenotype in the San Antonio Family Heart Study. Gene expression profiles of 1084 samples were measured by Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChips.

ND data set. This blood lymphocyte data set consisted of 346 samples from patients with neurological diseases. Illumina HumanRef-8 v3.0 Expression BeadChip were used to measure their gene expression profiles.

Yeast data set. The yeast microarray data set was composed of 44 samples from the Saccharomyces Genome Database (http://db.yeastgenome.org/cgi-bin/SGD/expression/expressionConnection.pl). Original experiments were designed to study the cell cycle [63]. A detailed description of the data set can be found in [64].

Tissue-specific mouse data sets. This study uses 4 tissue-specific gene expression data from a large F_2 mouse intercross (B×H) previously described in [65, 66]. Specifically, the surveyed tissues include adipose (239 samples), whole brain (221 samples), liver (272 samples) and muscle (252 samples).

B. Definition of entropy among categorical variables

Consider the categorical random variable DX for which the frequency of the r-th level is given by $p(ldx_r)$. Then the entropy of (the frequency distribution of) DX is defined as

$$Entropy(DX) = -\sum_{r=1}^{no.lDX} p_{DX}(ldx_r)\log(p_{DX}(ldx_r))\,, \tag{1.23}$$

where we set $0 * \log(0) = 0$ when the probability equals 0.

Assume now that a second categorical variable DY is available, which takes on the values $ldy_1, \ldots, ldy_{no.lDY}$ with probabilities $p_{DY}(ldy_1)$, ..., $p(ldy_{no.lDY})$, respectively. Denote the joint probability distribution between DX and DY by $p_{DX,DY}(ldx_r, ldy_c)$. The joint entropy of DX and DY is defined as

$$Entropy(DX, DY) = -\sum_r \sum_c p_{DX,DY}(ldx_r, ldy_c)\log(p_{DX,DY}(ldx_r, ldy_c)). \tag{1.24}$$

C. Miller-Madow estimators for entropy

Consider a continuous variable X with length m. We obtain DX, the discretized version of X, by the equal-width discretization method. The equal-width discretization method results in a vector of relative frequencies $p = (p_1, \ldots, p_{no.bins})$ where p_r denotes the frequency of the r-th bin. Using these relative frequencies, the Miller-Madow estimator is given by

$$Entropy^{MM}(DX) = -\sum_{r=1}^{no.bins} p_r \log(p_r) + \frac{no.bins - 1}{2m}. \tag{1.25}$$

D. Definition of Biweight Midcorrelation

Biweight midcorrelation (bicor) is considered to be a good alternative to Pearson correlation since it is more robust to outliers [67]. In order to define the biweight midcorrelation of two numeric vectors $x = (x_1, ..., x_m)$ and $y = (y_1, ..., y_m)$, one first defines u_i, v_i with $i = 1, ..., m$:

$$\begin{aligned} u_i &= \frac{x_i - med(x)}{9mad(x)} \\ v_i &= \frac{y_i - med(y)}{9mad(y)} \end{aligned} \tag{1.26}$$

where $med(x)$ is the median of x, and $mad(x)$ is the median absolute deviation of x. This leads us to the definition of weight w_i for x_i, which is,

$$w_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|) \tag{1.27}$$

where the indicator $I(1 - |u_i|)$ takes on value 1 if $1 - |u_i| > 0$ and 0 otherwise. Therefore, $w_i^{(x)}$ ranges from 0 to 1. It decreases as x_i gets away from $med(x)$, and stays at 0 when x_i differs from $med(x)$ by more than $9mad(x)$. An analogous weight $w_i^{(y)}$ can be defined for y_i. Given the weights, we can define biweight midcorrelation of x and y as:

$$bicor(x, y) = \frac{\sum_{i=1}^{m} (x_i - med(x)) w_i^{(x)} (y_i - med(y)) w_i^{(y)}}{\sqrt{\sum_{j=1}^{m} \left[(x_j - med(x)) w_j^{(x)}\right]^2} \sqrt{\sum_{k=1}^{m} \left[(y_k - med(y)) w_k^{(y)}\right]^2}}. \tag{1.28}$$

A modified version of biweight midcorrelation is implemented as function *bicor* in the WGCNA R package [5, 20]. One major argument of the function is "maxPOutliers", which caps the maximum proportion of outliers with weight $w_i = 0$. Practically, we find that $maxPOutliers = 0.02$ detects outliers efficiently while preserving most data. Therefore, 0.02 is the value we utilize in this study.

E. Types of correlation based gene co-expression networks

Given the expression profile x,the co-expression similarity s_{ij} between genes i and j can be defined as:

$$s_{ij} = |cor(\boldsymbol{x}_i, \boldsymbol{x}_j)|.$$

An unweighted network adjacency A_{ij} between gene expression profiles $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be defined by hard thresholding the co-expression similarity s_{ij} as follows

$$A_{ij} = \begin{cases} 1 & \text{if } s_{ij} \geq \tau \\ 0 & \text{otherwise,} \end{cases} \tag{1.29}$$

where τ is the 'hard' threshold parameter. Hard thresholding of the correlation leads to simple network concepts (e.g., the gene connectivity equals the number of direct neighbors) but it may lead to a loss of information.

To preserve the continuous nature of the co-expression information, we define the weighted network adjacency between 2 genes as a power of the absolute value of the correlation coefficient [4, 61]:

$$A_{ij} = s_{ij}^{\beta}, \tag{1.30}$$

with $\beta \geq 1$. This soft thresholding approach emphasizes strong correlations, punishes weak correlations, and leads to a weighted gene co-expression network.

An important choice in the construction of a correlation network concerns the treatment of strong negative correlations. In signed networks negatively correlated nodes are considered unconnected. In contrast, in unsigned networks nodes with high negative correlations are considered connected (with the same strength as nodes with high positive correlations). As detailed in [4, 22], a signed weighted

adjacency matrix can be defined as follows

$$A_{ij} = (0.5 + 0.5\text{cor}(x_i, x_j))^{\beta} \tag{1.31}$$

and an unsigned adjacency by

$$A_{ij} = |cor(x_i, x_j)|^{\beta}. \tag{1.32}$$

β is default to 6 for unsigned adjacency and 12 for signed adjacency. The choice of signed vs. unsigned networks depends on the application; both signed [22] and unsigned [30, 61, 65] weighted gene networks have been successfully used in gene expression analysis.

F. Adjacency function based on topological overlap

The topological overlap matrix (TOM) based adjacency function A_{TOM} maps an original adjacency matrix $A^{original}$ to the corresponding topological overlap matrix, i.e.

$$A_{TOM}(A^{original})_{ij} = \frac{\sum_{l \neq i,j} A_{il}^{original} A_{l,j}^{original} + A_{ij}^{original}}{min(\sum_{l \neq i} A_{il}^{original}, \sum_{l \neq j} A_{jl}^{original}) - A_{ij}^{original} + 1}. \tag{1.33}$$

The TOM based adjacency function A_{TOM} is particularly useful when the entries of $A^{original}$ are sparse (many zeroes) or susceptible to noise. This replaces the original adjacencies by a measure of interconnected that is based on shared neighbors. The topological overlap measure can serve as a filter that decreases the effect of spurious or weak connections and it can lead to more robust networks [17, 18, 68].

G. Mutual-information based network inference methods

There are 4 commonly used mutual-information based network inference methods: RELNET, CLR, MRNET and ARACNE. In order to identify pairwise interactions between numeric variables x_i, x_j, all methods start by estimating mutual information $MI(x_i, x_j)$.

RELNET

The relevance network (RELNET) approach [6, 28] thresholds the pairwise measures of mutual information by a threshold τ. However, this method suffers from a significant limitation that vectors separated by one or more intermediaries (indirect relationships) may have high mutual information without implying a direct interaction.

CLR

The CLR algorithm [26] is based on the empirical distribution of MI. It first defines a score z_i given the mutual information $MI(x_i, x_j)$ and the sample mean μ_i and standard deviation σ_i of the empirical distribution of mutual information $MI(x_i, x_k), k = 1, \ldots, n$:

$$z_i = \max\left(0, \frac{MI(x_i, x_j) - \mu_i}{\sigma_i}\right). \tag{1.34}$$

z_j can be defined analogously. In terms of z_i, z_j, the score used in CLR algorithm can be expressed as $z_{ij} = \sqrt{z_i^2 + z_j^2}$.

MRNET

MRNET [27] infers a network by repeating the maximum relevance/minimum redundancy (MRMR) feature selection method for all variables. The MRMR method starts by selecting the variable x_i having the highest mutual information with target y. Next, given a set S of selected variables, the criterion updates S by choosing the variable x_k that maximizes $u_j - r_j$ where u_j is a relevance term and r_j is a redundancy term. In particular,

$$u_j = MI(x_k, y) \tag{1.35}$$

$$r_j = \frac{1}{|S|} \sum_{x_i \in S} MI(x_k, x_i) \tag{1.36}$$

The score of each pair x_i and x_j will be the maximum score of the one computed when x_i is the target and the one computed when x_j is the target.

ARACNE

The ARACNE [9] (Algorithm for the Reconstruction of Accurate Cellular Networks) developed by Andrea Califano's group is an extension of RELNET. Given the limitation of RELNET, ARACNE removes the vast majority of indirect candidate interactions using a well-known information theoretic property, the data processing inequality (DPI). The DPI applied to association networks states that if variables x_i and x_j interact only through a third variable x_k, then

$$MI(x_i, x_j) \leq min(MI(x_i, x_k), MI(x_k, x_j)) \tag{1.37}$$

ARACNE starts with a network graph where each pair of nodes with $MI_{ij} > \tau$ is connected by an edge. The weakest edge of each triplet, e.g. the edge between i and j, is interpreted as an indirect interaction and is removed if the difference

between $min(MI(x_i, x_k), MI(x_k, x_j))$ and $MI(x_i, x_j)$ lies above a threshold ϵ, i.e. the edge is removed if

$$MI(x_i, x_j) \leq min(MI(x_i, x_k), MI(x_k, x_j)) - \epsilon. \tag{1.38}$$

The tolerance threshold ϵ could be chosen to reflect the variance of the MI estimator and should decrease with increasing sample size m. Using a non-zero tolerance $\epsilon > 0$ can lead to the persistence of some 3-vector loops.

The outputs from RELNET, CLR, MRNET or ARACNE are association matrices. They can be transformed into corresponding adjacencies based on the algorithm discussed in Introduction.

MIC

Another mutual information based method is the recently proposed the maximal information coefficient (MIC) [35]. The MIC is a type of maximal information-based nonparametric exploration (MINE) statistics [35]. In our empirical evaluations, we calculate the MIC using the *minerva* R package [69].

H. Fitting indices of polynomial regression models

While networks based on the Pearson correlation can only capture linear co-expression patterns there is clear evidence for non-linear co-expression relationships in transcriptional regulatory networks [70]. The following classical regression based approaches can be used for studying non-linear relationships. The polynomial regression model:

$$\begin{aligned} E(y) &= \beta_0 1 + \beta_1 x + \beta_2 x^2 \ldots + \beta_d x^d \\ &= M\beta\,, \end{aligned} \tag{1.39}$$

where

$$M = [1, x, \ldots, x^d]\,. \tag{1.40}$$

One can show that the least squares estimate of the parameter vector $\hat{\beta}$ is

$$\hat{\beta} = (M^\tau M)^- M^\tau\ y,$$

where$^-$ denotes the (pseudo) inverse, and $^\tau$ denotes the transpose of a matrix.

Given $\hat{\beta}$, we can calculate the fitting index R^2 as:

$$R^2 = cor(y, \hat{y})^2 = cor(y, M\hat{\beta})^2 \tag{1.41}$$

In the context of a regression model, R^2 is also known as the proportion of variation of y explained by the model.

I. Spline regression model construction

To investigate the relationship between variable x and y, one can use another textbook method from the arsenal of statisticians: spline regression models. Here knots are used to decide boundaries of the sub-intervals. They are typically pre-specified, e.g. based on quantiles of x. The choice of the knots will affect the model fit. It turns out that the values of the knots (i.e. their placement) is not as important as the number of knots. We use the following rule of thumb for the number of knots: if $m > 100$ use 5 knots, if $m < 30$ use 3 knots, otherwise use 4 knots.

To ensure that fit between y and x satisfies a continuous relationship, we review

the hockey stick function $()_+$ to transform x:

$$(s)_+ = \begin{cases} s & \text{if } s \geq 0 \\ 0 & \text{if } s < 0. \end{cases} \tag{1.42}$$

This function can also be applied to the components of a vector, e.g. $(x)_+$ denotes a vector whose negative components have been set to zero. So $(x - knot1)_+$ is a vector whose u-th component equals $x[u] - knot1$ if $x[u] - knot1 \geq 0$ and 0 otherwise.

We are now ready to describe cubic spline regression model, which fits polynomial of degree 3 to sub-intervals. The general form of a cubic spline with 2 knots is as follows

$$E(y) = \beta_0 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - knot_1)_+^3 + \beta_5 (x - knot_2)_+^3. \tag{1.43}$$

The knot parameters (numbers) $knot_1, knot_2, \ldots$ are chosen before estimating the parameter values. Analogous to polynomial regression, R^2 can be calculated as the association measure between x and y. This method guarantees the smoothness of the regression line and restrict the influence of each observation to its local sub-interval.

J. Availability of software

Project name: Adjacency matrix for non-linear relationships

Project home page: `http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA`

Operating system(s): Platform independent

Programming language: R

Licence: GNU GPL 3

The following functions described in this article have been implemented in the WGCNA R package [5]. Function *adjacency.polyReg* and *adjacency.splineReg* calculate polynomial and spline regression R^2 based adjacencies. Users can specify the R^2 symmetrization method. Function *mutualInfoAdjacency* calculates the mutual information based adjacencies $A^{MI,SymmetricUncertainty}$ (Eq. 1.14), $A^{MI,UniversalVersion1}$ (Eq. 1.15) and $A^{MI,UniversalVersion2}$ (Eq. 1.16). Function *AFcorMI* implements the F^{cor-MI} prediction function 1.19 for relating correlation with mutual information.

List of abbreviations

ARACNE: algorithm for the reconstruction of accurate cellular networks.

Bicor: biweight midcorrelation.

GO: gene ontology.

LRT: likelihoood ratio test.

MI: mutual information.

MIC: maximal information coefficient.

TOM: topological overlap matrix.

WGCNA: weighted correlation network analysis.

CHAPTER 1 FIGURES

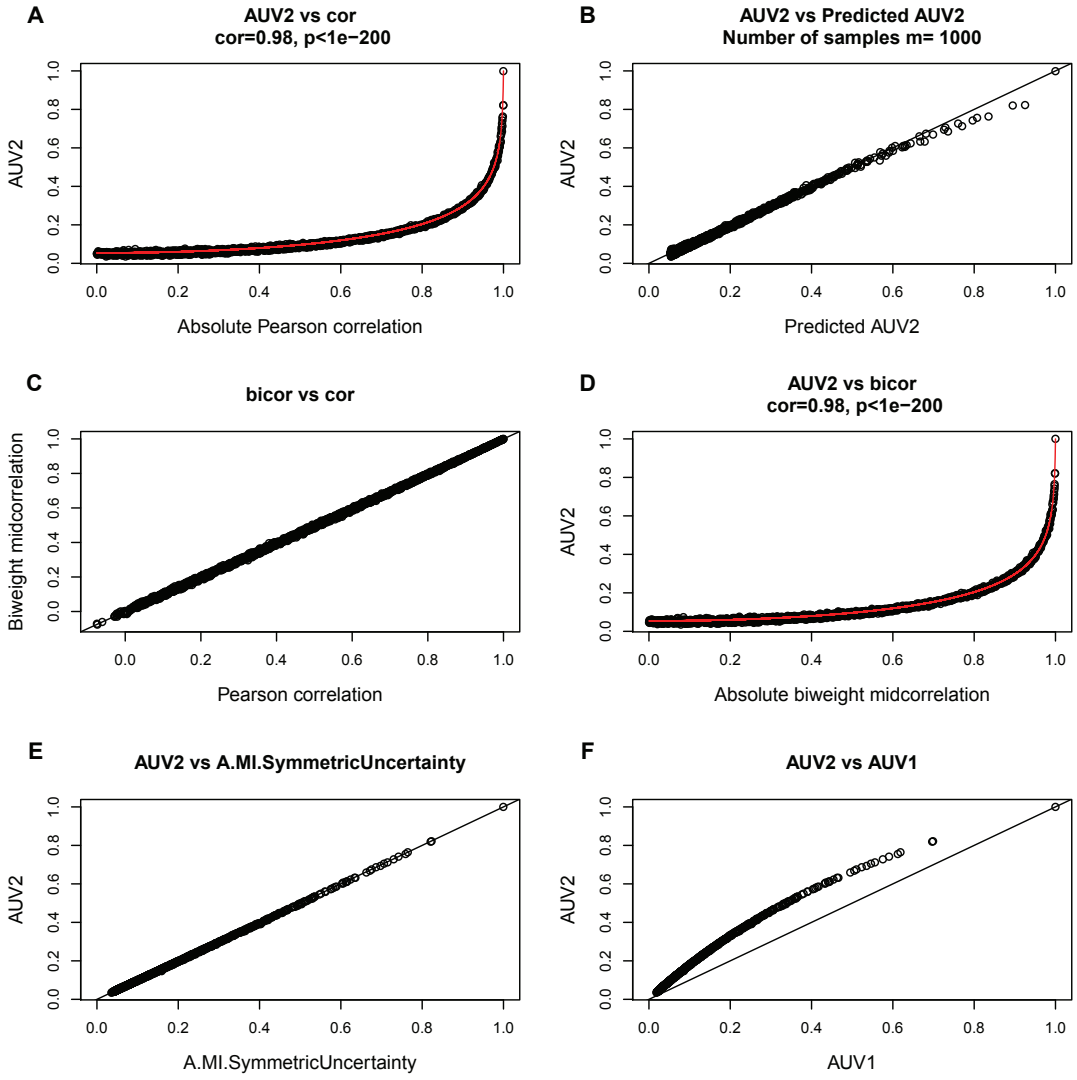

Figure 1.1: **Relating mutual information based adjacencies to the Pearson correlation and biweight midcorrelation in simulation.** **(A)** MI-based adjacency $A^{MI,UniversalVersion2}$ versus absolute Pearson correlation. Spearman correlation of the two measures and the corresponding p-value are shown at the top. The red line shows the predicted $A^{MI,UniversalVersion2}$. **(B)** Observed $A^{MI,UniversalVersion2}$ versus its predicted value. The straight line has slope 1 and intercept 0. **(C)** Observed Pearson correlation versus bicor values. **(D)** $A^{MI,UniversalVersion2}$ versus bicor. Spearman correlation and p-value of the 2 measurements are presented at the top, and predicted $A^{MI,UniversalVersion2}$ are shown as the red line. **(E)** $A^{MI,UniversalVersion2}$ versus $A^{MI.SymmetricUncertainty}$. **(F)** $A^{MI,UniversalVersion2}$ versus $A^{MI,UniversalVersion1}$.

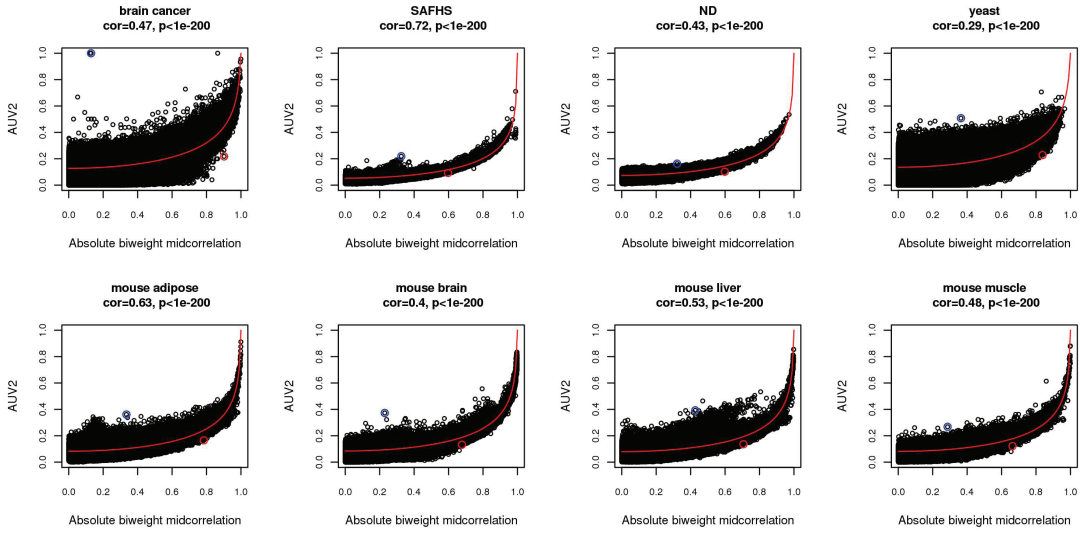

Figure 1.2: **Comparison of correlation and mutual information based co–expression measures in** 8 **empirical data sets.** Absolute value of bicor versus $A^{MI,UniversalVersion2}$ for all probe pairs in each data set. The Spearman correlation and corresponding p-value between the two measures are shown at the top. The red curve predicts $A^{MI,UniversalVersion2}$ from bicor based on Eq. 1.19. The blue circle highlights the probe pair with the highest $A^{MI,UniversalVersion2}$ z-score among those with insignificant bicor z-scores (less than 1.9); the red circle highlights the probe pair with the highest bicor z-score among those with insignificant $A^{MI,UniversalVersion2}$ z-scores (less than 1.9).

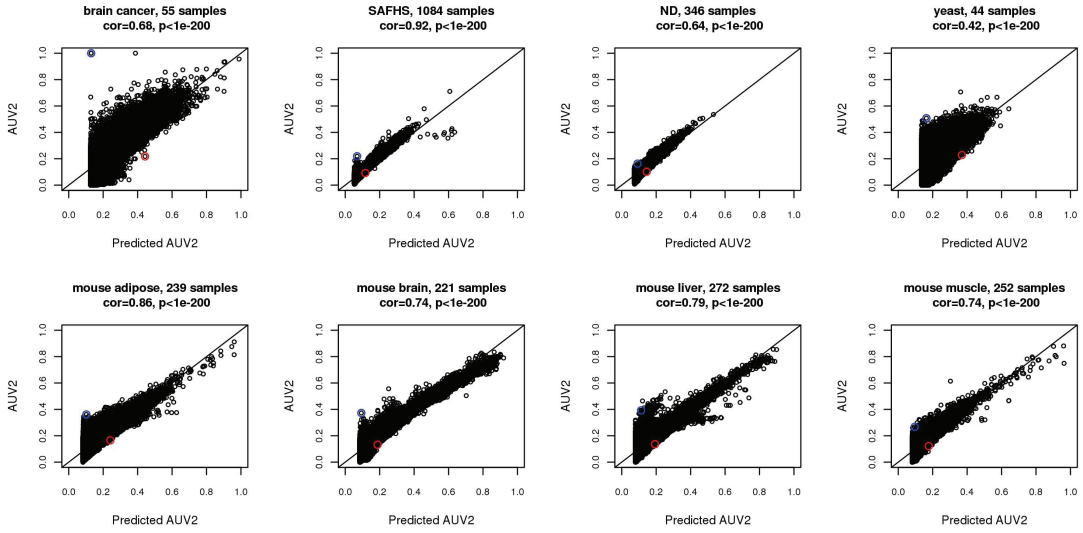

Figure 1.3: **Comparison of predicted and observed $A^{MI,UniversalVersion2}$ in** 8 **empirical data sets.** In all data sets, prediction from bicor based on Eq. 1.19 and observed $A^{MI,UniversalVersion2}$ are highly correlated (the Pearson correlation and corresponding p-value shown at top). Line y=x is added. Blue and red circles have the same meaning as in Figure 1.2.

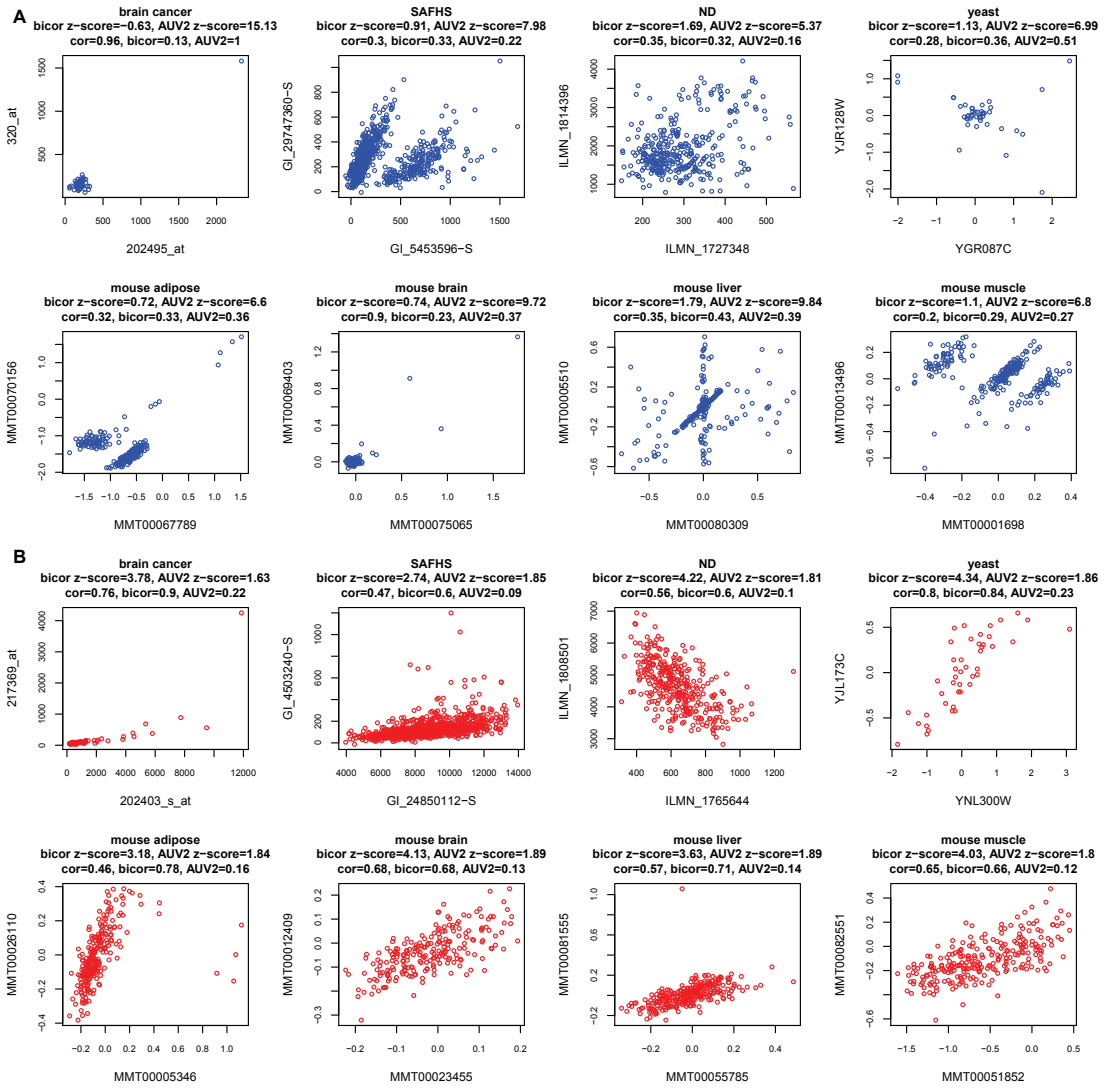

Figure 1.4: **Gene expression of example probe pairs for which the correlation and mutual information based measures disagree. (A)** Gene expression of probe pairs highlighted by blue circles in Figure 1.2. **(B)** Gene expression of probe pairs highlighted by red circles in Figure 1.2. The Pearson correlation, bicor, $A^{MI,UniversalVersion2}$ values and z-scores of the latter two measures are shown at the top. Mutual information is susceptible to outliers, sometimes detects unusual patterns that are hard to explain, and often misses linear relations that are captured by bicor.

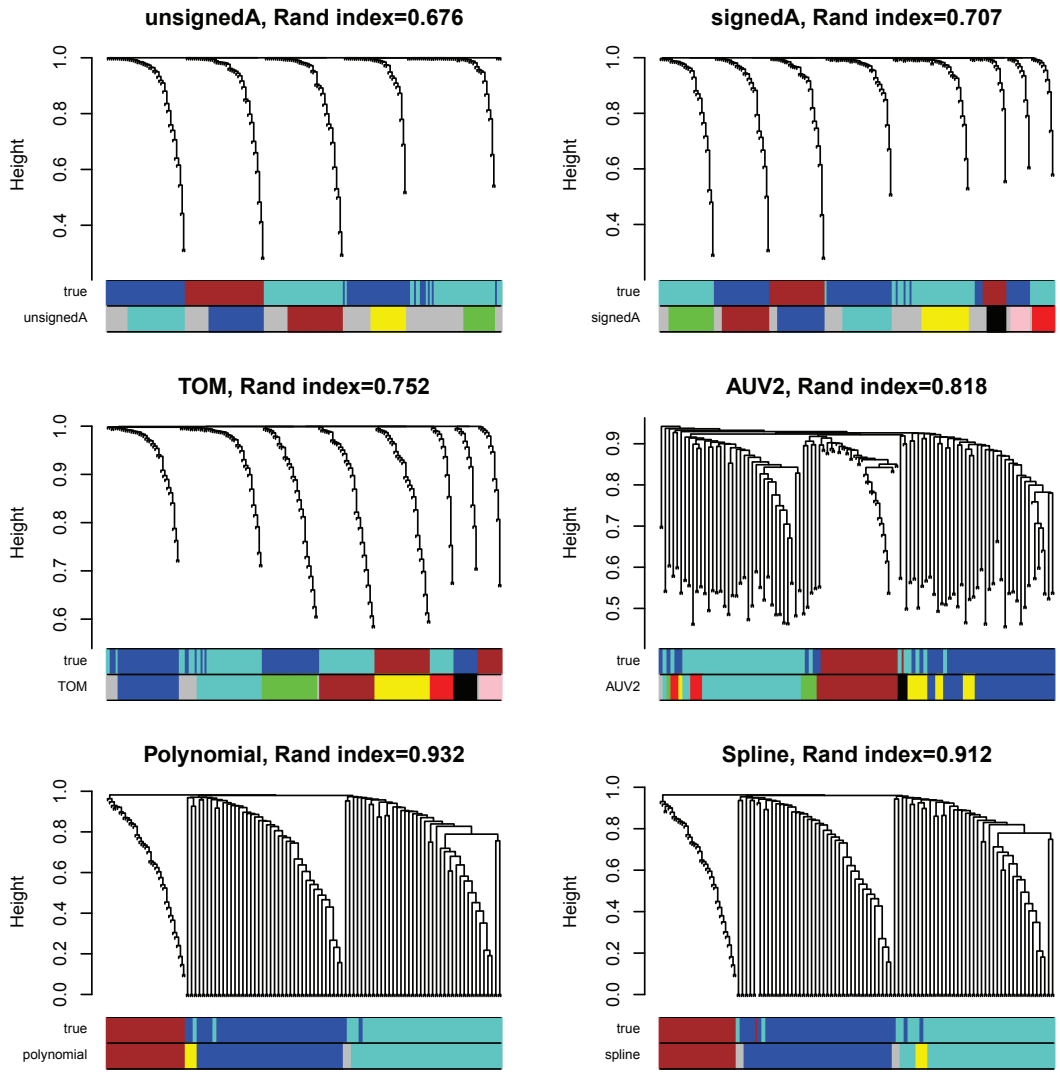

Figure 1.5: **Module identification based on various network inference methods in simulation with non-linear gene-gene relationships.** The data set is composed of 200 genes across 200 samples. 3 true modules are designed. Two of them, labeled with colors turquoise and blue, contain linear and non-linear (quadratic) gene-gene relationships. For each adjacency, the clustering tree and module colors are shown. True simulated module assignment is shown by the first color band underneath each tree. On top of each panel is the Rand index between inferred and simulated module assignments.

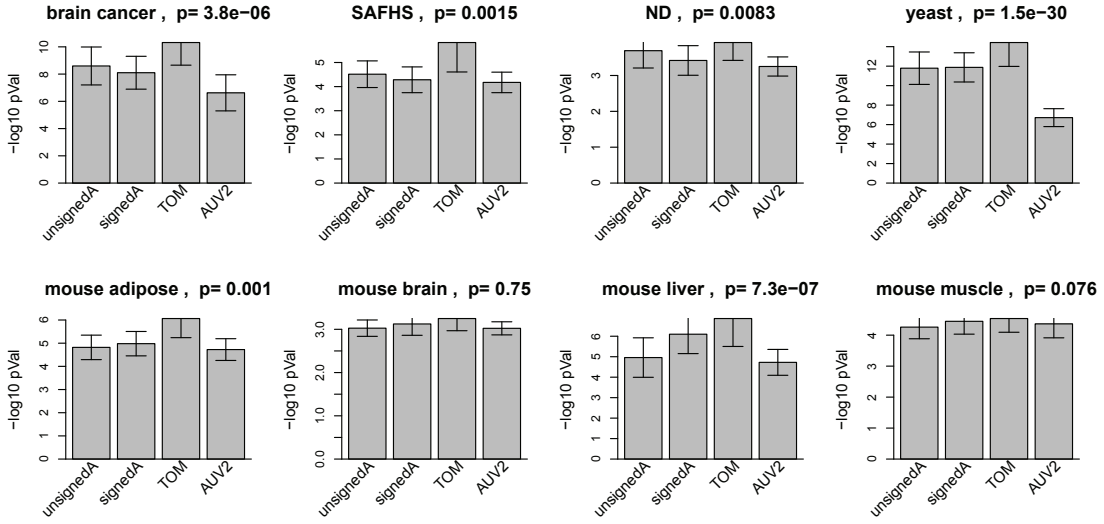

Figure 1.6: **Gene ontology enrichment analysis comparing $A^{MI,UniversalVersion2}$ with bicor based adjacencies in 8 empirical data sets.** 5 best GO enrichment p-values from all modules identified using each adjacency are log transformed, pooled together and shown as barplots. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. TOM outperforms the others in all 8 data sets.

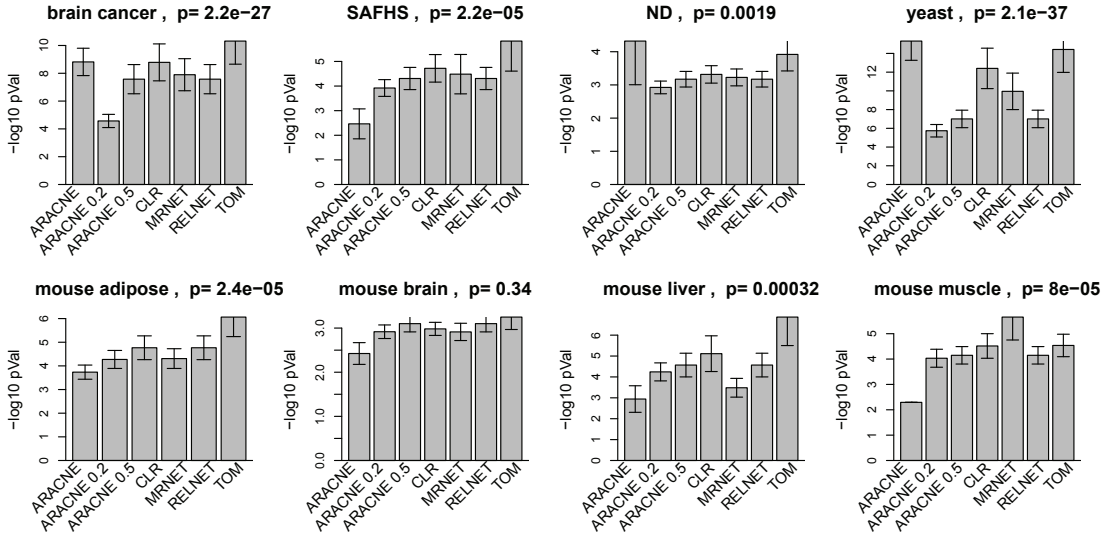

Figure 1.7: **Gene ontology enrichment analysis comparing TOM with MI based adjacencies in** 8 **empirical data sets.** 5 best GO enrichment p-values from all modules identified using each adjacency are log transformed, pooled together and shown as barplots. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. TOM outperforms the others in 5 data sets. ARACNE($\epsilon = 0$) wins in two data sets, making it the second best.

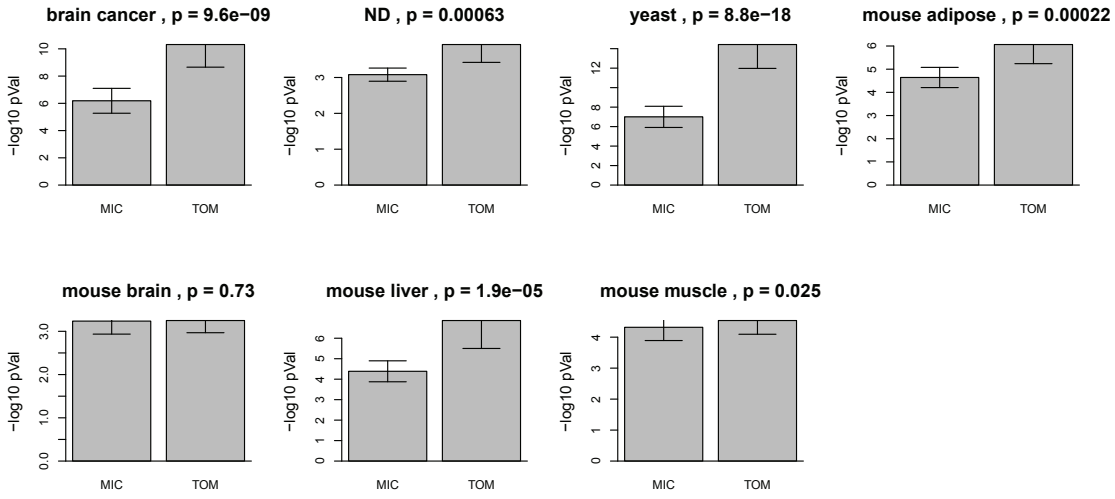

Figure 1.8: **Comparison of MIC and correlation based co-expression measures.** Comparison of MIC and correlation in our empirical gene expression data sets except SAFHS. 5 best GO enrichment p-values from all modules identified using MIC and TOM are log transformed, pooled together and shown as barplots. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. TOM outperforms MIC in all data sets except the mouse brain data.

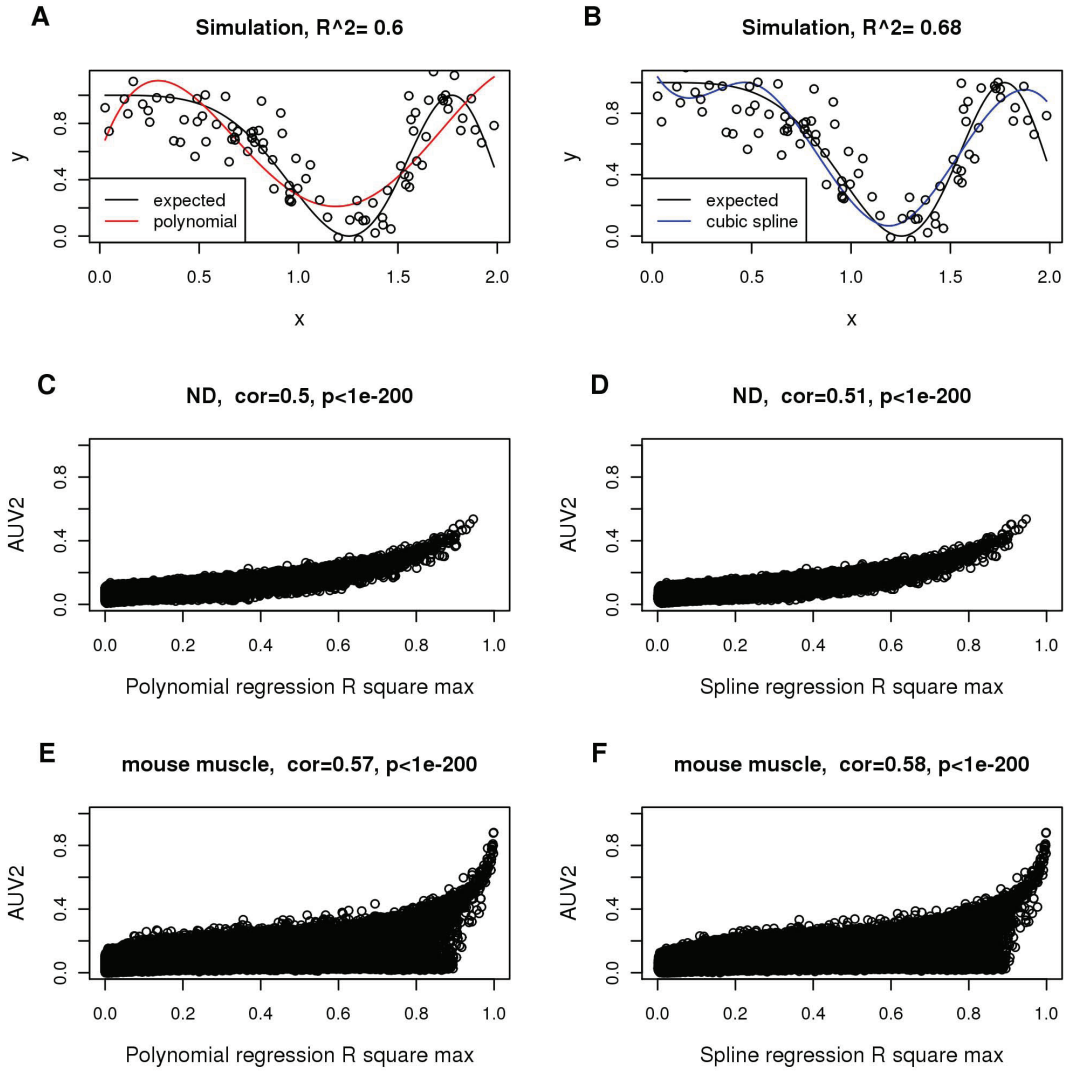

Figure 1.9: **Fitting polynomial and spline regression models to measure non-linear relationships. (A-B)** A pair of simulated data x, y (black dots) with the black curve illustrating the true expected value $E(y|x)$, where $E(y|x) = cos(x^2)^2$. Red curve: a polynomial regression model with degree $d = 4$. Blue curve: a cubic spline regression model with 2 knots. Fitting indices of the two models are shown at the top. **(C-D)** Comparisons of regression models and mutual information based co-expression measures in the ND data set. Co-expression of probe pairs is measured with polynomial (d=3)/cubic spline regressions and $A_{MI,UniversalVersion2}$. The Spearman correlation and p-value of the two measures are shown at the top. **(E-F)** Comparisons in the mouse muscle data set.

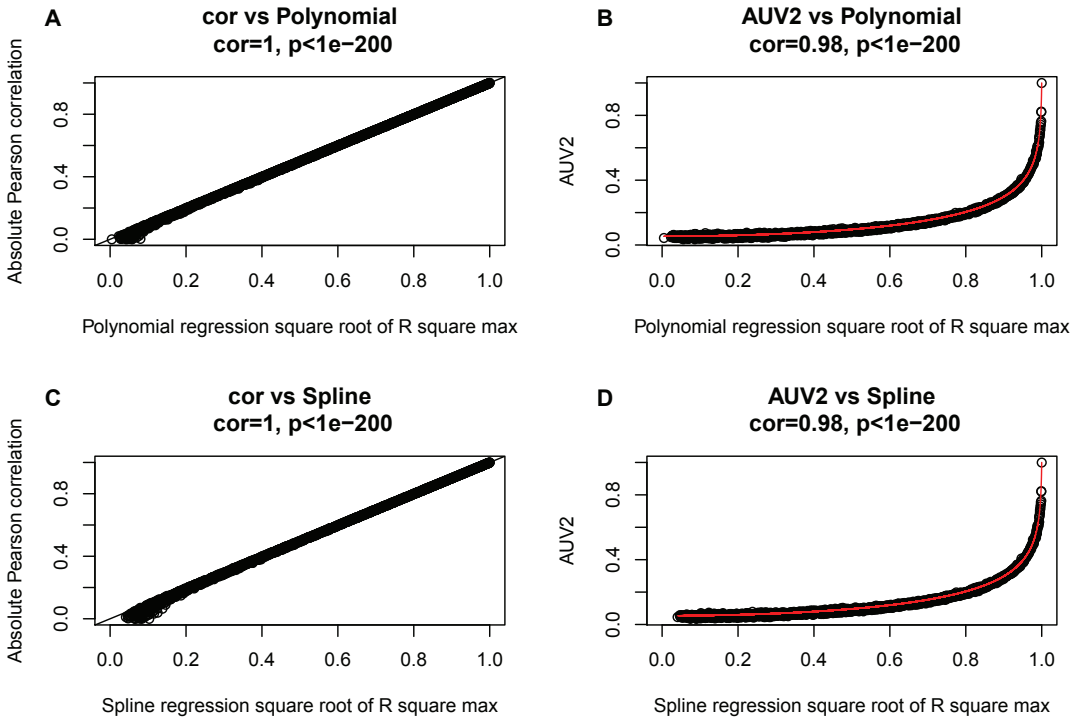

Figure 1.10: **Compare polynomial and spline regression models to correlation or mutual information based co-expression measures in simulation.** Each point corresponds to a pair of numeric vectors x and y with length $m = 1000$. **(A)** Square root of R^2 from polynomial regression symmetrized by Eq. 1.5 versus absolute Pearson correlation values. **(B)** R^2 from polynomial regression symmetrized by Eq. 1.5 versus $A^{MI,UniversalVersion2}$. The red line predicts $A^{MI,UniversalVersion2}$ from R^2. **(C-D)** Same plots for spline regression models.

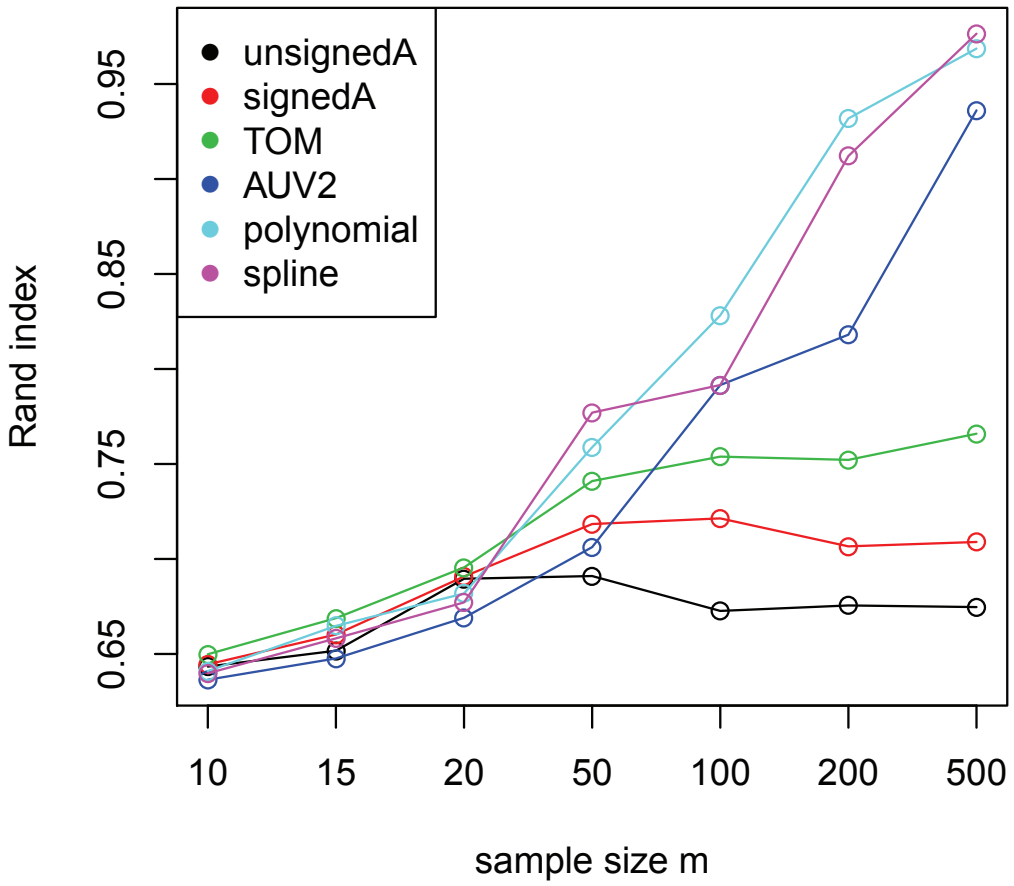

Figure 1.11: **Rand indices in simulations with various number of observations.** Simulation sample size versus Rand indices between inferred and simulated module assignments from different network inference methods. increase as the simulation data set contains more samples. Non-linear measures, especially polynomial and spline regression models, outperform other measures as sample size increases.

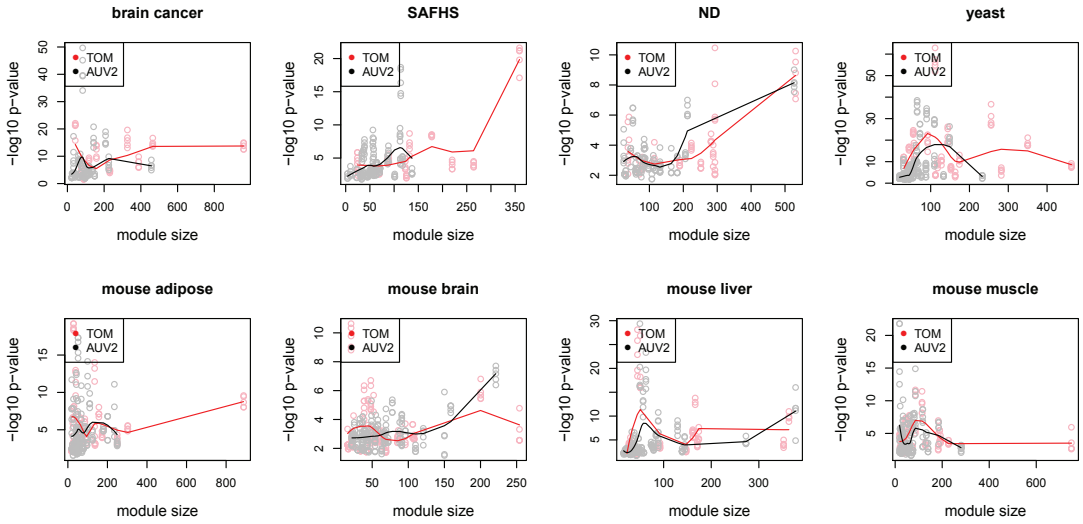

Figure 1.12: **The relationship between module size and gene ontology enrichment p-values in 8 real data applications.** In each panel, module size (x-axis) is plotted against $-\log 10$ GO enrichment p-values (y-axis)in dots. Loess regression lines are provided to show the trend. Red and black color represent network modules constructed using TOM and $A^{MI,UniversalVersion2}$ based measures, respectively. In most data sets, the enrichment of modules defined by TOM is better than that of comparably sized modules defined by $A^{MI,UniversalVersion2}$.

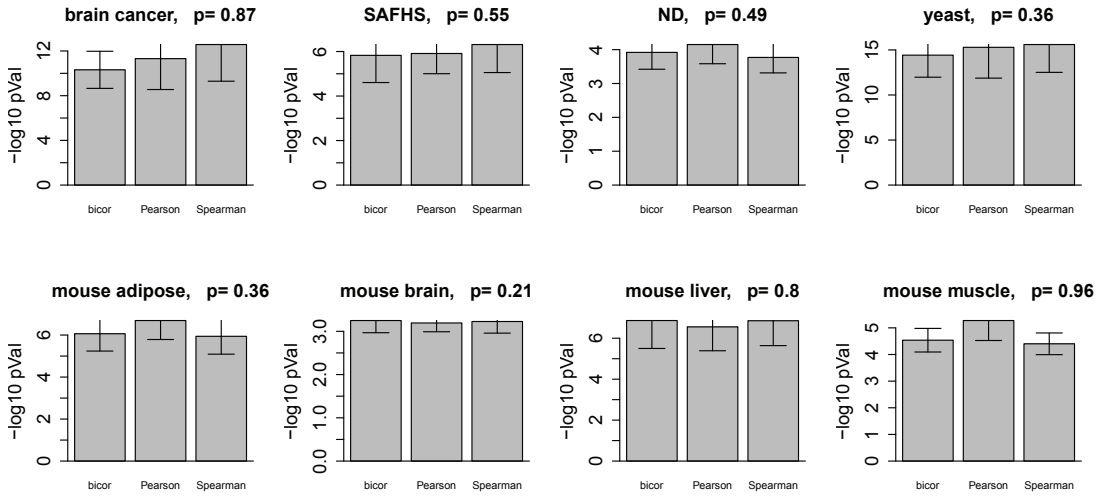

Figure 1.13: **Comparison of bicor, Pearson correlation and Spearman correlation based signed adjacency in** 8 **empirical data sets.** Each panel show the $-\log 10$ transformed 5 best gene ontology enrichment p-values of all modules identified using each type of adjacency. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. All three types of correlation are similar in terms of GO enrichment.

CHAPTER 1 TABLES

Table 1.1: **Types of networks and characteristics.**

Network type	**Examples**	**Variable types**	**Ease of estimation**	**Utility for modeling**						**Adjacency used here**	**Used in GO enrichment analysis**
				GRN	Reduce	Direct	Time	Nonlin.	Sign		
Correlation network	WGCNA [5]	Numeric	Easy	Yes	Yes	No	Maybe	No	Yes	unsignedA	Yes
										signedA	Yes
										TOM	Yes
Polynomial or Spline regression network	WGCNA [5]	Numeric	Moderate	Yes	Yes	No	Maybe	Yes	No	$polyR^2$	No
										$splineR^2$	No
Mutual information network	ARACNE [9], RELNET [6, 28], CLR [26], MRNET [27], MIC [35]	Discretized numeric, categorical	Moderate	Yes	Not clear	No	Maybe	Yes	No	ASU	No
										AUV1	No
										AUV2	Yes
										ARACNE	Yes
										ARACNE0.2	Yes
										ARACNE0.5	Yes
										CLR	Yes
										MRNET	Yes
										RELNET	Yes
										MIC	Yes
Boolean network	Boolean network [71]	Dichotomized numeric	Moderate	Yes	Not clear	Yes	Yes	NA	NA	No	No
Probabilistic network	Bayesian network [72, 73]	Any	Hard	Yes	Not clear	Yes	Yes	Yes	Yes	No	No

For each network method, the table reports what kinds of biological insights can be gained and what kind of data can be analyzed. Column "GRN" indicates whether the network has been (or can be) used for studying gene regulatory networks. Column "Reduce" indicates whether the method has been used for reducing high dimensional data (e.g. via modules and their representatives). Column "Direct" indicates whether the the network can encode directional information. Column "time" indicates whether the network method is suited for studying time series data. Column "Nonlin. " indicates whether the network can capture non-linear relationships between pairs of variables (represented as nodes). Column "Sign" indicates whether the network adjacency provides information on the sign of the relationship between two variables, e.g. a correlation coefficient can take on positive and negative values. The table entry "NA" stands for not applicable. Adjacency used here: unsignedA: unsigned bicor; signedA: signed bicor; TOM: TOM transformed signed bicor; ASU: $A^{MI,SymmetricUncertainty}$; AUV1: $A^{MI,UniversalVersion1}$; AUV2: $A^{MI,UniversalVersion2}$; ARACNE: ARACNE, $\epsilon = 0$; ARACNE0.2: ARACNE, $\epsilon = 0.2$; ARACNE0.5: ARACNE, $\epsilon = 0.5$.

CHAPTER 2

Random generalized linear model: a highly accurate and interpretable ensemble predictor

Introduction

Prediction methods (also known as classifiers, supervised machine learning methods, regression models, prognosticators, diagnostics) are widely used in biomedical research. For example, reliable prediction methods are essential for accurate disease classification, diagnosis and prognosis. Since prediction methods based on multiple features (also known as covariates or independent variables) can greatly outperform predictors based on a single feature [74], it is important to develop methods that can optimally combine features to obtain high accuracy. Introductory text books describe well known prediction methods such as linear discriminant analysis (LDA), K-nearest neighbor (KNN) predictors, support vector machines (SVM) [75], and tree predictors [76]. Many publications have evaluated popular prediction methods in the context of gene expression data [77–82].

Ensemble predictors are particularly attractive since they are known to lead to highly accurate predictions. An ensemble predictor generates and integrates multiple versions of a single predictor (often referred to as base learner), and arrives at a final prediction by aggregating the predictions of multiple base learners, e.g. via plurality voting across the ensemble. One particular approach for constructing an ensemble predictor is bootstrap aggregation (bagging) [83]. Here multiple versions of the original data are generated through bootstrapping, where observations from the training set are randomly sampled with replacement. An individual predictor (e.g. a tree predictor) is fitted on each bootstrapped data set. Thus, 100 bootstrapped data sets (100 bags) will lead to an ensemble of 100 tree predictors. In case of a class outcome (e.g. disease status), the individual predictors "vote" for each class and the final prediction is obtained by majority voting.

Breiman (1996) showed that bagging weak predictors (e.g. tree predictors or

forward selected linear models) often yields substantial gains in predictive accuracy [83]. But it seems that ensemble predictors are only very rarely used for predicting clinical outcomes. This fact points to a major weakness of ensemble predictors: they typically lead to "black box" predictions that are hard to interpret in terms of the underlying features. Clinicians and epidemiologists prefer forward selected regression models since the resulting predictors are highly interpretable: a linear combination of relatively few features can be used to predict the outcome or the probability of an outcome. But the sparsity afforded by forward feature selection comes at an unacceptably high cost: forward variable selection (and other variable selection methods) often greatly overfit the data which results in unstable and inaccurate predictors [84,85]. Ideally, one would want to combine the advantages of ensemble predictors with those of forward selected regression models. As discussed below, multiple articles describe ensemble predictors based on linear models including the seminal work by Breiman [83] who evaluated a bagged forward selected linear regression model. However, the idea of bagging forward selected linear models (or other GLMs) appears to have been set aside as new ensemble predictors, such as the random forest, became popular. A random forest (RF) predictor not only bags tree predictors but also introduces an element of randomness by considering only a randomly selected subset of features at each node split [86]. The number of randomly selected features, *mtry*, is the only parameter of the random forest predictor. The random forest predictor has deservedly received a lot of attention for the following reasons: First, the bootstrap aggregation step allows one to use out-of-bag (OOB) samples to estimate the predictive accuracy. The resulting OOB estimate of the accuracy often obviates the need for cross-validation and other resampling techniques. Second, the RF predictor provides several measures of feature (variable) importance. Several articles explore the use of these importance measures to select genes [78, 86, 87].

Third, it can be used to define a dissimilarity measure that can be used in clustering applications [86, 88]. Fourth, and most importantly, the RF predictor has superior predictive accuracy. It performs as well as alternatives in cancer gene expression data sets [78] but it really stands out when applied to the UCI machine learning benchmark data sets where it is as good as (if not better than) many existing methods [86]. While we confirm the truly outstanding predictive performance of the RF, the proposed RGLM method turns out to be even more accurate than the RF (e.g. across the disease gene expression data sets). Breiman and others have pointed out that the black box predictions of the RF predictor can be difficult to interpret. For this reason, we wanted to give bagged forward selected generalized linear regression models another careful look. After exploring different approaches for injecting elements of randomness into the individual GLM predictors, we arrived at a new ensemble predictor, referred to as random GLM predictor, with an astonishing predictive performance. An attractive aspect of the proposed RGLM predictor is that it combines the advantages of the RF with that of a forward selected GLM. As the name *generalized* linear model indicates, it can be used for a general outcome such as a binary outcome, a multi-class outcome, a count outcome, and a quantitative outcome. We show that several incremental (but important) changes to the original bagged GLM predictor by Breiman add to up to a qualitatively new predictor (referred to as random GLM predictor) that performs at least as well as the RF predictor on the UCI benchmark data sets. While the UCI data are the benchmark data for evaluating predictors, only a dozen such data sets are available for binary outcome prediction. To provide a more comprehensive empirical comparison of the different prediction methods, we also consider over 700 comparisons involving gene expression data. In these genomic studies, the RGLM method turns out to be slightly more accurate than the considered alternatives. While the improvements in accuracy afforded by the

RGLM are relatively small they are statistically significant.

This article is organized as follows. First, we present a motivating example that illustrates the high prediction accuracy of the RGLM. Second, we compare the RGLM with other state of the art predictors when it comes to binary outcome prediction. Toward this end, we use the UCI machine learning benchmark data, over 700 empirical gene expression comparisons, and extensive simulations. Third, we compare the RGLM with other predictors for quantitative (continuous) outcome prediction. Fourth, we describe several variable importance measures and show how they can be used to define a thinned version of the RGLM that only uses few important features. Even for data sets comprised of thousands of gene features, the thinned RGLM often involves fewer than 20 features and is thus more interpretable than most ensemble predictors.

Materials and Methods

A. Construction of the RGLM predictor

RGLM is an ensemble predictor based on bootstrap aggregation (bagging) of generalized linear models whose features (covariates) are selected using forward regression according to AIC criterion. GLMs comprise a large class of regression models, e.g. linear regression for a normally distributed outcome, logistic regression for binary outcome, multi-nomial regression for multi-class outcome and Poisson regression for count outcome [89]. Thus, RGLM can be used to predict binary-, continuous-, count-, and other outcomes for which generalized linear models can be defined. The "randomness" in RGLM stems results from two sources. First, a non-parametric bootstrap procedure is used which randomly selects samples with replacement from the original data set. Second, a random subset of features (specified by input parameter $nFeaturesInBag$) is selected for each bootstrap sample. This amounts to a random sub-space method [90] applied to each bootstrap sample separately.

The steps of the RGLM construction are presented in Figure 2.1. First, starting from the original data set another equal-sized data set is generated using the non-parametric bootstrap method, i.e. samples are selected with replacement from the original data set. The parameter *nBags* (default value 100) determines how many of such bootstrap data sets (referred to as bags) are being generated. Second, a random set of features (determined by the parameter *nFeaturesInBag*) is randomly chosen for each bag. Thus, the GLM predictor for bag 1 will typically involve a different set of features than that for bag 2. Third, the *nFeaturesInBag* of randomly selected features per bag are rank-ordered according to their individual association with the outcome variable y in each bag. For a quantitative outcome y, one can simply use the absolute value of the correlation coefficient between

the outcome and each feature to rank the features. More generally, one can fit a univariate GLM model to each feature to arrive at an association measure (e.g. a Wald test statistic or a likelihood ratio test). Only the top ranking features (i.e. features with the most significant univariate significance levels) will become candidate covariates for forward selection in a multivariate regression model. The top number of candidate features is determined by the input parameter *nCandidateCovariates* (default value 50). Fourth, forward variable (feature) selection is applied to the *nCandidateCovariates* of each bag to arrive at a multivariable generalized linear model per bag. The forward selection procedure used by RGLM is based on the *stepAIC* R function in the *MASS* R library where method is set to "forward". Fifth, the predictions of each forward selected multivariate model (one per bag) are aggregated across bags to arrive at a final ensemble prediction. The aggregation method depends on the type of outcome. For a continuous outcome, predicted values are simply averaged across bags. For a binary outcome, the adjusted Majority Vote (aMV) strategy [91] is used which averages predicted probabilities across bags. Given the estimated class probabilities one can get a binary prediction by choosing an appropriate threshold (default value 0.5).

Importantly, RGLM also has a parameter *maxInteractionOrder* (default value 1) for creating interactions up to a given order among features in the model construction. For example, RGLM.inter2 results from setting *maxInteractionOrder*=2, i.e. considering pairwise (also known as 2-way) interaction terms. As example, consider the case when only pairwise interaction terms are used. For each bag a random set of features is selected (similar to the random subspace method, RSM) from the original covariates, i.e. covariates without interaction terms. Next, all pairwise interactions among the *nFeaturesInBag* randomly selected features are generated. Next, the usual RGLM candidate feature selection steps will be applied to the combined set of pairwise interaction terms and the

nFeaturesInBag randomly selected features per bag resulting in *nCandidateCovariates* top ranking features per bag, which are subsequently subjected to forward feature selection.

These methods are implemented in our R software package *randomGLM* which allows the user to input a training set and optionally a test set. It automatically outputs out-of-bag estimates of the accuracy and variable importance measures.

Parameter choices for the RGLM predictor

As discussed below, we find that it is usually sufficient to consider only $nBags = 100$ bootstrap data sets. The default value for *nFeaturesInBag* depends on the total number of features. It is easier to explain it in terms of the proportion of features randomly selected per bag, $nFeaturesInBag/N$, where N is the total number of features of the training set. Apart from N it is also valuable to consider the effective number of features which equals the number of features N plus the number of interaction terms, e.g. $N^* = N + N(N-1)/2$ in case of pairwise interactions. Using this notation, the default value of nFeaturesInBag can be arrived at by solving equations presented in Table 2.1. These equations were found by empirically evaluating various choices of $nFeaturesInBag$ values (e.g. $\sqrt{N}, N/5, N/3, N/2, 2N/3, N$). In particular, we found that in case of $N^* <= 10$, then using all features (i.e setting $nFeaturesInBag/N = 1$) is often a good choice, whereas if $N^* > 300$ then setting $nFeaturesInBag/N = 0.2$ works well. The default value $nFeaturesInBag/N = 1.0276 - 0.00276N^*$ in the intermediate case ($10 < N^* <= 300$) results from fitting an interpolation line through the two points (10,1) and (300, 0.2). We find that RGLM is quite robust with respect to the parameter *nFeaturesInBag*. To limit the number of covariates considered in forward selection (which is computationally intensive), the default value of *nCandidateCovariates* is set to 50. Overall, the default values perform well in our

simulations, empirical gene expression and machine learning benchmark studies. But we recommend to use the OOB estimate of predictive accuracy to inform the choice of the parameter values.

B. Relationship with related prediction methods

As discussed below, RGLM can be interpreted as a variant of a bagged predictor [83]. In particular, it is similar to the bagged forward linear regression model [83] but differs in the following aspects:

1. RGLM allows for interaction terms between features which greatly improve the performance on some data sets (in particular the UCI benchmark data sets). We refer to RGLM involving two-way or three way interactions as RGLM.inter2 and RGLM.inter3, respectively.

2. RGLM has a parameter $nFeaturesInBag$ that allows one to restrict the number of features used in each bootstrap sample. This parameter is conceptually related to the $mtry$ parameter of the Random Forest predictor. In essence, this parameter allows one to use a random subspace method (RSM, [90]) in each bootstrap sample.

3. RGLM has a parameter $nCandidateCovariates$ that allows one to restrict the number of features in forward regression, which not only has computational advantages but also introduces additional instability into the individual predictors, which is a desirable characteristic of an ensemble predictor.

4. RGLM optimizes the AIC criterion during forward selection.

5. RGLM has a "thinning threshold" parameter which allows one to reduce the number of features involved in prediction while maintaining good prediction

accuracy. Since a thinned RGLM involves far fewer features, it facilitates the understanding how the ensemble arrives at its predictions.

RGLM is not only related to bagging but also to the random subspace method (RSM) proposed by [90]. In the RSM, the training set is also repeatedly modified as in bagging but this modification is performed in the feature space (rather than the sample space). In the RSM, a subset of features is randomly selected which amounts to restricting attention to a subspace of the original feature space. As one of its construction steps, RGLM uses a RSM on each bootstrap sample. Future research could explore whether random partitions as opposed to random subspaces would be useful for constructing an RGLM. Random partitions of the feature space are similar to random subspaces but they divide the feature space into mutually exclusive subspaces [92, 93]. Random partition based predictors have been shown to perform well in high-dimensional data ([92]). Both RSM and random partitions have more general applicability than RGLM since these methods can be used for any base learner. There is a vast literature on ensemble induction methods but a property worth highlighting is that RGLM uses forward variable selection of GLMs. Recall that RGLM goes through the following steps: 1) bootstrap sampling, 2) RSM (and optionally creating interaction terms), 3) forward variable selection of a GLM, 4) aggregation of votes. Empirical studies involving different base learners (other than GLMs) have shown that combining bootstrap sampling with RSM (steps 1 and 2) leads to ensemble predictors with comparable performance to that of the random forest predictor [94].

Another prediction method, random multinomial logit model (RMNL), also shares a similar idea with RGLM. It was recently proposed for multi-class outcome prediction [91]. RMNL bags multinomial logit models with random feature selection in each bag. It can be seen as a special case of RGLM, except that no forward model selection is carried out.

C. Software implementation

The RGLM method is implemented in the freely available R package *randomGLM*. The R function *randomGLM* allows the user to output training set predictions, out-of-bag predictions, test set predictions, coefficient values, and variable importance measures. The *predict* function can be used arrive at test set predictions. Tutorials can be found at the following webpage: http://labs.genetics.ucla.edu/horvath/RGLM.

D. Short description of alternative prediction methods

Forward selected generalized linear model predictor (forwardGLM). We denote by *forwardGLM* the (single) generalized linear model predictor whose covariates were selected using forward feature selection (according to the AIC criterion). Thus, forwardGLM does not involve bagging, random feature selection, and is not an ensemble predictor.

Random forest (RF). RF is an ensemble predictor that consists of a collection of decision trees which vote for the class of observations [86]. The RF is known for its outstanding predictive accuracy. We used the *randomForest* R package in our studies. We considered two choices for the RF parameter *mtry*: i) the default RF predictor where *mtry* equals the square root of the number of features and ii) *RFbigmtry* where *mtry* equals the total number of features. We always generated at least 500 trees per forest but used 1000 trees when calculating variable importance measures.

Recursive partitioning and regression trees (Rpart). Classification and regression trees were generated using the default settings *rpart* R package. Tree methods are described in [76].

Linear discriminant analysis (LDA). LDA aims to find a linear combina-

tion of features (referred to as discriminant variables) to predict a binary outcome (reviewed in [95, 96]). We used the *lda* R function in the *MASS* R package with parameter choice $method = moment$.

Diagonal linear discriminant analysis (DLDA). DLDA is similar to LDA but it ignores the correlation patterns between features. While this is often an unrealistic assumption, DLDA (also known as gene voting) has been found to work well in in gene expression applications [77]. Here we used the default parameters from the *supclust* R package [97].

K nearest neighbor (KNN). We used the *knn* R function in the *class* R package [95, 96], which chose the parameter k of nearest neighbors using 3-fold cross validation (CV).

Support vector machines (SVM). We used the default parameters from the *e1071* R package to fit SVMs [75]. Additional details can be found in [98].

Shrunken centroids (SC). The SC predictor is known to work well in the context of gene expression data [99]. Here we used the implementation in the *pamr* R package [99] which chose the optimal level of shrinkage using cross validation.

Penalized regression models. Various convex penalties can be applied to generalized linear models. We considered ridge regression [100] corresponding to an ℓ_2 penalty, the lasso corresponding to an ℓ_1 penalty [101], and elastic net corresponding to a linear combination of ℓ_1 and ℓ_2 penalties [102]. We used the *glmnet* R function from the *glmnet* R package [103, 104] with alpha parameter values of $0, 1$, and 0.5 respectively. *glmnet* also involves another parameter (lambda) which was chosen as the median of the lambda sequence output resulting from *glmnet*. For UCI benchmark data sets, pairwise interaction between features were considered.

E. 20 disease-related gene expression data sets

We use 20 disease related gene expression data sets involving cancer and other human diseases (described in Table 2.2). The first 10 data sets involving various cancers were previously used by [78]. These data can be downloaded from the author's webpage at http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html. The BrainTumor2 and DLBCL data sets were downloaded from http://www.gems-system.org/. The remaining 8 data sets (lung1 – MSdiagnosis2) were downloaded from either the Gene Expression Omnibus (GEO) database or the ArrayExpress data base in raw form and subsequently preprocessed using MAS5 normalization and quantile normalization. Only the top 10000 probes (features) with highest mean expression were considered for outcome prediction. We briefly point out that Diaz et al (2006) report prediction error rates estimated using a bootstrap method. In contrast, we report 3-fold cross validation estimates (averaged over 100 random partitions of the data into 3 folds), which may explain minor numerical differences between our study and that of Diaz et al (2006).

F. Empirical gene expression data sets

For all data sets below, we considered 100 randomly selected gene traits, i.e. 100 randomly selected probes. They were directly used as continuous outcomes or dichotomized according to the median value (top half = 1, bottom half = 0) to generate binary outcomes. For all data sets except "Brain cancer", $\frac{2}{3}$ of the observations (arrays) were randomly chosen as the training set, while the remaining samples were chosen as test set. We focused on the 5000 genes (probes) with the highest mean expression levels in each data set.

Brain cancer data sets. These two related data sets contain 55 and 65 microarray samples of glioblastoma (brain cancer) patients, respectively. Gene

expression profiles were measured using Affymetrix U133 microarrays. A detailed description can be found in [61]. The first data set (comprised of 55 samples) was used as a training set while and the second data set (comprised of 65 samples) was used as a test set.

SAFHS blood lymphocyte data set. This data set [62] was derived from blood lymphocytes of randomly ascertained participants enrolled in the San Antonio Family Heart Study. Gene expression profiles were measured with the Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChips. After removing potential outliers (based on low interarray correlations), 1084 samples remained in the data set.

WB whole blood gene expression data set. This is the whole blood gene expression data from healthy controls. Peripheral blood samples from healthy individuals were analyzed using Illumina Human HT-12 microarrays. After preprocessing, 380 samples remained in the data set.

Mouse tissue gene expression data sets. The 4 tissue specific gene expression data sets were generated by the lab of Jake Lusis at UCLA. These data sets measure gene expression levels (Agilent array platform) from adipose (239 samples), brain (221 samples), liver (272 samples) and muscle (252 samples) tissue of mice from the B×H F_2 mouse intercross described in [65, 66]. In addition to gene traits, we also predicted 21 quantitative mouse clinical traits including mouse weight, length, abdominal fat, other fat, total fat, adiposity index (total fat*100/weight), plasma triglycerides, total plasma cholesterol, high-density lipoprotein fraction of cholesterol, plasma unesterified cholesterol, plasma free fatty acids, plasma glucose, plasma low-density lipoprotein and very low-density lipoprotein cholesterol, plasma MCP-1 protein levels, plasma insulin, plasma glucose-insulin ratio, plasma leptin, plasma adiponectin, aortic lesion size (measured by histological examination using a semi-quantitative scoring methods),

aneurysms (semi-quantitative scoring method), and aortic calcification in the lesion area.

G. Machine learning benchmark data sets

The 12 machine learning benchmark data sets used in this article are listed in Table 2.3. Note that only eight of the 12 data sets have a binary outcomes. The multi-class outcomes of the 4 remaining data sets were turned into binary outcomes by considering the most prevalent class versus all other classes combined. Missing data were imputed using nearest neighbor averaging. For each data set and prediction method, we report the average 3-fold CV estimate of prediction accuracy over 100 random partitions of the data into 3 folds.

H. Simulated gene expression data sets

We simulated an outcome variable y and gene expression data that contained 5 modules (clusters). Only 2 of the modules were comprised of genes that correlated with the outcome y. 45% of the genes were background genes, i.e. these genes were outside of any module. The simulation scheme is implemented in the R function *simulateDatExpr5Modules* from the *WGCNA* R package [5]. This R function was used to simulate pairs of training and test data sets. The simulation study was used to evaluate prediction methods for continuous outcomes and for binary outcomes. For binary outcome prediction, the continuous outcome y was thresholded according to its median value.

We considered 180 different simulation scenarios involving varying sizes of the training data (50, 100, 200, 500, 1000 or 2000 samples) and varying numbers of genes $(60, 100, 500, 1000, 5000$ or 10000 genes) that served as features. Test sets contained the same number of genes as in the corresponding training set and 1000

samples. For each simulation scenario, we simulate 5 replicates resulting from different choices of the random seed.

Results

A. Motivating example: disease-related gene expression data sets

We compare the prediction accuracy of RGLM with that of other widely used methods on 20 gene expression data sets involving human disease related outcomes. Many of the 20 data sets (Table 2.2) are well known cancer data sets, which have been used in other comparative studies [77, 78, 124, 125]. A brief description of the data sets can be found in Materials and Methods.

To arrive at an unbiased estimate of prediction accuracy, we used 3-fold cross validation (averaged over 100 random partitions of the data into 3 folds). Note that the accuracy equals 1 minus the median misclassification error rate. Table 2.4 reports the prediction accuracy of different methods including RGLM, random forest (RF, with default value for its *mtry* parameter), random forest (RFbigmtry, with *mtry* equal to the total number of features), tree predictor (also known as recursive partitioning, Rpart), linear discriminant analysis (LDA), diagonal linear discriminant analysis (DLDA), k nearest neighbor (KNN), support vector machine (SVM) and shrunken centroid (SC). A short description of these prediction methods is provided in Materials and Methods.

As seen from Table 2.4, RGLM achieves the highest mean accuracy in these disease data sets, followed by RFbigmtry and SC. Note that the standard random forest predictor (with default parameter choice) performs worse than RGLM. The accuracy difference between RGLM and alternative methods is statistically significant (Wilcoxon signed rank test < 0.05) for all predictors except for RFbigmtry, DLDA and SC. Since RFbigmtry is an ensemble predictor that relies on thousands of features it would be difficult to interpret its predictions in terms of the underlying genes.

Our evaluations focused on the accuracy (and misclassification error). How-

ever, a host of other accuracy measures could be considered. We calculated sensitivity and specificity. The top 3 methods with highest sensitivity are: RF (median sensitivity= 0.969), SVM (0.969) and RGLM (0.960). The top 3 methods with highest specificity are: SC (0.900), RGLM (0.857) and KNN (0.848).

A strength of this empirical comparison is that it involves clinically or biologically interesting data sets but a severe limitation is that it only involves 20 comparisons. Therefore, we now turn to more comprehensive empirical comparisons.

B. Binary outcome prediction

Empirical study involving dichotomized gene traits

Many previous empirical comparisons of gene expression data considered fewer than 20 data sets. To arrive at 700 comparisons, we use the following approach: We started out with 7 human and mouse gene expression data sets. For each data set, we randomly chose 100 genes as gene traits (outcomes) resulting in 7×100 possible outcomes. We removed the gene corresponding to the gene trait from the feature set. Next, each gene trait was dichotomized by its median value to arrive at a binary outcome y. The goal of each prediction analysis was to predict the dichotomized gene trait y based on the other genes. At first sight, this artificial outcome is clinically uninteresting but it is worth emphasizing that clinicians often deal with dichotomized measures of gene products, e.g. high serum creatinine levels may indicate kidney damage, high PSA levels may indicate prostate cancer, and high HDL levels may indicate hypercholesterolemia. To arrive at unbiased estimates of prediction accuracy, we split each data set into a training and test set. Figure 2.2 (A) shows boxplots of the accuracies across the 700 comparisons. Similar performance patterns are observed for the individual data sets (Figure

2.2 (B-H)). The figure also reports pairwise comparisons of the RGLM method versus alternative methods. Specifically, it reports the two-sided Wilcoxon signed rank test p-values for testing whether the accuracy of the RGLM predictor is higher than that of the considered alternative method. Strikingly, RGLM is more accurate than the other methods overall. While the increase in accuracy are often minor, they are statistically significant as can be seen by comparing RGLM to RF (median difference $= 0.02, p = 2.1 \times 10^{-51}$), RFbigmtry (median difference $= 0.01, p = 7.3 \times 10^{-16}$), LDA (median difference $= 0.06, p = 2.4 \times 10^{-53}$), SVM (median difference $= 0.03, p = 1.8 \times 10^{-62}$) and SC (median difference $= 0.04, p = 4.3 \times 10^{-71}$). Other predictors perform even worse, and the corresponding p-values are not shown.

The fact that RFbigmtry is more accurate in this situation than the default version of RF probably indicates that relatively few genes are informative for predicting a dichotomized gene trait. Also note that RGLM is much more accurate than the unbagged forward selected GLM which reflects that forward selection greatly overfits the training data. In conclusion, these comprehensive gene expression studies show that RGLM has outstanding prediction accuracy.

Machine learning benchmark data analysis

Here we evaluate the performance of RGLM on the UCI machine learning benchmark data sets which are often used for evaluating prediction methods [83, 86, 126–129]. We consider 12 benchmark data sets from the *mlbench* R package: 9 UCI data sets and 3 synthetic data sets (Table 2.3). We choose these data sets for two reasons. First, these 12 data sets were also used in the original evaluation of the random forest predictor [86]. Second, these data include all of the available data sets with binary outcomes in the *mlbench* R package. A detailed description of these data sets can be found in Materials and Methods. In his orig-

inal publication on the random forest, Breiman found that the RF outperformed bagged predictors on the UCI benchmark data which may explain why bagged GLMs have not received much attention. We hypothesize that the relatively poor performance of a bagged logistic regression model on these data sets could be ameliorated by considering interaction terms between the features. Table 2.5 confirms our hypothesis. RGLM.inter2 (corresponding to pairwise interaction terms) has superior or tied accuracy compared to RGLM in 10 out of 12 benchmark data sets. In particular, pairwise interactions greatly improve the prediction accuracy in the ringnorm data set. Higher order interactions (RGLM.inter3) do not perform better than RGLM.inter2 but dramatically increase computational burden (data not shown).

Overall, we find that RGLM.inter2 ties with SVM ($diff = -0.001, p = 0.96$) and RF ($diff = 0.001, p = 0.26$) for the first place in the benchmark data. Moreover, RGLM.inter2 achieves the highest sensitivity and specificity (data not shown), which also support its good performance in the benchmark data sets.

A potential limitation of these comparisons is that we considered pairwise interaction terms for the RGLM predictor but not for the other predictors. To address this issue, we also considered pairwise interactions among features for other predictors. Table 2.6 shows that no method surpasses RGLM.inter2 when pairwise interaction terms are considered. In particular, interaction terms between features do not improve the performance of the random forest predictor. A noteworthy disadvantage of RGLM.inter in case of many features is the computational burden that may result from adding interaction terms. In applications where interaction terms are needed for RGLM, faster alternatives (e.g. RF) remain an attractive choice.

Simulation study involving binary outcomes

As described in Materials and Methods, we simulated 180 gene expression data sets with binary outcomes. The number of features (genes) ranged from 60 to 10000. The sample sizes (number of observations) of the training data ranged from 50 to 2000. To robustly estimate the test set accuracy we chose a large size for the corresponding test set data, $n = 1000$. Figure 2.3 shows the boxplots of the test set accuracies of different predictors. The accuracy of the forwardGLM is much lower than that of RGLM, demonstrating the benefit of creating an ensemble predictor. Overall, RGLM delivers significantly higher median accuracy than all other methods except the random forest (with default parameter setting).

C. Continuous outcome prediction

In the following, we show that RGLM also performs exceptionally well when dealing with continuous quantitative outcomes. We not only compare RGLM to a standard forward selected linear model predictor (forwardGLM) but also a random forest predictor (for a continuous outcome). We do not report the findings for the k-nearest neighbor predictor of a continuous outcome since it performed much worse than the above mentioned approaches in our gene expression applications (the accuracy of a KNN predictor was decreased by about 30 percent). We again split the data into training and test sets. We use the correlation between test set predictions and truly observed test set outcomes as measure of predictive accuracy. Note that this correlation coefficient can take on negative values (in case of a poorly performing prediction method).

Empirical study involving continuous gene traits

Here we used the same 700 gene expression comparisons as described above (100 randomly chosen gene traits from each of 7 gene expression data sets) but did not dichotomize the gene traits. Incidentally, prediction methods for gene traits are often used for imputing missing gene expression values. Our results presented in Figure 2.4 indicate that for the majority of genes high accuracies can be achieved. But for some gene traits, the accuracy measure, which is defined as a correlation coefficient, takes on negative values indicating that there is no signal in the data. Note that the forward selected linear predictor ties with the random forest irrespective of the choice of the *mtry* parameter and both methods perform significantly worse than the RGLM predictor.

Mouse tissue expression data involving continuous clinical outcomes

Here we used the mouse liver and adipose tissue gene expression data sets to predict 21 clinical outcomes (detailed in Materials and Methods). Again, RGLM achieved significantly higher median prediction accuracy compared to the other predictors (Figure 2.5).

Simulation study involving continuous outcomes

180 gene expression data sets are simulated in the same way as described previously (for evaluating a binary outcome) but here the outcome y was not dichotomized. As shown in Figure 2.6, RGLM yields significantly higher prediction accuracy than other predictors, although the differences are minor. The forwardGLM accuracy trails both RGLM and RF, reflecting again the fact that forward regression overfits the data.

D. Comparing RGLM with penalized regression models

In our previous comparisons, we found that RGLM greatly outperforms forward selected GLM methods based on the AIC criterion. Many powerful alternatives to forward variable selection have been developed in the literature, in particular penalized regression models. Here, we compare RGLM to 3 major types of penalized regression models: ridge regression [100], elastic net [102], and the lasso [101]. The predictive accuracies of these penalized regression models were compared to those of the RGLM predictor using the same data sets described above for evaluating binary outcome and quantitative outcome prediction methods. Wilcoxon's signed rank test was used to determine whether differences in predictive accuracy were significant. Figure 2.7 (A) shows that RGLM outperforms penalized regression models when applied to binary outcomes. For all comparisons, the paired median difference (median of RGLM accuracy minus penalized regression accuracy) is positive which indicates that RGLM is at least as good if not better than any of these 3 penalized regression models. In particular, RGLM is significantly better than ridge regression ($diff = 0.025, p = 2 \times 10^{-52}$) and the lasso ($diff = 0.011, p = 7 \times 10^{-10}$) on the 700 dichotomized gene expression trait data. Also, RGLM is significantly better than elastic net ($diff = 0.022, p = 2 \times 10^{-27}$) and lasso ($diff = 0.03, p = 3 \times 10^{-28}$) in simulations with binary outcomes. Figure 2.7 (B) shows that RGLM outperforms penalized regression models for continuous outcome prediction as well. Positive accuracy differences again imply that RGLM is at least as good as these penalized regression models. In particular, it significantly outperforms ridge regression ($diff = 0.035, p = 2 \times 10^{-86}$) in the 700 continuous gene expression traits data and outperforms elastic net ($diff = 0.029, p = 4 \times 10^{-25}$) and lasso ($diff = 0.034, p = 8 \times 10^{-27}$) in simulations with continuous outcomes.

As a caveat, we mention that cross validation methods were not used to in-

form the parameter choices of the penalized regression models since the RGLM predictor was also not allowed to fine tune its parameters. By only using default parameter choices we ensure a fair comparison. In a secondary analysis , however, we allowed penalized regression models to use cross validation for informing the choice of the parameters. While this slightly improved the performance of the penalized regression models (data not shown), it did not affect our main conclusion. RGLM outperforms penalized regression models in these comparisons.

E. Feature selection

Here we briefly describe how RGLM naturally gives rise to variable (feature) importance measures. We compare the variable importance measures of RGLM with alternative approaches and show how variable importance measures can be used for defining a thinned RGLM predictor with few features.

Variable importance measure

There is a vast literature on using ensemble predictors and bagging for selecting features. For example, Meinshausen and Bühlmann describe "stability selection" based on variable selection employed in regression models [130]. The method involves repetitive sub-sampling, and variables that occur in a large fraction of the resulting selection set are chosen. Li et al. use a random k-nearest neighbor predictor (RKNN) to carry out feature selection [125]. The Entropy-based Recursive Feature Elimination (E-RFE) method of Furlanello et al. ranks features in high dimensional microarray data [131]. RGLM, like many ensemble predictors, gives rise to several measures of feature (variable) importance. For example, the number of times a feature is selected in the forward GLM across bags, $timesSelectedByForwardRegression$, is a natural measure of variable importance (similar to that used in stability selection [130]). Another variable

importance measure is the number of times a feature is selected as candidate covariate for forward regression, $timesSelectedAsCandidates$. Note that both $timesSelectedByForwardRegression$ and $timesSelectedAsCandidates$ have to be $\leq nBags$. Finally, one can use the sum of absolute GLM coefficient values, $sumAbsCoefByForwardRegression$, as a variable importance measure. We prefer $timesSelectedByForwardRegression$, since it is more intuitive and points to the features that directly contribute to outcome prediction.

To reveal relationships between different types of variable importance measures, we present a hierarchical cluster tree of RGLM measures, RF measures and standard marginal analysis based on correlations in Figure 2.8. As expected, the marginal association measures (standard Pearson correlation and the Kruskal-Wallis test which can both be used for a binary outcome) cluster together. The same holds for the random forest based importance measures ("mean decreased accuracy" and "mean decreased node purity") and the 3 RGLM based importance measures.

RGLM predictor thinning based on a variable importance measure

Both RGLM and random forest have superior prediction accuracy but they differ with respect to how many features are being used. Recall that the random forest is composed of individual trees. Each tree is constructed by repeated node splits. The number of features considered at each node split is determined by the RF parameter $mtry$. The default value of $mtry$ is the square root of the number of features. In case of 4999 gene features in our empirical studies, the default value is $mtry = 71$. For RFbigmtry, we choose all possible features, i.e. $mtry = 4999$. We find that a random forest predictor typically uses more than 40% of the features (i.e. more than 2000 genes) in the empirical studies. In contrast, RGLM typically only involves a few hundred genes in these studies. There are several reasons why

RGLM uses far fewer features in its construction. First, and foremost, it uses forward selection (coupled with the AIC criterion) to select features in each bag. Second, the number of candidate covariates considered for forward regression is chosen to be low, i.e. $nCandidateCovariates = 50$.

In RGLM, the number of times a feature is selected by forward regression models among all bags, $timesSelectedByForwardRegression$, follows a highly skewed distribution. Only few features are repeatedly selected into the model while most features are selected only once (if at all). It stands to reason that an even sparser, highly accurate predictor can be defined by refitting the GLM on each bag without considering these rarely selected features. We refer to this feature removal process as RGLM predictor thinning. Thus, features whose value of $timesSelectedByForwardRegression$ lies below a pre-specified thinning threshold will be removed from the model fit a posteriori.

Figure 2.9 presents the effects of predictor thinning in our empirical study. Here $nFeaturesInBag$ is chosen to equal the total number of features. To ensure a fair comparison, we constructed and thinned the resulting RGLM in the training set only. Next, we evaluated the accuracy of the resulting thinned predictor in a test data set. Results were averaged across the 700 studies used in Figure 2.2 (A). Figure 2.9 (A) shows that the mean (and median) test set accuracies across 700 tests gradually decreases as the thinning threshold becomes more stringent. This is expected since the predictor loses potentially informative features with increasing values of the thinning threshold. Because the number of bags, $nBags$, is chosen to be 100, $timesSelectedByForwardRegression$ takes on a value ≤ 100. Note that for a thinning threshold of 70 or larger, the median accuracy is constant at 0.5 which indicates that for at least 50% of comparisons the prediction is no longer informative. This reflects the fact that for large thinning thresholds, no covariates remain in the GLM models and the resulting predictor reduces to the

"naive predictor" which assigns a constant outcome to all observations.

Interestingly, the accuracy diminishes very slowly for initial, low threshold values. But even low threshold values lead to a markedly sparser ensemble predictor (Figure 2.9 (B)). In other words, the average fraction of features (genes) remaining in the thinned RGLM declines drastically as the thinning threshold increases.

We have found that the following empirical function accurately describes the relationship between thinning threshold ($timesSelectedByForwardRegression$ threshold) and proportion of features left in the thinned RGLM predictor:

$$propLeft = F(x) = \begin{cases} 1 & x = 0 \\ exp\{-e(ex)^{0.775} nBags^{0.0468(1-log(x))}\} & 0 < x \leq 1 \end{cases} \quad (2.1)$$

where $x = \frac{thinning\ threshold}{nBags}$ and e denotes Euler's constant $e \approx 2.718$. Eq. 2.1 was found by log transforming the data and using optimization approaches for estimating the parameters. No mathematical derivation was used. One can easily show that $F(x)$ (Eq. 2.1) is a monotonically decreasing function which accurately describes the proportion of remaining features as can be seen from Figure 2.9 (B). Since the proportion of remaining variables depends not only on the thinning threshold but also on the number of bags $nBags$, we also study how these results depend on the choice of $nBags$. Toward this end, we varied $nBags$ from 20 to 500 for predicting the 100 dichotomized gene traits in the mouse adipose data set. The predicted values (red curve) based on Eq. 2.1 overlaps almost perfectly with the observed values (black curve) for all considered choices of $nBags$ (data not shown), which indicates that Eq. 2.1 accurately estimates the proportion of remaining features for range of different values of $nBags$.

Our results demonstrate that the number of required features decreases rapidly even for low values of the thinning threshold without compromising the prediction accuracy of the thinned predictor. Figure 2.9 (C) shows that a thinning

threshold of 20, leads to a thinned predictor whose accuracy is negligibly lower (difference in median accuracy=0.009) than that of the original RGLM predictor but it involves less than 20% of the original number of variables. Recall that even the original number of variables is markedly lower than that of the RF predictor. These results demonstrate that the thinned RGLM combines the advantages of an ensemble predictor (high accuracy) with that of a forward selected GLM model (few features, interpretability).

RGLM thinning versus RF thinning

The idea behind RGLM thinning is to remove features with low values of the variable importance measure. Of course, a similar idea can be applied to other predictors. Here we briefly evaluate the performance of a thinned random forest predictor which removed variables based on a low value of its importance measure ("mean decreased accuracy"). To arrive at an unbiased comparison, both RGLM and RF are thinned based on results obtained in the training data. Next, accuracies of the thinned predictors are evaluated in the test set data. Figure 2.10 compares thinned RGLM versus thinned RF in our disease related data sets and also the empirical studies. Numbers that connect dashed lines are RGLM thinning thresholds. For a pre-specified threshold, the number of features used in the thinned random forest is matched to that used in the thinned RGLM (except for the threshold 0). Without thinning, RF uses a lot more features than RGLM as mentioned previously. As expected, the median number of genes left for prediction and the corresponding median prediction accuracy generally decrease as the thinning threshold becomes more stringent. Overall, a thinned RGLM yields a significantly higher median accuracy than a thinned RF across different thinning thresholds (see the paired Wilcoxon signed rank test p-values). In clinical practice, a thinned predictor with very few features and good accuracy can be

very useful and interpretable. For example, choosing a threshold of 5 in panel Figure 2.10 (A) and a threshold of 35 in panel (B) would result in very sparse predictors. In both cases, especially in panel (A), the thinned RGLM has higher median accuracy than that of the thinned RF.

Discussion

A. Why was the RGLM not discovered earlier?

After Breiman proposed the idea of bagged linear regression models in 1996 [83], many authors have explored the utility of bagging logistic regression models [133–139]. Most previous studies report that bagging does not improve the accuracy of logistic regression. Bühlman and Yu showed theoretically that bagging helps for "hard threshold" methods but not for "soft threshold" methods (such as logistic regression) [140]. These studies indicate that bagged logistic regression models are not beneficial since the individual predictors (logistic regression models) are too stable. Overall, we agree with these results. But our comprehensive evaluations show that by injecting elements of randomness and instability into a bagged logistic regression model one arrives at a state of the art prediction method that often outperforms existing methods. Figure 2.11 describes why the construction of the RGLM runs counter to conventional wisdom. As indicated by the upper right hand panel of Figure 2.11, the RGLM is based on two seemingly bad modifications to a GLM. As indicated by the top left panel of Figure 2.11, forward selection of a GLM is typically a bad idea since it overfits the data and thus degrades the prediction accuracy of a single GLM predictor. As indicated by the bottom right panel of Figure 2.11, bagging a full logistic regression (i.e. without variable selection) is also a bad idea since it leads to a complicated (ensemble) predictor without clear evidence for increased accuracy (see related articles by [133–138]). But these two seemingly bad modifications add up to a superior prediction method. Breiman already noted that the instability afforded by variable selection is important for constructing a bagged linear model based predictor [83]. In order to define an accurate GLM based ensemble predictor, we also find that it is important to introduce additional elements of randomness and

instability, which is also reflected in the name *random* GLM. Our results show that the proposed changes (allowing for interaction terms, forward variable selection using AIC, restricting the number of features per bag and the number of candidate features) results in a more accurate predictor that involves surprisingly few features (especially when thinning is used).

Additional reasons why the merits of RGLM have not been recognized earlier may be the following. First, it may be a historical accident. Bagging was quickly over-shadowed by other seemingly more accurate ways of constructing ensemble predictors, such as boosting [141] and the RF [86], both of which have markedly better performance on the UCI benchmark data. We find that RGLM.inter2 ties with SVM and RF for the top spot in UCI benchmark data set (Table 2.5). Incidentally, RGLM performs significantly better than SVM and RF on the disease data sets (Table 2.4) and in the 700 gene expression comparisons (Figure 2.2).

Second, previous comparisons of bagged predictors in the context of genomic data were based on limited empirical evaluations. Many comparisons involved fewer than 20 microarray data sets when comparing predictors [77, 78]. While the comparisons involved clinically important data sets from cancer applications, these studies were simply not comprehensive enough.

Third, previous studies probably did not consider enough bootstrap samples (bags). While previous studies used 10 to 50 bags, we always used 100 bags when constructing the RGLM. To illustrate how prediction accuracy depends on the number of bags, we evaluate the brain cancer data with 1 to 500 bags using 5 gene traits randomly selected from those used in our binary and continuous outcome prediction, respectively. The results are shown in Figure 2.12. Most improvement is gained in the first several dozens of bags. 100 bags is generally enough although fluctuations remain. More bags may lead to slightly better predictions but at the expense of longer computation time.

B. Strengths and limitations

RGLM shares many advantages of bagged predictors including a nearly unbiased estimate of the prediction accuracy (the out-of-bag estimate) and several variable importance measures. While our empirical studies focus on binary and continuous outcomes, it is straightforward to define RGLM for count outcomes (resulting in a random Poisson regression model) and for multi-class outcomes (resulting in a random multinomial regression model).

A noteworthy limitation of RGLM is computational complexity since the forward selection process (e.g. by the function *stepAIC* [95] from the *MASS* R package) is particularly time-consuming. The total time depends on the number of candidate features, the order of interaction terms, and the number of bags. Our R software implementation allows the user to specify three-way or higher order interaction terms but it is not clear how much they add beyond pairwise interaction terms. Our R implementation allows the user to use parallel processing for speeding up the calculations.

Our empirical studies demonstrate that RGLM compares favorably with the random forest, support vector machines, penalized regression models, and many other widely used prediction methods. As a caveat, we mention that we chose default parameter choices for each of these methods in order to ensure a fair comparison. Future studies could evaluate how these prediction methods compare when resampling schemes (e.g. cross validation) are used to inform parameter choices. Our *randomGLM* R package will allow the reader to carefully evaluate the method.

List of abbreviations

AIC: Akaike information criteria.

aMV: adjusted majority vote.

CV: cross validation.

DLDA: diagonal linear discriminant analysis.

E-RFE: entropy-based recursive feature elimination.

forwardGLM: forward selected generalized linear model.

GLM: generalized linear model.

KNN: K nearest neighbor.

LDA: linear discriminant analysis.

RF: random forest with default mtry.

RFbigmtry: random forest with mtry equal to the total number of features.

RGLM: random generalized linear model.

RGLM.inter2: RGLM considering pairwise interactions between features.

RGLM.inter3: RGLM considering two-way and three-way interactions between features.

RKNN: random K nearest neighbor.

RMNL: random multinomial logit model.

Rpart: recursive partitioning.

RSM: random subspace method.

SC: shrunken centroids.

SVM: support vector machine.

CHAPTER 2 FIGURES

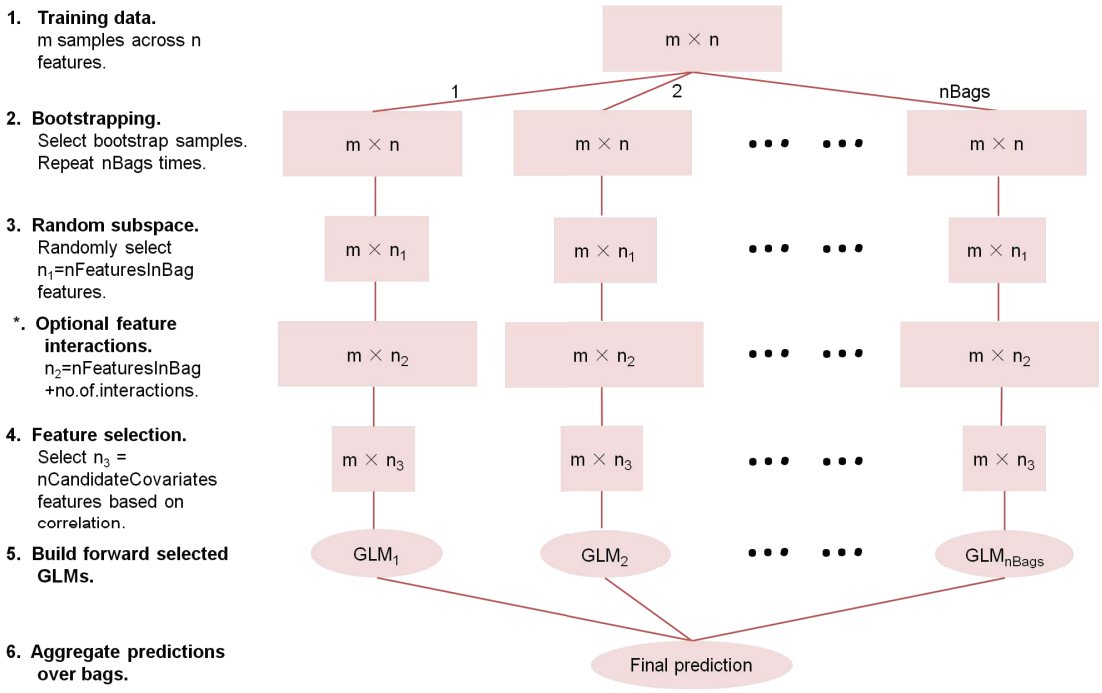

Figure 2.1: **Overview of the RGLM construction.** The figure outlines the steps used in the construction of the RGLM. The pink rectangles represent data matrices at each step. Width of a rectangle reflects the number of remaining features.

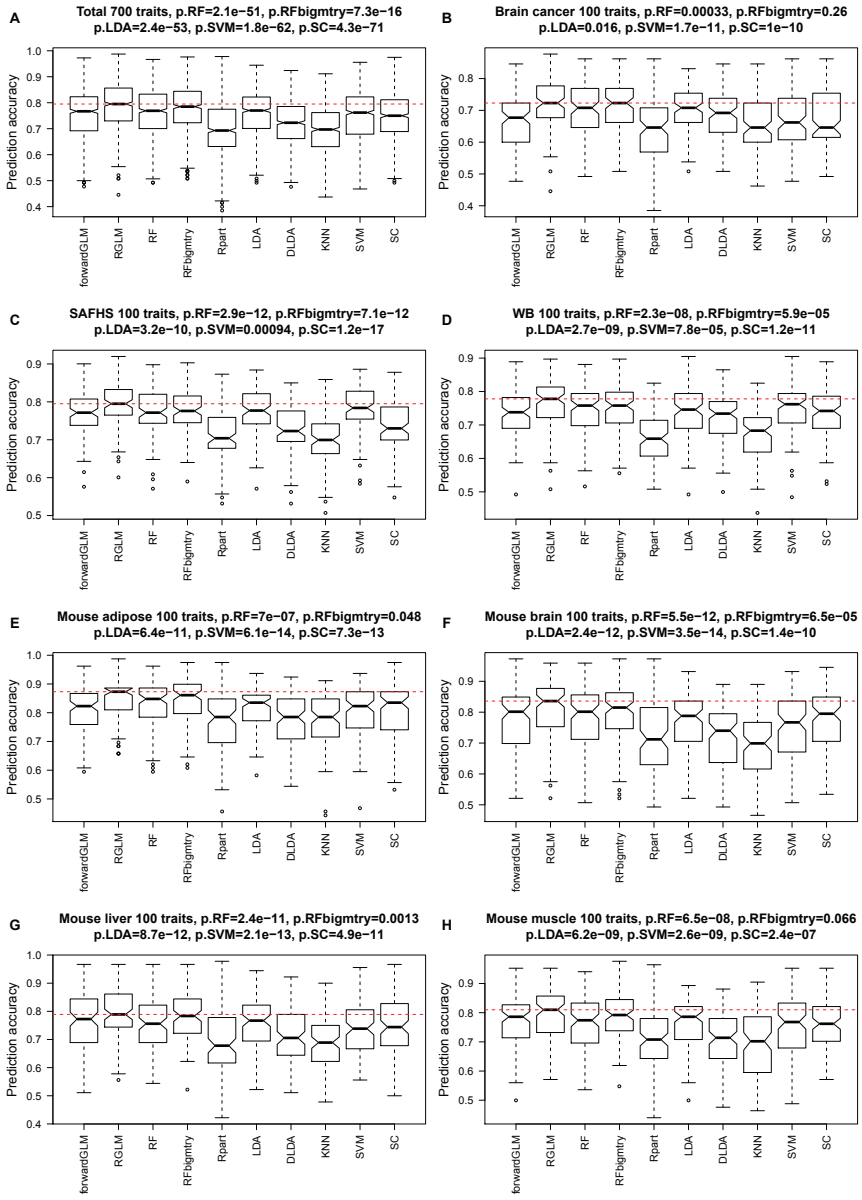

Figure 2.2: **Binary outcome prediction in empirical gene expression data sets.** The boxplots show the test set accuracies across 700 comparisons. The horizontal line inside each box represents the median accuracy. The horizontal dashed red line is the median accuracy of RGLM. P-values result from using the two-sided Wilcoxon signed rank test for evaluating whether the median accuracy of RGLM is the same as that of the mentioned method. For example, p.RF results from testing whether the median accuracy of RGLM is the same as that of RF. (A) summary across 7 data sets. (B-H) results for individual data sets.

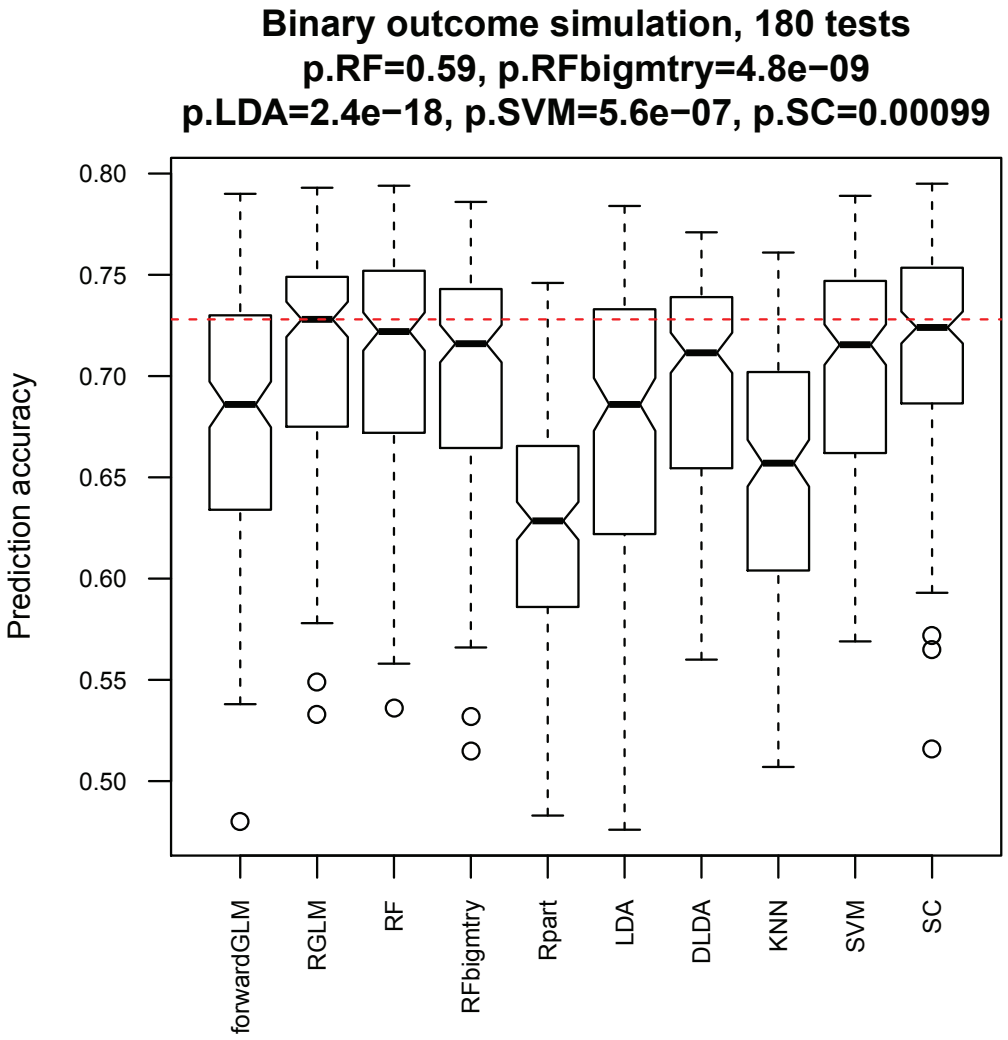

Figure 2.3: **Binary outcome prediction in simulation.** This boxplot shows the test set prediction accuracies across the 180 simulation scenarios.The red dashed line indicates the median accuracy of the RGLM. P-values result from using the two-sided Wilcoxon signed rank test for evaluating whether the median accuracy of RGLM is the same as that of the mentioned method.

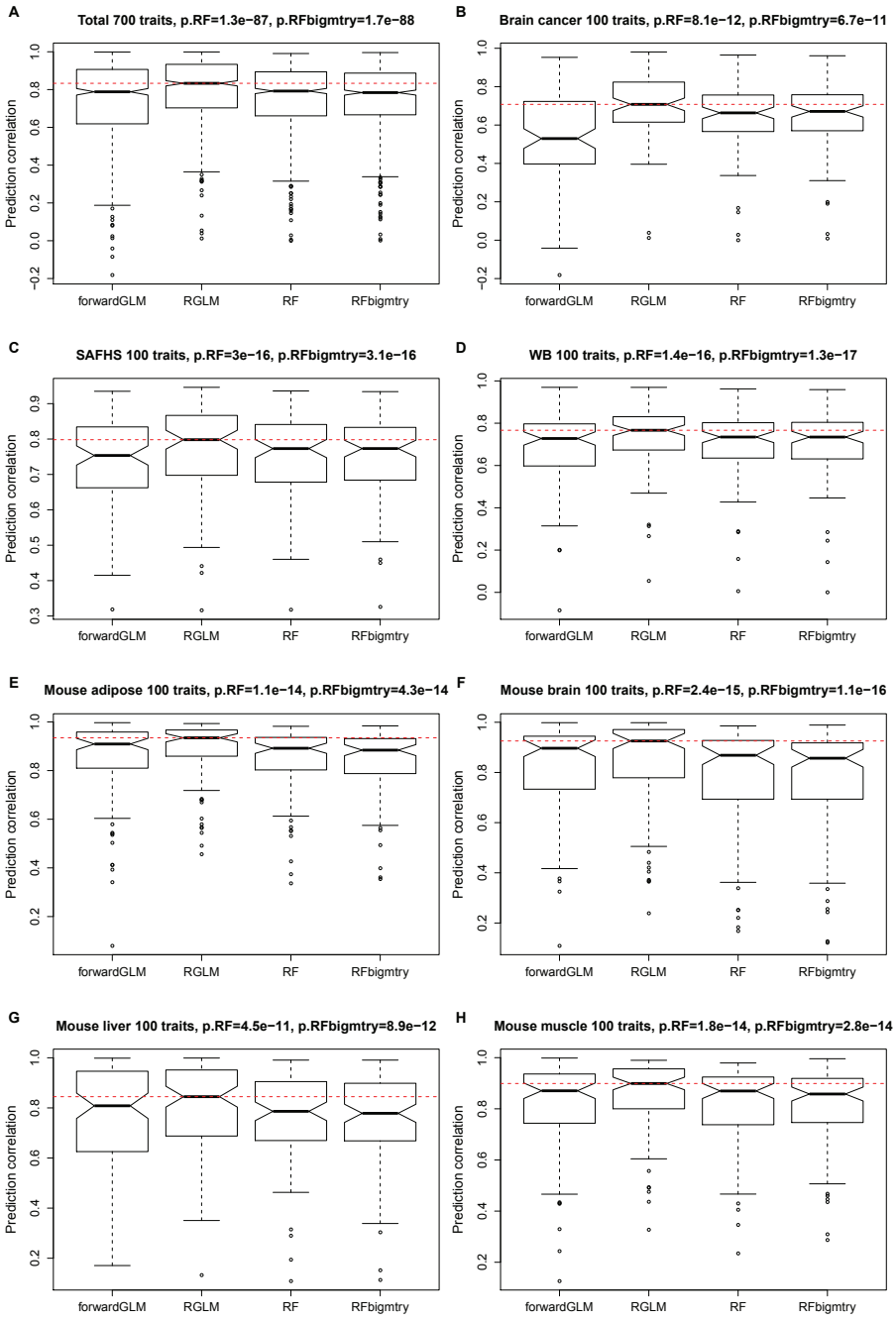

Figure 2.4: **Continuous outcome prediction in empirical gene expression data sets.** The boxplots show the test set prediction correlation in 700 applications. (A) summary of 7 expression data set. (B-H)results for individual data sets.

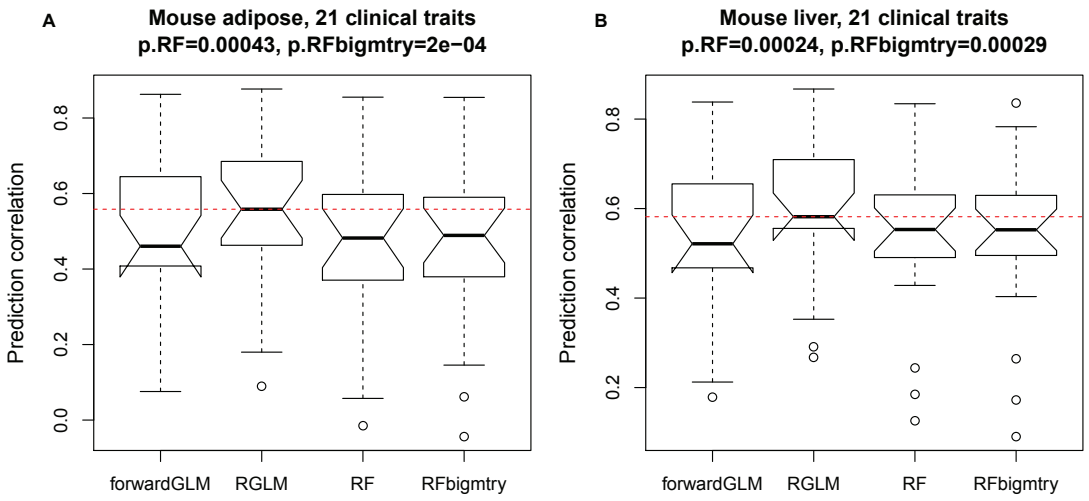

Figure 2.5: **Continuous clinical outcome prediction in mouse adipose and liver data sets.** The boxplots show the test set prediction correlation for predicting 21 clinical outcomes in (A) mouse adipose and (B) mouse liver. The red dashed line indicates the median correlation for RGLM. P-values result from using the two-sided Wilcoxon signed rank test for evaluating whether the median accuracy of RGLM is the same as that of the mentioned method.

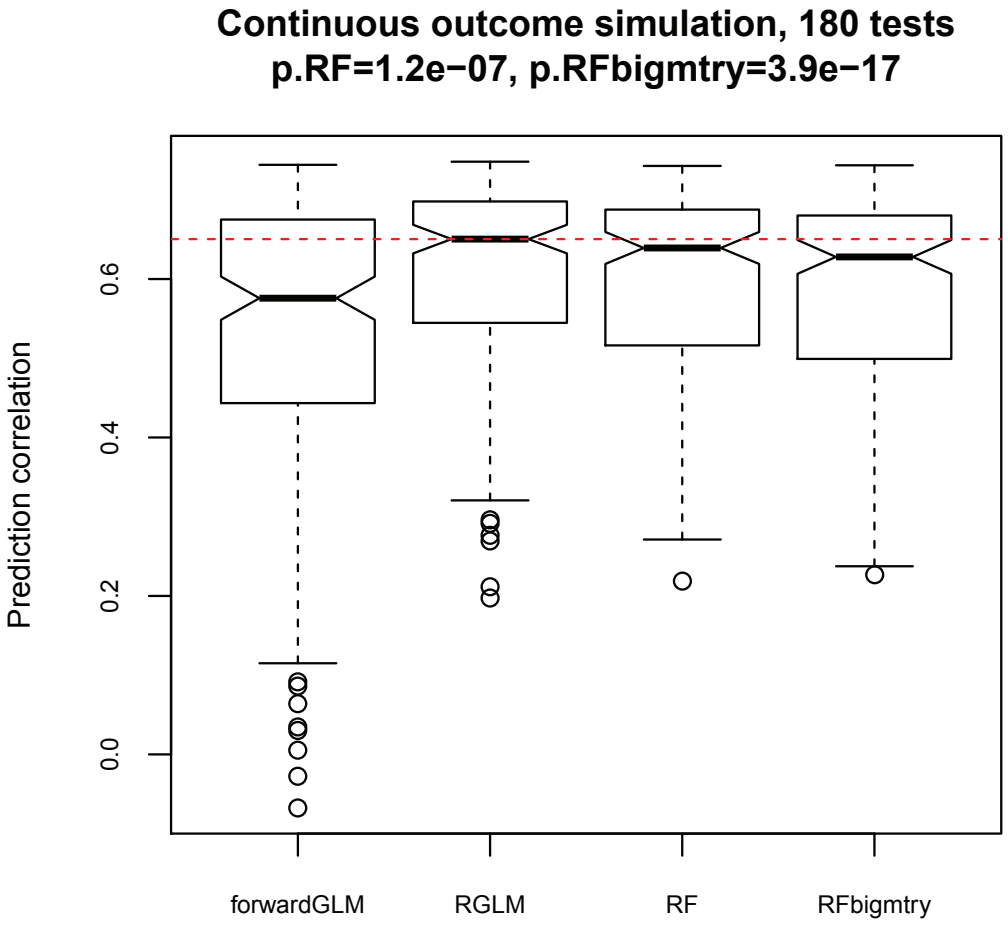

Figure 2.6: **Continuous outcome prediction in simulation studies.** This boxplot shows the test set prediction accuracy across the 180 simulation scenarios. The red dashed line indicates the median accuracy for the RGLM. Wilcoxon signed rank test p-values are presented.

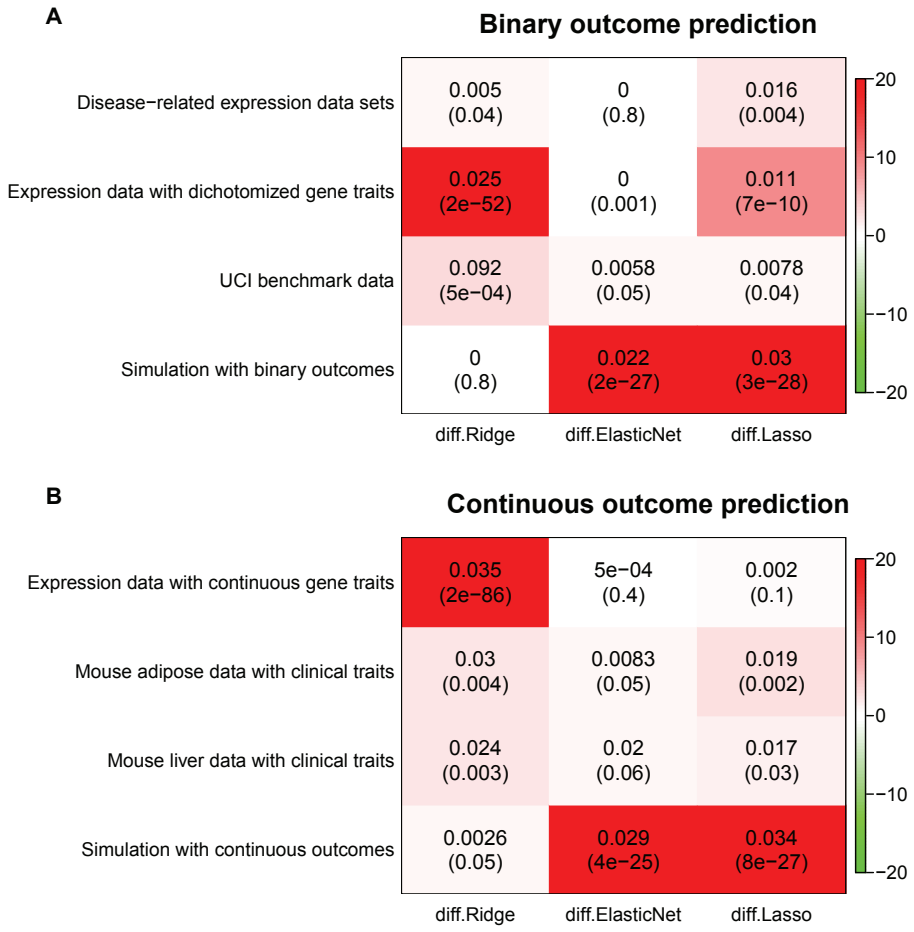

Figure 2.7: **Penalized regression models versus RGLM.** The heatmap reports the median difference in accuracy between RGLM and 3 types of penalized regression models in (A) binary outcome prediction and (B) continuous outcome prediction. Each cell entry reports the paired median difference in accuracy (upper number) and the Wilcoxon signed rank test p-value (lower number). The cell color indicates the significance of the finding, where red implies that RGLM outperforms penalized regression model and green implies the opposite. The color panel on the right side shows how colors correspond to $-log10$(p-values). $diff.Ridge = median(RGLM.accuracy - RidgeRegression.accuracy)$. $diff.ElasticNet = median(RGLM.accuracy - ElasticNet.accuracy)$. $diff.Lasso = median(RGLM.accuracy - Lasso.accuracy)$.

Figure 2.8: **Relationship between variable importance measures based on the Pearson correlation across 70 tests.** This figure shows the hierarchical cluster tree (dendrogram) of 7 variable importance measures. $absPearsonCor$ is the absolute Pearson correlation between each gene and the dichotomous trait. $KruskalWallis$ stands for the $-log10$ p-value of the Kruskal-Wallis group comparison test (which evaluates whether the gene is differentially expressed between the two groups defined by the binary trait). $RFdecreasedAccuracy$ and $RFdecreasedPurity$ are variable importance measures of the RF. $timesSelectedAsCandidates, timesSelectedByForwardRegression$ and $sumAbsCoefByForwardRegression$ are RGLM measures. These measures are evaluated in 10 tests from each of the 7 empirical expression data sets. In every test, different measures independently score genes for their relationship with a specific dichotomized gene trait. A Pearson correlation matrix was calculated by correlating the scores of different variable importance methods. Matrices across the 70 tests were averaged and the result was transformed to a dissimilarity measure that was subsequently used as input of hierarchical clustering.

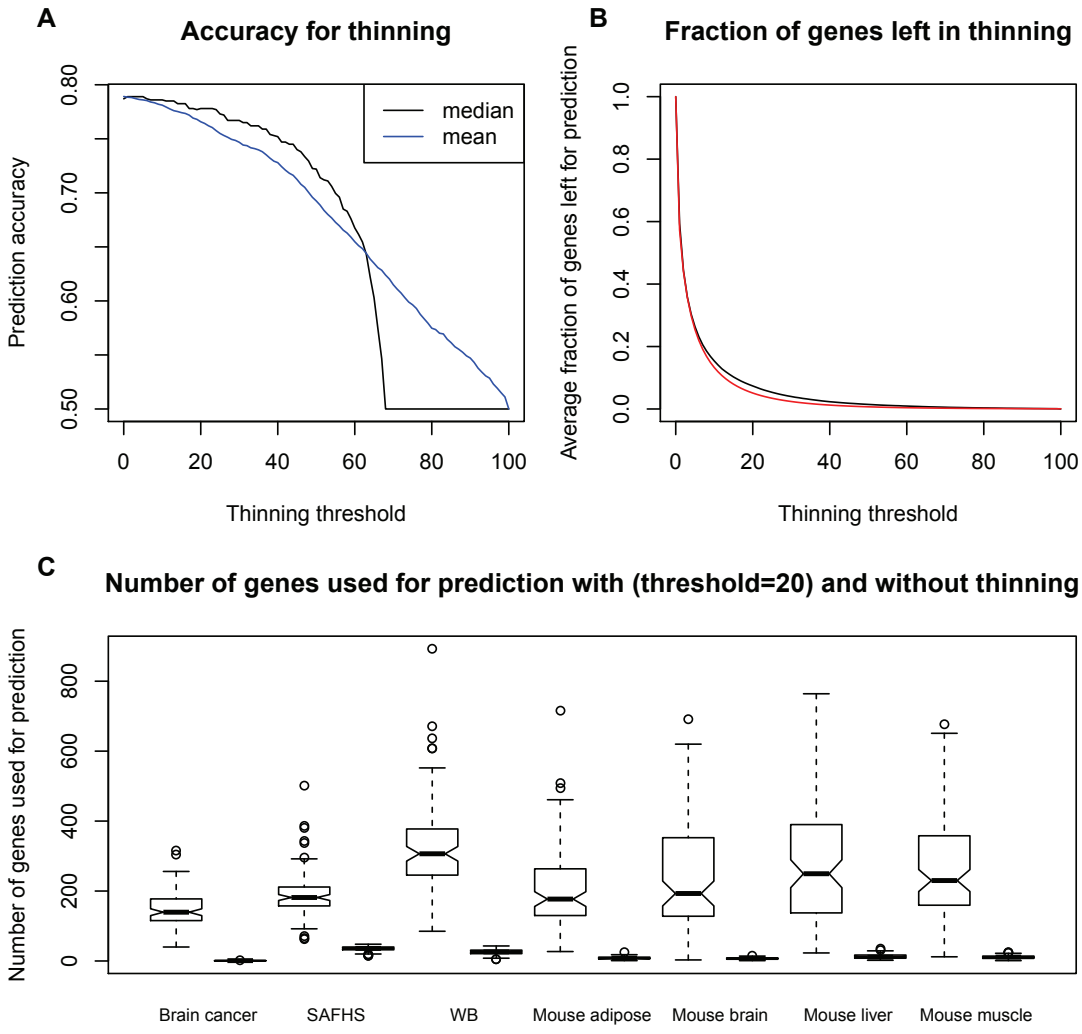

Figure 2.9: **RGLM predictor thinning.** This figure averages the thinning results of 700 applications (predicting 100 gene traits from each of 7 empirical data set). (A) Accuracies decrease as the thinning threshold increases. The black and blue lines represent the median and mean accuracies, respectively. (B) The average fraction of genes left in final models (y-axis) drops quickly as the thinning threshold increases as shown in the black line. The function in Eq. 2.1 approximates the relationship between the two variables as shown in the red line. (C) Number of genes used in prediction for no thinning versus thinning threshold equal to 20. On average, less than 20% of genes remain.

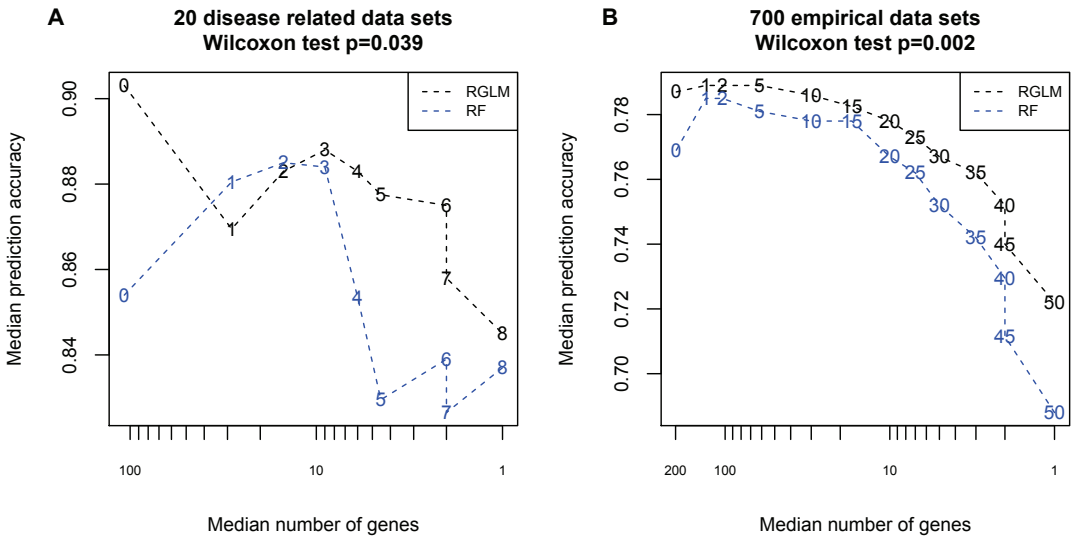

Figure 2.10: **RGLM thinning versus RF thinning.** This figure compares the thinned RGLM with the thinned RF in (A) the 20 disease related data sets and (B) the 700 gene expression traits. Numbers that connect dashed lines are RGLM thinning thresholds. For a pre-specified threshold, the number of features used for a thinned random forest is matched with that for the thinned RGLM (except for a threshold of 0). The $x-axis$ (log-scaled) and the $y-axis$ report the median number of genes left for prediction and the median accuracy across data sets, respectively. The Wilcoxon signed rank test was used to test whether the median accuracy of the thinned RGLM equals that of the thinned RF. Note that the thinned RGLM consistently yields higher accuracies than the thinned RF (according to the 2-sided test p-values).

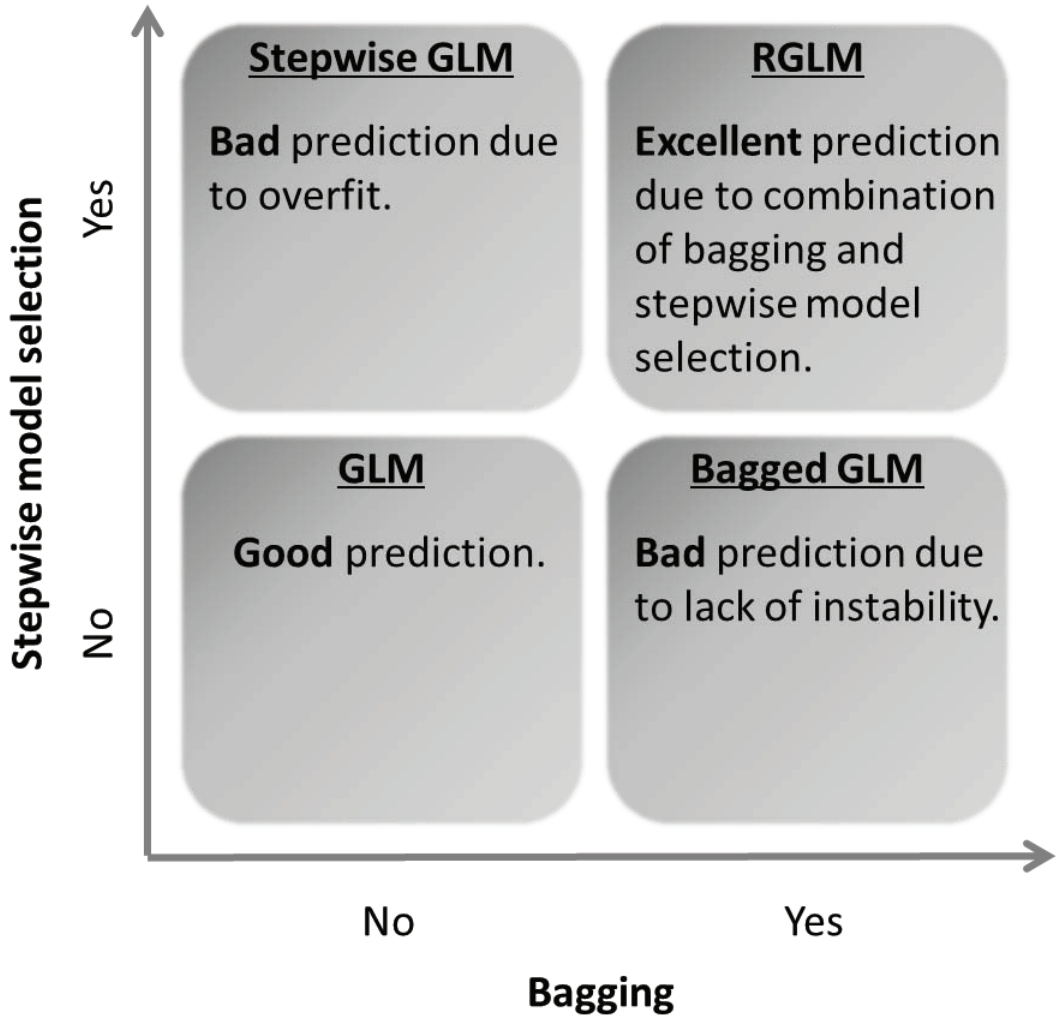

Figure 2.11: **How do modifications of a GLM affect the prediction accuracy.** The figure illustrates how two bad modifications to a GLM add up to a superior predictor (RGLM). In general, bagging or forward model selection alone lower the prediction accuracy of generalized linear models (such as logistic regression models). However, combining these two bad modifications leads to the superior prediction accuracy of the RGLM predictor. The figure may also explain why the benefits of RGLM type predictors were not previously recognized.

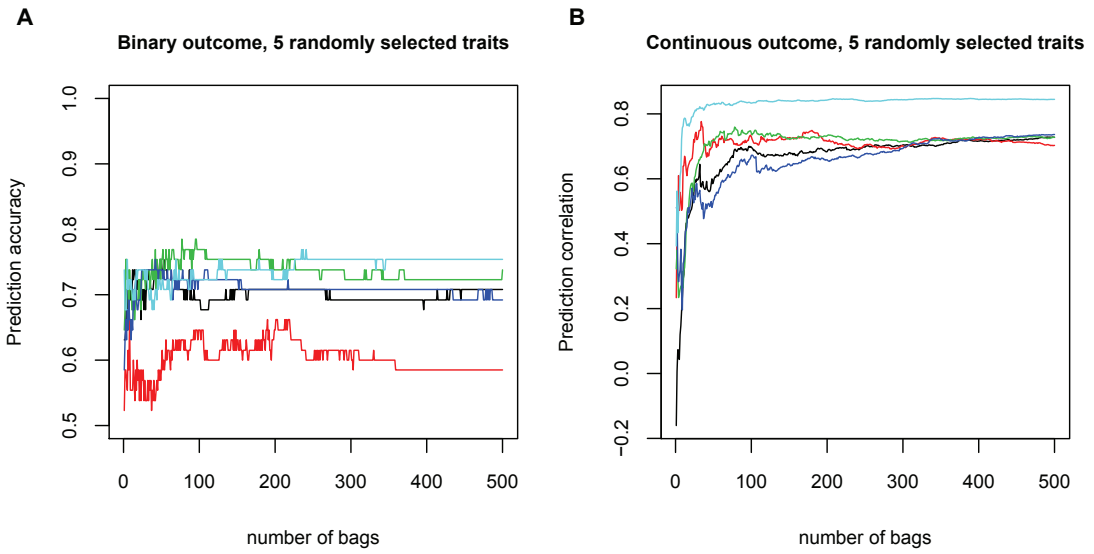

Figure 2.12: **Prediction accuracy versus number of bags used for RGLM.** This figure presents the results for predicting 5 gene traits in the brain cancer data set when different numbers of bags (bootstrap samples) are used for constructing the RGLM. Each color represents one gene trait. (A) Binary outcome prediction. The 5 gene traits were randomly selected from all 100 gene traits used in the binary outcome prediction section. (B) Continuous outcome prediction. The 5 gene traits were randomly selected from all 100 gene traits used in the continuous outcome prediction.

CHAPTER 2 TABLES

Table 2.1: **Default setting of nFeaturesInBag.**

	N	**nFeaturesInBag/N**	**N***	**nFeaturesInBag/N**
No interaction	$1-10$	1	$1-10$	1
	$11-300$	$1.0276-0.00276N$	$11-300$	$1.0276-0.00276N^*$
	>300	0.2	>300	0.2
2-way interaction	$1-4$	1	$1-10$	1
	$5-24$	$1.0276-0.00276N(N+1)/2$	$11-300$	$1.0276-0.00276N^*$
	>24	0.2	>300	0.2
3-way interaction	$1-3$	1	$1-10$	1
	$4-12$	$1.0276-0.00276(N^3+5N)/6$	$11-300$	$1.0276-0.00276N^*$
	>12	0.2	>300	0.2

This table shows the default values of $nFeaturesInBag$ in terms of $nFeaturesInBag/N$ for RGLM, RGLM.inter2 and RGLM.inter3. N is the total number of features of the training data. N^* is the effective number of features which equals the number of features N plus the number of interaction terms. Formulas are shown in terms of both N and the corresponding N^*. 1.0276 and 0.00276 are obtained by interpolating a straight line between (10,1) and (300, 0.2).

Table 2.2: **Description of the 20 disease expression data sets.**

Data set	Samples	Features	Reference	Data set ID	Binary outcome
adenocarcinoma	76	9868	[105]	NA	most prevalent class vs others
brain	42	5597	[106]	NA	most prevalent class vs others
breast2	77	4869	[107]	NA	most prevalent class vs others
breast3	95	4869	[107]	NA	most prevalent class vs others
colon	62	2000	[108]	NA	most prevalent class vs others
leukemia	38	3051	[109]	NA	most prevalent class vs others
lymphoma	62	4026	[110]	NA	most prevalent class vs others
NCI60	61	5244	[111]	NA	most prevalent class vs others
prostate	102	6033	[112]	NA	most prevalent class vs others
srbct	63	2308	[113]	NA	most prevalent class vs others
BrainTumor2	50	10367	[114]	NA	Anaplastic oligodendrogliomas vs Glioblastomas
DLBCL	77	5469	[115]	NA	follicular lymphoma vs diffuse large B-cell lymphoma
lung1	58	10000	[116]	GSE10245	Adenocarcinoma vs Squamous cell carcinoma
lung2	46	10000	[117]	GSE18842	Adenocarcinoma vs Squamous cell carcinoma
lung3	71	10000	[118]	GSE2109	Adenocarcinoma vs Squamous cell carcinoma
psoriasis1	180	10000	[119, 120]	GSE13355	lesional vs healthy skin
psoriasis2	82	10000	[121]	GSE14905	lesional vs healthy skin
MSstage	26	10000	[122]	E-MTAB-69	relapsing vs remitting RRMS
MSdiagnosis1	27	10000	[123]	GSE21942	RRMS vs healthy control
MSdiagnosis2	44	10000	[122]	E-MTAB-69	RRMS vs healthy control

Sample size, number of features, original reference, data set IDs and outcomes for the 20 disease related gene expression data sets.

Table 2.3: **Description of the UCI benchmark data.**

Data set	Samples	Features
BreastCancer	699	9
HouseVotes84	435	16
Ionosphere	351	34
diabetes	768	8
Sonar	208	60
ringnorm	300	20
threenorm	300	20
twonorm	300	20
Glass	214	9
Satellite	6435	36
Vehicle	846	18
Vowel	990	10

Sample size and number of features for the 12 UCI machine learning benchmark data sets.

Table 2.4: **Prediction accuracy in the 20 disease gene expression data sets.**

Data set	RGLM	RF	RFbigmtry	Rpart	LDA	DLDA	KNN	SVM	SC
adenocarcinoma	0.842	0.842	0.842	0.737	0.842	0.744	0.842	0.842	0.803
brain	0.881	0.810	0.833	0.762	0.810	0.929	0.881	0.786	0.929
breast2	0.623	0.610	0.636	0.584	0.610	0.636	0.584	0.558	0.636
breast3	0.705	0.695	0.716	0.611	0.695	0.705	0.669	0.674	0.700
colon	0.855	0.823	0.823	0.726	0.855	0.839	0.774	0.774	0.871
leukemia	0.921	0.895	0.921	0.816	0.868	0.974	0.974	0.763	0.974
lymphoma	0.968	1.000	1.000	0.903	0.960	0.984	0.984	1.000	0.984
NCI60	0.902	0.869	0.869	0.738	0.885	0.902	0.852	0.869	0.918
prostate	0.931	0.892	0.902	0.853	0.873	0.627	0.804	0.853	0.912
srbct	1.000	0.944	0.984	0.921	0.857	0.905	0.952	0.873	1.000
BrainTumor2	0.760	0.750	0.740	0.620	0.760	0.700	0.700	0.660	0.720
DLBCL	0.909	0.851	0.883	0.831	0.922	0.779	0.870	0.792	0.857
lung1	0.931	0.931	0.931	0.828	0.914	0.931	0.931	0.897	0.914
lung2	0.935	0.935	0.935	0.826	0.957	0.978	0.935	0.848	0.978
lung3	0.901	0.901	0.887	0.803	0.873	0.859	0.831	0.859	0.887
psoriasis1	0.989	0.994	0.989	0.978	0.994	0.989	0.989	0.983	0.989
psoriasis2	0.963	0.988	0.976	0.963	0.976	0.963	0.963	0.963	0.963
MSstage1	0.846	0.846	0.846	0.423	0.769	0.769	0.808	0.769	0.769
MSdiagnosis1	0.963	0.926	0.926	0.556	0.889	0.889	0.963	0.926	0.926
MSdiagnosis2	0.591	0.614	0.614	0.568	0.545	0.568	0.568	0.568	0.523
MeanAccuracy	0.871	0.856	0.863	0.752	0.843	0.833	0.844	0.813	0.863
Rank	1	4	2.5	9	6	7	5	8	2.5
Pvalue	NA	0.029	0.079	0.00014	0.0075	0.05	0.014	0.00042	0.37

For each data set, the prediction accuracy was estimated using $3 - fold$ cross validation across 100 random partitions of the data into 3 folds. Mean accuracies across the 20 data sets and the resulting ranks are summarized at the bottom. Two sided paired Wilcoxon test p-values can be used to determine whether the accuracy of RGLM is significantly different from that of other predictors. Note that the RGLM yields the highest mean accuracy.

Table 2.5: **Prediction accuracy in the UCI machine learning benchmark data.**

Data set	RGLM	RGLM.inter2	RF	RFbigmtry	Rpart	LDA	DLDA	KNN	SVM	SC
BreastCancer	0.964	0.959	0.969	0.961	0.941	0.957	0.959	0.966	0.967	0.956
HouseVotes84	0.961	0.963	0.958	0.954	0.954	0.951	0.914	0.924	0.958	0.938
Ionosphere	0.883	0.946	0.932	0.917	0.875	0.863	0.809	0.849	0.940	0.829
diabetes	0.768	0.759	0.759	0.754	0.741	0.768	0.732	0.740	0.757	0.743
Sonar	0.769	0.837	0.817	0.788	0.707	0.726	0.697	0.812	0.822	0.726
ringnorm	0.577	0.973	0.940	0.910	0.770	0.567	0.570	0.590	0.977	0.535
threenorm	0.803	0.827	0.807	0.777	0.653	0.817	0.825	0.815	0.853	0.817
twonorm	0.937	0.953	0.947	0.920	0.733	0.957	0.960	0.947	0.953	0.960
Glass	0.636	0.743	0.827	0.799	0.729	0.659	0.531	0.808	0.748	0.645
Satellite	0.986	0.987	0.988	0.985	0.961	0.985	0.734	0.990	0.988	0.803
Vehicle	0.965	0.986	0.986	0.973	0.944	0.967	0.729	0.909	0.974	0.752
Vowel	0.936	0.986	0.983	0.976	0.950	0.938	0.853	0.999	0.991	0.909
MeanAccuracy	0.849	0.910	0.909	0.893	0.830	0.846	0.776	0.862	0.911	0.801
Rank	6	2	2	4	8	7	10	5	2	9
Pvalue	0.0093	NA	0.26	0.042	0.00049	0.0093	0.0067	0.11	0.96	0.0015

For each data set, the prediction accuracy was estimated using $3-fold$ cross validation across 100 random partitions of the data into 3 folds. RGLM.inter2 incorporates pairwise interaction between features into the RGLM predictor. Mean accuracies and the resulting ranks are summarized at the bottom. The Wilcoxon signed rank test was used to test whether accuracy differences between RGLM.inter2 and other predictors are significant. SVM yields the highest mean accuracy.

Table 2.6: **Prediction accuracy when including pairwise interactions between features in the UCI machine learning benchmark data.**

Data set	RF	RFbigmtry	Rpart	LDA	DLDA	KNN	SVM	SC
BreastCancer	0.969	0.964	0.954	0.963	0.940	0.963	0.969	0.949
HouseVotes84	0.958	0.951	0.944	0.903	0.914	0.924	0.961	0.944
Ionosphere	0.932	0.923	0.880	0.707	0.764	0.832	0.929	0.852
diabetes	0.757	0.750	0.732	0.760	0.736	0.728	0.755	0.751
Sonar	0.812	0.830	0.760	0.755	0.686	0.812	0.812	0.707
ringnorm	0.953	0.922	0.760	0.753	0.873	0.557	0.970	0.943
threenorm	0.777	0.753	0.612	0.620	0.853	0.665	0.807	0.847
twonorm	0.937	0.897	0.727	0.717	0.933	0.732	0.923	0.953
Glass	0.813	0.790	0.732	0.710	0.559	0.788	0.734	0.617
Satellite	0.988	0.986	0.962	0.988	0.708	0.990	0.988	0.762
Vehicle	0.984	0.978	0.946	0.984	0.771	0.892	0.969	0.779
Vowel	0.991	0.986	0.949	0.982	0.899	1.000	0.997	0.912
MeanAccuracy	0.906	0.894	0.830	0.820	0.803	0.823	0.901	0.835

The table shows the prediction accuracy of predictors other than RGLM when considering pairwise interactions between features in the same UCI mlbench data sets. Although several predictors show improvement, none of them beats RGLM.inter2.

CHAPTER 3

Predicting COPD status with the Random Generalized Linear Model

Introduction

Sample classification, especially disease status prediction, is an important area of investigation for gene expression studies. While the high number of features measured by microarray expression data promises to result in accurate multivariate predictors of disease outcomes [74], relatively few genomic predictors have been found to be useful in clinical practice. Many machine learning methods have been used for predictor construction, such as Random Forest (RF) [86], Support Vector Machine [75], K-Nearest Neighbor [95, 96] and Linear Discriminant Analysis [95, 96].

We recently developed a highly accurate and interpretable predictor: the Random Generalized Linear Model (RGLM) predictor [142]. This ensemble predictor is based on bootstrap aggregation (bagging) of generalized linear models whose features (covariates) are selected using a random subspace method coupled with forward regression. RGLM has shown high prediction accuracy in many gene expression applications [142].

Although we and others have evaluated predictors in real applications [77, 78], the evaluations are often unrealistic in the sense that training and test data are subsets from the same original data set, making them very comparable to each other. In practice, however, a test set is usually composed of new samples that have not measured using exactly the same methods or equipment as the available training samples. The 2012 Improver Challenge [143] provides a much more realistic scenario for evaluating prediction methods in that training and test data are generated by different groups/labs. Participants in this data analysis contest aim to predict the disease status of a test set of de-identified samples using any publicly available training data of their choice. In this article, we focus on one of the sub-challenge data sets, namely the chronic obstructive pulmonary

disease (COPD) data.

COPD is a leading cause of death worldwide [144]. It is estimated to affect about $9 - 10\%$ of adults aged ≥ 40 years of age [144]. COPD causes irreversible damage to the lungs, with the airways becoming narrower over time. A major cause of the disease is cigarette smoking [145]. In clinical practice, no single symptom or sign can adequately confirm or exclude the diagnosis of COPD [146] and no widely recognized molecular biomarker exists to date. If feasible, gene expression based molecular predictors would greatly facilitate the early detection and diagnosis of COPD. This article is organized as follows. We first describe the data used for prediction and the corresponding pre-processing procedures in the Materials and Methods section. Next, we compare the training set out-of-bag (OOB) prediction performances of RGLM and RF. We then evaluate the test set prediction and identify gene signatures. We further test which genes, if any, are indispensable for accurate prediction. Finally, we discuss the potential reasons for the superior performance of RGLM in the COPD sub-challenge.

Results

The data pre-processing steps are detailed in the Materials and Methods. Demographic data for 235 training set samples and 40 test set samples included in this analysis after pre-processing are shown in Table 3.1. The training set consisted of 26 COPD cases and 209 healthy controls. Samples were combined from different GEO data sets to increase sample size and thus increase power. Within training set samples, cases were on average almost 10 years older than controls (51.7 compared with 41.5); the percentage of males was higher in cases (80.8%) than in controls (67.0%); additionally, the percentage of smokers was much higher in cases (100%) than in controls (57.4%). In addition, all training set cases were sampled from the small airways of lungs, while all test set individuals were sampled from the large airways. Small and large airways may contain slightly different cell types at different abundances, increasing the difficulty for prediction.

We considered the RGLM predictor and the RF predictor for classification. Since both methods are ensemble predictors, they naturally provide training set OOB predictions that can be used to evaluate prediction performance. We therefore compared the OOB predictive probability of being affected for true cases and true controls in the training set (Figure 3.1). The Kruskal-Wallis test p-values on top of each panel can be used to quantify to what extent each ensemble predictor can distinguish true cases from controls. As shown in Figure 3.1 (A-B), RGLM ($p = 7.4 \times 10^{-10}$) performed much better than RF ($p = 9.3 \times 10^{-7}$) for COPD classification in the training set. To further increase the predictive accuracy of RGLM, we chose ("tuned") parameter values. A noteworthy parameter of RGLM is called "mandatoryCovariates" since it allows one to force specified features into the prediction models for all bags (detailed in Materials and Methods). This is particularly beneficial when certain features are known to be associated with the outcome. For example, smoking status and age are known to be associated

with COPD status [145, 147, 148] and we confirmed this in our training data (Table 3.1). When including smoking status and age as mandatory covariates into RGLM, the predictive performance was further improved ($p = 2.3 \times 10^{-11}$) as indicated in Figure 3.1 (C). There are also important discussions in the literature on gender differences in susceptibility to smoking effects and lung function reduction in COPD [149]. When including gender as an extra mandatory covariate into RGLM, we observed no further performance improvement ($p = 3.8 \times 10^{-11}$, Figure 3.1 (D)). A plausible explanation is that gender may indirectly affect COPD status through smoking: men smoke more than women, therefore an association between smoking and COPD would lead to an association between gender and COPD (Table 3.1). As aside, we mention that gender is not recognized as a COPD risk factor according to the GOLD (Global Initiative for Chronic Obstructive Lung Disease) document for COPD [148]. Based on these results, we chose to use the RGLM predictor with smoking status and age as mandatory covariates for test set predictions on the Improver challenge data set.

Figure 3.1 also reveals that all predictors were poorly calibrated. Most samples, even true COPD cases, were assigned very low probabilities of being affected. This poor calibration reflects the fact that the training data were highly unbalanced: 26 COPD cases versus 209 controls. As a result, most true cases would be classified as controls if the predictive probabilities are thresholded at 0.5. To solve this issue in test set classification, we re-calibrated the test set predictive class probabilities assuming half of the test set smokers were COPD cases (as detailed in the Materials and Methods section).

After the Improver Challenge organizers released the true COPD status of the test set data, we were able to assess the test set accuracy of the RGLM predictor. The test set contained 24 true COPD patients and 16 healthy controls. The RGLM predictor resulted in an accuracy of 0.825, which was much higher than

that of a naive predictor (0.600) obtained by assigning all test samples to the most prevalent class (Table 3.2). The RGLM predictor had optimal specificity (estimated to be 1) but sub-optimal sensitivity (0.708). The receiver operating curve (ROC) is shown in Figure 3.2 with an impressive area under the ROC curve of 0.939. Overall, the RGLM predictor performed best among all methods proposed by the different teams who participated in the COPD sub-challenge.

Variable importance measures provided by the RGLM predictor allowed us to identify 355 genes (421 probes) relevant for COPD classification in addition to the mandatory covariates smoking status and age. As detailed in the Materials and Methods section, these probes served as covariates (dependent variables) in the forward selected GLMs. Most of these probes were selected only once or twice across the 100 bags. The top eight most frequently selected "signature" genes (selected at least five times across the 100 bags) are listed in Table 3.3. Interestingly, four out of these eight genes are involved in transcriptional regulation. Although not statistically significant, this is consistent with a previous finding that COPD biomarker genes were enriched for functions related to transcriptional regulation [150].

In the original RGLM article [142], we introduced "RGLM predictor thinning", where we constructed a sparser but still highly accurate predictor by refitting the GLM in each bag only considering frequently selected features. This method can be applied here to elucidate which gene features are indispensable in preserving predictive accuracy in addition to the mandatory covariates. In Figure 3.3, we only considered gene features that were selected more times than a set "thinning threshold". Since gene features were selected up to 8 times among all bags, the thinning threshold could vary from 0 to 8. We found that the number of gene features used for modeling declined rapidly from 421 to 0 as the thinning threshold increased (Figure 3.3 (A)), but the prediction accuracies were not affected at

all (Figure 3.3 (B)). This indicates that gene features have little effect on the prediction accuracy. Instead, the high accuracy is achieved via the mandatory variables (smoking status, age). While the signature genes listed in Table 3.3 are not needed for predicting COPD status when the mandatory variables are also included in the model, we list them nevertheless since they may be of interest to biologists.

Discussion

In this article, we successfully employed the recently developed RGLM predictor to predict the COPD status of 40 de-identified individuals whose demographic information and gene expression profiles were available. We pre-processed the publicly available training sets and the Improver test set together to make them comparable. Out of bag estimates of predictive accuracy in the training data led us to favor RGLM over the random forest predictor. The RGLM predictor with smoking status and age as mandatory covariates achieved the highest predictive accuracy not only in the training data but also in the Improver Challenge test data set. Variable importance measures of the RGLM narrowed down eight gene features for COPD classification, but high prediction accuracy can be retained by using only smoking status and age as features.

While RGLM was the best performing method, it is important to note that the prediction accuracy of different methods cannot be compared directly because the different Improve Challenge teams used different normalization strategies and slightly different training data. The success of the RGLM method in the COPD sub-challenge may reflect our pre-processing steps that aimed to make training and test sets as comparable as possible. Toward this end, we used the following pre-processing steps. First, we only used training sets which were measured on the same Affymetrix microarray platform as the test set because most prediction methods are vulnerable to platform differences. Second, we used raw Affymetrix CEL files as opposed to normalized data so that we could apply the same normalization method (MAS5) to all data. Here we chose MAS5 normalization to facilitate comparisons by others using their own data. And we point out that the Improver Challenge organizers found no significant difference in prediction performance between teams that used MAS5 and those who used other microarray normalization methods. Third, we combined training and test sets in our pre-processing steps to

ensure high comparability. Fourth, we used the SampleNetwork R function [151] because it implements powerful methods for finding array outliers and for carrying out quantile normalization and ComBat batch effect correction [152].

The success of the RGLM method also reflects the merits of the predictor itself. In the majority of applications, we find that RGLM is substantially better than the non-ensembled version of the GLM as shown in our recent comparative study [142]. Breiman showed already that a bagged version of an unstable learner is more accurate than the individual learner [83]. RGLM gained accuracy by aggregating unstable forward selected GLMs which tended to overfit training data. In order to quantify the advantage brought by the ensemble, we compared the test set accuracy of RGLM versus a forwardGLM (the unbagged version of RGLM, detailed in Materials and Methods) in Table 3.2. As expected, RGLM was much more accurate than the unbagged forwardGLM (0.825 vs. 0.575), which probably reflects the well known fact that forward selection greatly overfits the data [153].

As indicated by the Kruskal-Wallis test p-values in Figure 3.1, RGLM outperformed the RF, which is a highly accurate ensemble predictor as well. One may suspect that by tuning the RF parameter mtry, one can improve the performance of the RF. To some extent, this is indeed the case. The Kruskal-Wallis test p-value for RF could be improved from 9.3×10^{-7} to 6.4×10^{-8} by setting mtry to N instead of its default value $\sqrt{N}$, where N is the total number of features. Large values for mtry work well when relatively few gene features are informative for predicting COPD. In this application, RGLM outperformed the RF because of the following reasons. First, it focused on the most associated genes since it only considered the top 50 most associated features for forward variable selection in each bag. Second, the RGLM parameter "mandatoryCovariates" allowed us to force known risk factors (smoking status and age) into each individual GLM. We expect that the performance of the RF would be greatly improved if the same

mandatory covariates (smoking status, age) would be included in the construction of each tree predictor.

A major limitation of our analysis was that we did not distinguish between small and large airway origin in our study because of lack of pertinent training set data. As shown in Table 3.1, airway origin was confounded with data set and COPD status. Thus adjusting for airway origin in our analysis (e.g. by conditioning) would have removed the desired signal as well. Therefore, we could only predict large airway samples based on a predictor trained on small airway samples. Given this limitation, it is surprising, that the prediction was fairly accurate. We think this reflects that the mandatory covariates used in our study (smoking status, age) relate to both types of diseases.

A second limitation of the analysis is that we assumed that half of the smokers are COPD cases. We used this ad-hoc assumption for a technical reason: it allowed us to counter the effect caused by highly unbalanced training data. An alternative and statistically superior approach is to sub-sample training set controls so that they are balanced with respect to the number of cases. In response to a reviewer comment, we implemented this alternative approach by frequency matching using age (as detailed in Materials and Methods). As shown in Table 3.2, the test set prediction performance was not as good as that of our original predictor, perhaps because the resulting sample sizes were insufficient (sub-sampling resulted in 26 controls down from 209 controls).

The GOLD 2011 document has classified COPD patients into 4 categories according to airflow limitation severity [148]. Our training cases likely fall into several of these categories. In future studies, it would be beneficial to distinguish these disease categories when training the predictor since it would reduce the disease heterogeneity and thus increase the predictive signal.

Our study identified 355 genes (421 probes) useful for COPD prediction. Al-

though they have little effects on predictive accuracy when mandatory covariates smoking status and age are used, they may be of interest to biologists. Further examination may provide information regarding genetic determinants of COPD. The overall aim of our study was the development of a predictor of COPD status. If, instead, we wanted to learn about the biology underlying COPD, we would have used signed weighted correlation network analysis (WGCNA) because it facilitates a systems biologic module based analysis of microarray data [5].

Materials and Methods

A. Data description

The test data set contains gene expression profiling data for large airway samples of lungs from 40 de-identified individuals, measured using the Affymetrix Human Genome U133 plus 2.0 platform. As part of the data analysis challenge, we were allowed to use training data from any publicly available source. We downloaded the following raw gene expression data from the Gene Expression Omnibus (GEO) database as our training sets for COPD: GSE10006 [154], GSE10135 [155], GSE11906 [156], GSE11952 [157], GSE13933 [158], GSE19667 [159], GSE20257 [160], GSE5058 [161, 162], GSE5059 [161], GSE7832 [162] and GSE8545 [163]. Arrays present in more than one data set were used only once. All training data were generated from Affymetrix Human Genome U133 plus 2.0 array with 54675 probes. Training cases are small airway samples of lungs, while controls contain both small and large airway samples.

B. Data pre-processing

The raw training and test data were first MAS5 normalized and log2 transformed using the R package simpleaffy. Next, subject NS047 was removed from the training data because the smoking status of this individual in GSE10135 and GSE11906 data sets was contradictory. Finally, all training and test data were pooled together because quantile normalization and batch effect correction (see below) make training and test data more comparable. Our network based approach for preprocessing the data was implemented in the SampleNetwork R function [151], which has powerful methods for identifying array outliers and batch effects. This user-friendly R function sequentially carries out outlier removal, quantile normal-

ization and ComBat batch effect correction [152] in an interactive and automatic manner. No severe array outliers were found in terms of low inter-array correlations. Array batches represented by different data set IDs were corrected without controlling for other covariates. No feature selection was performed. After preprocessing, all training sets were combined into one large training set ($n = 235$ on 54675 probes). Clinical information such as age, gender, race and smoking status were available in both training and test samples.

C. Classification methods

Random forest Developed by Leo Breiman, RF is a highly accurate ensemble predictor that consists of an ensemble of individual decision trees which vote on the final outcome prediction [86]. It has a parameter mtry, i.e. the number of features considered at each node split. The default mtry value equals the square root of the number of features. Here we used the randomForest function from the randomForest R package based on the original Fortran code by Leo Breiman.

Random generalized linear model We recently developed the RGLM predictor. Similar to the RF, it is an ensemble of weak individual learners. However, different from the RF, the individual learners are generalized linear models whose features (covariates) are selected using forward regression according to AIC criteria [142]. In particular, RGLM uses individual logistic regression models for binary outcome classification. We used the randomGLM function from the randomGLM R package [142].

Briefly, RGLM for COPD classification was constructed as follows. First, 100 versions of bootstrap samples of observations (referred to as bags) were generated from the training data. Second, 20% of randomly chosen features (transcription probes) were selected for each bootstrap data set. Third, feature selection was carried out in each bag based on correlating each feature with the outcome. Only

the top 50 features with the most significant correlation would be considered as candidate covariates for forward selection in a logistic regression model. Fourth, a forward selected logistic regression model was fitted in each bag to arrive at one logistic model based prediction per bag. The forward selection procedure used by RGLM was based on the stepAIC R function in the MASS R library. Fifth, test set covariate values were used as input to the prediction models to arrive at a predicted class probability per bag. Sixth, predicted class probabilities were averaged across bags to arrive at a final test set class probability estimate. By thresholding the class probabilities we obtained the final binary outcome prediction.

RGLM has a parameter "mandatoryCovariates" that forces specified features into the prediction models of all learners (bags). As a result, it overweighs those features over others. For the COPD classification, we evaluated RGLM without mandatory covariates, RGLM with mandatory covariates age and smoking status, and RGLM with mandatory covariates age, smoking status and gender.

forwardGLM We denote by forwardGLM the (single) generalized linear model predictor whose covariates are selected using forward feature selection according to the AIC criterion. Thus, forwardGLM does not involve bagging, random feature selection, and is not an ensemble predictor. In this article, we use forwardGLM with age and smoking status as mandatory covariates.

D. Prediction re-calibration

It is difficult to choose a threshold for the predicted class probabilities since this depends on the prevalence of the disease. For simplicity, we assumed that half of the smokers in the test set had COPD while half were healthy controls. Therefore, the test set predictive probabilities were re-calibrated so that half of the test set smokers were classified as COPD cases. This could also be achieved by rescaling the original class probabilities on the log scale as follows: log(new.P) =

0.417*log(P) and choosing the usual threshold of 0.5.

E. Training controls sub-sampling

There were 26 COPD cases and 209 controls in the training data. Here, we aimed to sub-sample 26 controls to match the 26 cases by age. Frequency matching was used. Training samples were grouped into age intervals, i.e. $36-40, 41-45 ... 71-75$. In each age interval, we sub-sampled the same number of controls as the number of cases. After sub-sampling, training controls had mean age 52 and standard deviation 8.6, very similar to the statistics of training cases. Age was not used as a mandatory covariate in RGLM, since it was used in matching.

List of abbreviations

AIC: Akaike information criterion.

AUC: Area under the curve.

COPD: Chronic obstructive pulmonary disease.

GEO: Gene Expression Omnibus.

OOB: Out-of-bag.

RF: Random forest.

RGLM: Random generalized linear model.

ROC: Receiver operating characteristic.

CHAPTER 3 FIGURES

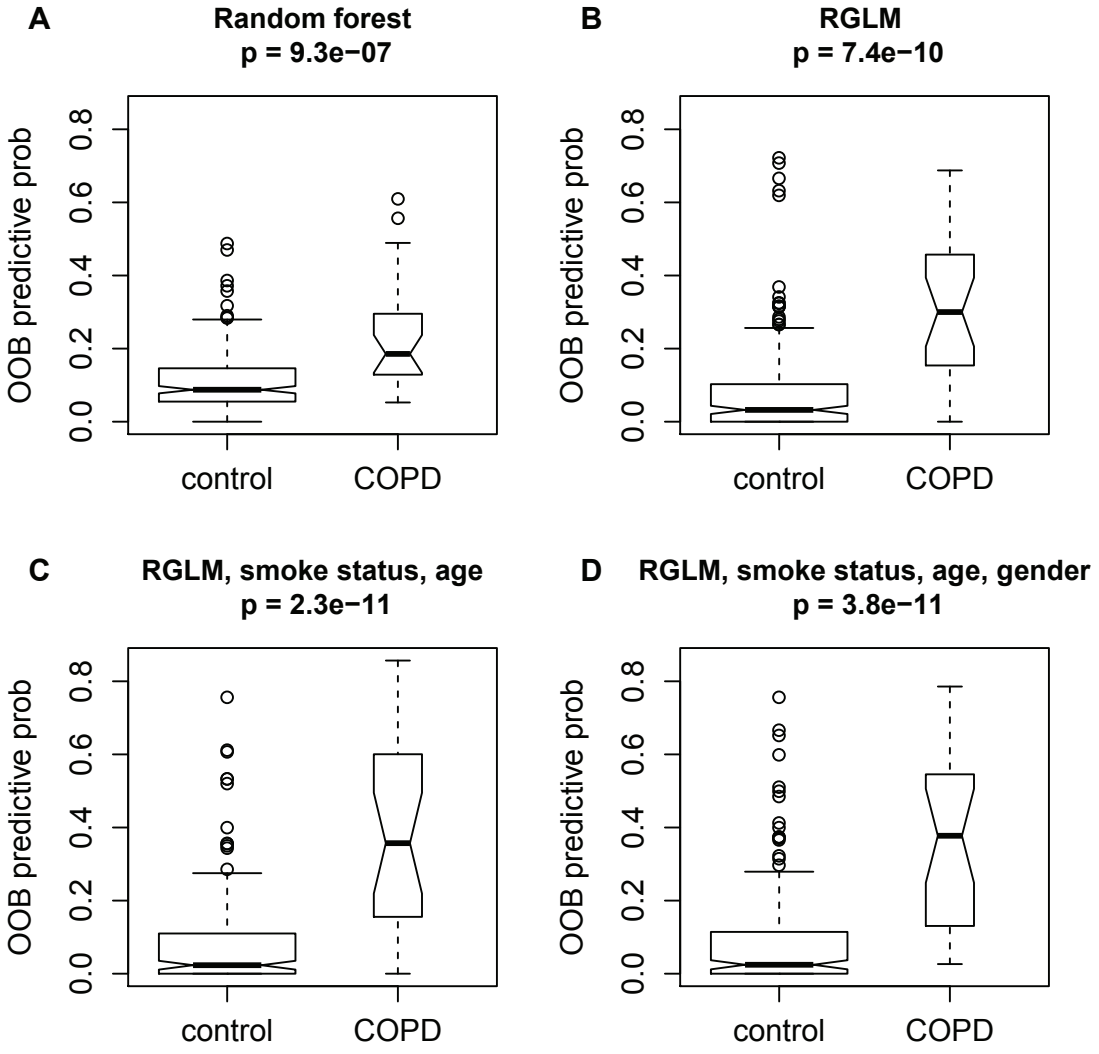

Figure 3.1: **Training set OOB prediction of 4 predictors.** The boxplots show the OOB predictive probability of being affected in the true COPD group and true control group. P-values at the top of each panel are derived from Kruskal-Wallis tests that compare the median predictive probabilities between the two groups. Four predictors are considered. (A) Random forest. (B) RGLM. (C) RGLM with smoking status and age as mandatory covariates. (D) RGLM with smoking status, age and gender as mandatory covariates.

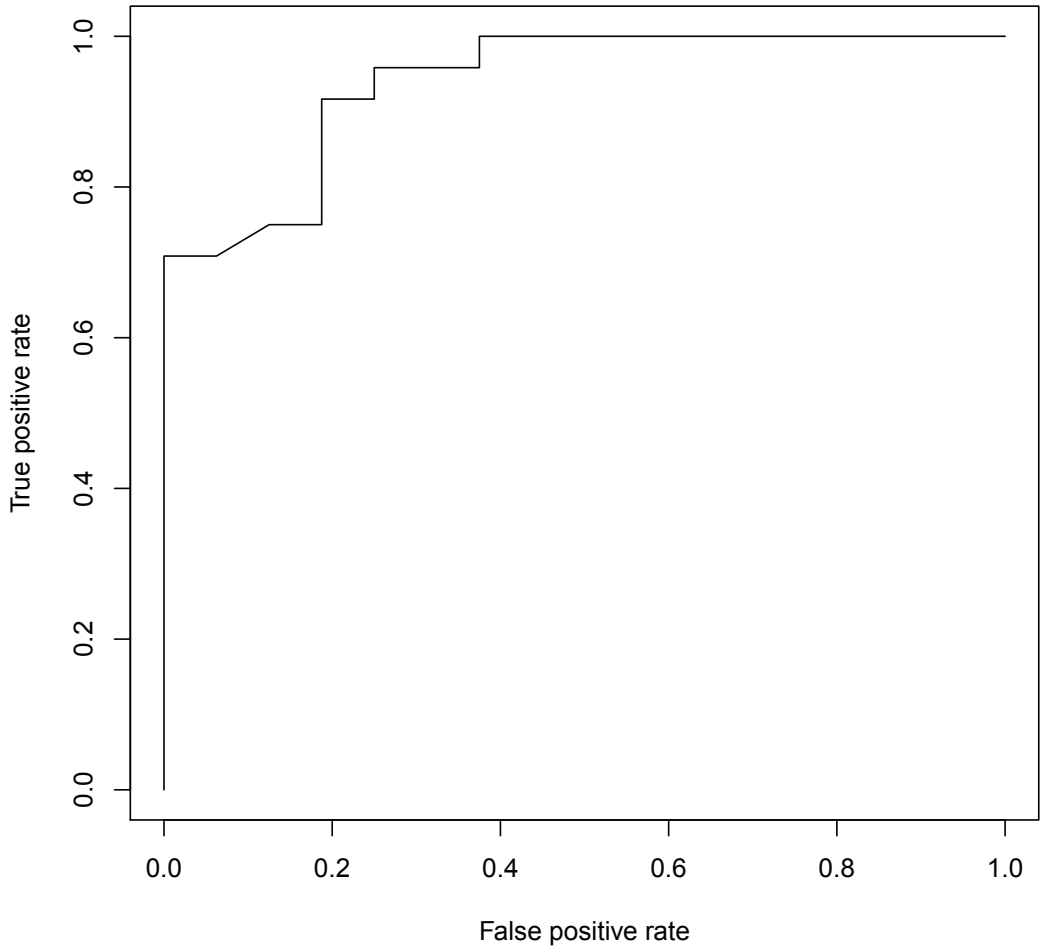

Figure 3.2: **ROC curve of test set prediction.** The predictor used is RGLM with smoking status and age as mandatory covariates.

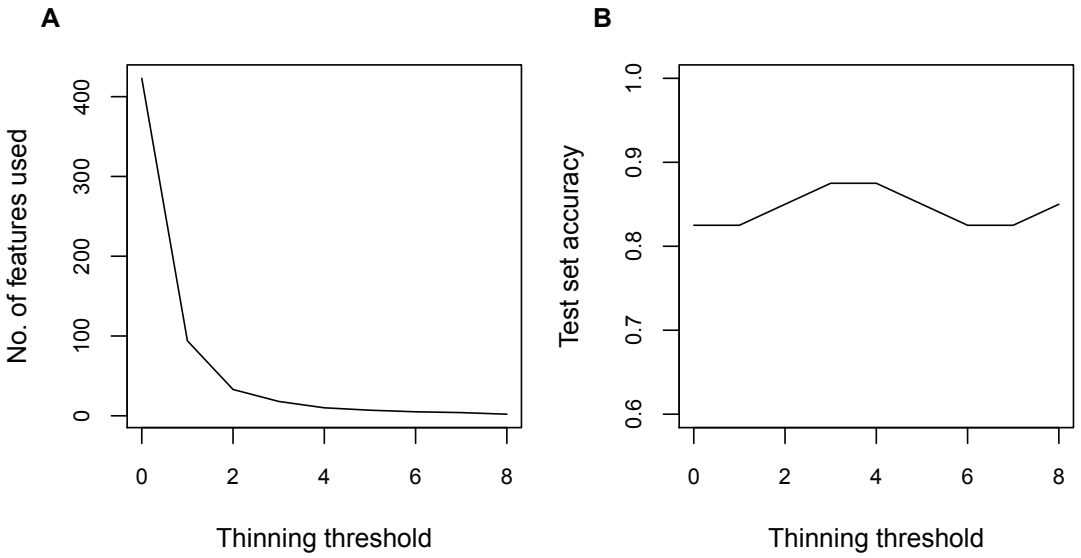

Figure 3.3: **RGLM predictor thinning.** (A) Number of features left in RGLM drops quickly as the thinning threshold increases. (B) Prediction accuracies fluctuate as the thinning threshold increases.

CHAPTER 3 TABLES

Table 3.1: **Sample statistics.**

	Training set COPD cases (n=26)	**Training set controls (n=209)**	**Test set samples (n=40)**
Age Mean (SD)	51.7 (8.1)	41.5 (9.4)	51.0 (10.3)
Gender			
Male n (%)	21 (80.8%)	140 (67.0%)	32 (80.0%)
Female n (%)	5 (19.2%)	69 (33.0%)	8 (20.0%)
Smoker			
Yes n (%)	26 (100%)	120 (57.4%)	32 (80.0%)
No n (%)	0 (0)	89 (42.6%)	8 (20.0%)
Tissue			
Small airway n (%)	26 (100%)	149 (71.3%)	0 (0)
Large airway n (%)	0 (0)	60 (28.7%)	40 (100%)

Table 3.2: **Evaluation of test set COPD classification by RGLM, forwardGLM, and RGLMsubSamp with smoking status and age as mandatory covariates.**

Predictor	Naïve Accuracy	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
RGLM	0.600	0.825	0.708	1	1	0.700	0.939
forwardGLM	0.600	0.575	0.292	1	1	0.485	0.646
RGLMsubSamp	0.600	0.700	0.750	0.625	0.750	0.625	0.700

Naïve Accuracy: accuracy achieved by assigning all test samples as COPD cases; PPV: Positive Predictive Value; NPV: Negative Predictive Value; AUC: area under the ROC curve.

Table 3.3: **Top eight signature genes.**

Times	**Probe ID**	**Gene symbol**	**Chr.**	**Description**	**Function**
8	219041_s_at	REPIN1	7	Replication initiator 1	Initiation of DNA replication
	1560328_s_at	NA	2	NA	NA
7	231126_at	C2orf70	2	Chr2 open reading frame 70	Unknown
6	228305_at	ZNF565	19	Zinc finger protein 565	Transcriptional regulation
	242758_x_at	KDM3A	2	Lysine (K)-specific demethylase 3A	Transcriptional activation
5	1553336_a_at	MIER3	5	Mesoderm induction early response 1, family member 3	Transcriptional repressor
	219517_at	ELL3	15	Elongation factor RNA polymerase II-like 3	RNA polymerase II catalysis
		SERINC4		Serine incorporator 4	Lipid synthesis
	228268_at	FMO2	1	Flavin containing monooxygenase 2	Oxidation catalysis

Only gene features that are selected at least five times in 100 bags are included.

References

[1] Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863–14868.

[2] Zhou X, Kao M, Wong W: **Transitive Functional Annotation By Shortest Path Analysis of Gene Expression Data**. *Proc Natl Acad Sci U S A* 2002, **99**(20):12783–88.

[3] Stuart JM, Segal E, Koller D, Kim SK: **A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules**. *Science* 2003, **302**(5643):249–255.

[4] Zhang B, Horvath S: **General framework for weighted gene coexpression analysis**. *Stat Appl Genet Mol Biol* 2005, **4**:17.

[5] Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.

[6] Butte A, Tamayo P, Slonim D, Golub T, Kohane I: **Discovering Functional Relationships Between RNA Expression and Chemotherapeutic Susceptibility Using Relevance Networks**. *Proc Natl Acad Sci U S A* 2000, **97**:12182–12186.

[7] Daub C, Steuer R, Selbig J, Kloska S: **Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data**. *BMC Bioinformatics* 2004, **5**:118.

[8] Basso K, Margolin A, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**(4):382–390.

[9] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. *BMC Bioinformatics* 2006, **7**(Suppl 1):S7.

[10] Priness I, Maimon O, Ben-Gal I: **Evaluation of gene-expression clustering via mutual information distance measure**. *BMC Bioinformatics* 2007, **8**:111, [http://www.biomedcentral.com/1471-2105/8/111].

[11] Meyer P, Lafitte F, Bontempi G: **minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information**. *BMC Bioinformatics* 2008, **9**:461.

[12] Cadeiras M, Bayern MV, Sinha A, Shahzad1 K, Lim WK, Grenett H, Tabak E, Klingler T, Califano A, Deng MC: **Drawing networks of rejection - a systems biological approach to the identification of candidate genes in heart transplantation**. *Journal of Cellular and Molecular Medicine* 2010, **15**(4):949–956.

[13] Allen JD, Xie Y, Chen M, Girard L, Xiao G: **Comparing Statistical Methods for Constructing Large Scale Gene Networks**. *PLoS ONE* 2012, **7**:e29348, [http://dx.doi.org/10.1371%2Fjournal.pone.0029348].

[14] Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: Detecting and evaluating dependencies between variables**. *Bioinformatics* 2002, **18**(Suppl 2):S231–240.

[15] Lindlof A, Lubovac Z: **Simulations of simple artificial genetic networks reveal features in the use of Relevance Networks**. *In Silico Biology* 2005, 5(3):239-250.

[16] Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**(5586):1551–5.

[17] Yip A, Horvath S: **Gene Network Interconnectedness and the Generalized Topological Overlap Measure**. *BMC Bioinformatics* 2007, **8**(8):22.

[18] Li A, Horvath S: **Network neighborhood analysis with the multi-node topological overlap measure**. *Bioinformatics* 2007, **23**(2):222–231.

[19] Hardin J, Mitani A, Hicks L, VanKoten B: **A robust measure of correlation between two genes on a microarray**. *BMC Bioinformatics* 2007, **8**:220.

[20] Langfelder P, Horvath S: **Fast R Functions For Robust Correlations And Hierarchical Clustering**. *Journal of Statistical Software* 2012, **46**(i11).

[21] Horvath S: *Weighted Network Analysis. Applications in Genomics and Systems Biology*. New York: Springer Book 2011.

[22] Mason M, Fan G, Plath K, Zhou Q, Horvath S: **Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells**. *BMC Genomics* 2009, **10**:327.

[23] Cover T, Thomas J: *Elements of information theory*. John Wiley Sons, New York 1991.

[24] Paninski L: **Estimation of entropy and mutual information**. *Neural Computation* 2003, **15**(6):1191–1253.

[25] Kraskov A, Stögbauer H, andrzejak R, Grassberger P: **Hierarchical Clustering Using Mutual Information**. *EPL (Europhysics Letters)* 2007, **70**(2):278.

[26] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles**. *PLoS Biol* 2007, **5**:e8, [http://dx.doi.org/10.1371%2Fjournal.pbio.0050008].

[27] Meyer PE, Kontos K, Lafitte F, Bontempi G: **Information-Theoretic Inference of Large Transcriptional Regulatory Networks**. *EURASIP Journal on Bioinformatics and Systems Biology* 2007, **2007**:79879.

[28] Butte A, Kohane I: **Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurments**. *Pac Symp Biocomput* 2000, :418–429.

[29] Moon YI, Rajagopalan B, Lall U: **Estimation of mutual information using kernel density estimators**. *Phys. Rev. E* 1995, **52**(3):2318–2321.

[30] Oldham M, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind D: **Functional organization of the transcriptome in human brain**. *Nat Neurosci* 2008, **11**(11):1271–1282.

[31] Wolfe C, Kohane I, Butte A: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks**. *BMC Bioinformatics* 2005, **6**:227.

[32] Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R**. *Bioinformatics* 2007, **24**(5):719–20.

[33] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Sherlock GMRG: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25–29.

[34] Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang Y, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**:R80.

[35] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC: **Detecting Novel Associations in Large Data Sets**. *Science* 2011, **334**(6062):1518–1524, [http://www.sciencemag.org/content/334/6062/1518.abstract].

[36] Faraway J: **Practical Regression and Anova using R**. *R pdf file at http://cranr-projectorg/doc/contrib/Faraway-PRApdf* 2002.

[37] D'Haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering**. *Bioinformatics* 2000, **16**(8):707–726, [http://dx.doi.org/10.1093/bioinformatics/16.8.707].

[38] Markowetz F, Spang R: **Inferring cellular networks–a review.** *BMC bioinformatics* 2007, **8 Suppl 6**(Suppl 6):S5+, [http://dx.doi.org/10.1186/1471-2105-8-S6-S5].

[39] Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: **How to infer gene networks from expression profiles**. *Molecular Systems Biology* 2007, **3**:78, [http://dx.doi.org/10.1038/msb4100120].

[40] De Smet R, Marchal K: **Advantages and limitations of current network inference methods**. *Nat Rev Micro* 2010, **8**(10):717–729, [http://dx.doi.org/10.1038/nrmicro2419].

[41] Stolovitzky G, MONROE D, Califano A: **Dialogue on Reverse-Engineering Assessment and Methods**. *Annals of the New York Academy of Sciences* 2007, **1115**(1):1–22.

[42] Stolovitzky G, Prill RJ, Califano A: **Lessons from the DREAM2 Challenges**. *Annals of the New York Academy of Sciences* 2009, **1158**:159–195.

[43] Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G: **Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges**. *PLoS ONE* 2010, **5**(2):e9202.

[44] Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data**. *J Comput Biol* 2000, **7**(3):601–20.

[45] Perrin B, Ralaivola L: **Gene networks inference using dynamic Bayesian networks**. *Bioinformatics* 2003, **19**(Suppl 2):II138–II148.

[46] Friedman N: **Inferring cellular networks using probabilistic graphical models**. *Science* 2004, **303**(5659):799–805.

[47] Li P, Zhang C, Perkins E, Gong P, Deng Y: **Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks**. *BMC Bioinformatics* 2007, **8**(Suppl 7):S13, [http://www.biomedcentral.com/1471-2105/8/S7/S13].

[48] Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED: **Advances to Bayesian network inference for generating causal networks from observational biological data**. *Bioinformatics* 2004, **20**(18):3594–3603, [http://bioinformatics.oxfordjournals.org/content/20/18/3594.abstract].

[49] Zhu J, Lum P, Lamb J, HuhaThakurta D, Edwards S, Thieringer R, Berger J, Wu M, Thompson J, Sachs A, Schadt E: **An integrative genomics approach to the reconstruction of gene networks in segregating populations**. *Cytogenet Genome Res* 2004, **105**:363–374.

[50] Schadt E, Lamb J, Yang X, Zhu J, Edwards J, GuhaThakurta D, Sieberts S, Monks S, Reitman M, Zhang C, Lum P, Leonardson A, Thieringer R, Metzger J, Yang L, Castle J, Zhu H, Kash S, Drake T, Sachs A, Lusis A: **An integrative genomics approach to infer causal associations between gene expression and disease**. *Nature Genetics* 2005, **37**(7):710–717.

[51] Sima C, Hua J, Jung S: **Inference of Gene Regulatory Networks Using Time-Series Data: A Survey**. *Curr Genomics* 2009, **10**(6):416–429.

[52] Shmulevich I, Dougherty ER, Kim S, Zhang W: **Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks**. *Bioinformatics* 2002, **18**(2):261–274, [http://bioinformatics.oxfordjournals.org/content/18/2/261.abstract].

[53] Lahdesmki H, Hautaniemi S, Shmulevich I, Yli-Hrja O: **Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks**. *Signal Processing* 2006, **86**(4):814–834.

[54] Schmitt WA, Raab RM, Stephanopoulos G: **Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data**. *Genome Research* 2004, **14**(8):1654–1663, [http://genome.cshlp.org/content/14/8/1654.abstract].

[55] Fernandes JS, Sternberg PW: **The tailless Ortholog nhr-67 Regulates Patterning of Gene Expression and Morphogenesis in the C. elegans Vulva**. *PLoS Genet* 2007, **3**(4):e69, [http://dx.plos.org/10.1371%2Fjournal.pgen.0030069].

[56] Yan J, Wang H, Liu Y, Shao C: **Analysis of Gene Regulatory Networks in the Mammalian Circadian Rhythm**. *PLoS Comput Biol* 2008, **4**(10):e1000193, [http://dx.doi.org/10.1371%2Fjournal.pcbi.1000193].

[57] Altay G, Emmert-Streib F: **Revealing differences in gene network inference algorithms on the network-level by ensemble methods**. *Bioinformatics* 2010, **26**(14):1738–1744.

[58] Chaitankar V, Ghosh P, Perkins E, Gong P, Zhang C: **Time lagged information theoretic approaches to the reverse engineering of gene regulatory networks**. *BMC Bioinformatics* 2010, **11**(Suppl 6):S19.

[59] Horvath S, Dong J: **Geometric interpretation of Gene Co-expression Network Analysis**. *PloS Comput Biol* 2008, **4**(8):e1000117.

[60] Wiggins C, Nemenman I: **Process pathway inference via time series analysis**. *Experimental Mechanics* 2003, **43**(3):361–370.

[61] Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, Laurance M, Zhao W, Shu Q, Lee Y, Scheck A, Liau L, Wu H, Geschwind D, Febbo P, Kornblum H, TF C, Nelson S, Mischel P: **Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target**. *Proc Natl Acad Sci U S A* 2006, **103**(46):17402–7.

[62] Goring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J: **Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes**. *Nat Genet* 2007, **39**:1208 – 1216.

[63] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization**. *Mol Biol Cell* 1998, **9**(12):3273–3297.

[64] Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson SF: **Gene Connectivity, Function, and Sequence Conservation: Predictions from Modular Yeast Co-expression Networks**. *BMC Genomics* 2006, **7**(7):40.

[65] Ghazalpour A, Doss S, Zhang B, Plaisier C, Wang S, Schadt E, Thomas A, Drake T, Lusis A, Horvath S: **Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight**. *PloS Genetics* 2006, **2**(2):8.

[66] Fuller T, Ghazalpour A, Aten J, Drake T, Lusis A, Horvath S: **Weighted gene coexpression network analysis strategies applied to mouse weight**. *Mamm Genome* 2007, **18**(6-7):463–472.

[67] Wilcox R: *Introduction to Robust Estimation and Hypothesis Testing*. San Diego: Academic Press 1997.

[68] Dong J, Horvath S: **Understanding Network Concepts in Modules**. *BMC Syst Biol* 2007, **1**:24.

[69] Albanese D, Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C: **cmine, minerva and minepy: a C engine for the MINE suite and its R and Python wrappers**. *ArXiv e-prints* 2012.

[70] Li H, Zhan M: **Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data**. *Bioinformatics* 2008, **24**(17):1874–1880.

[71] Kauffman S: **Metabolic stability and epigenesis in randomly connected nets**. *J.Theoret.Biol.* 1969, **22**:437–467.

[72] Chen X, Chen M, Ning K: **BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network.** *Bioinformatics* 2006, [http://view.ncbi.nlm.nih.gov/pubmed/17005537].

[73] Werhli AV, Grzegorczyk M, Husmeier D: **Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks**. *Bioinformatics* 2006, **22**(20):2523–2531, [http://dx.doi.org/10.1093/bioinformatics/btl391].

[74] Pinsky P, Zhu C: **Building multi-marker algorithms for diesease prediction: the role of correlations among markers.** *Biomarker insights* 2011, **6:**83–93.

[75] Vapnik V: *The nature of statistical learning theory.* Springer, New York; 2000.

[76] Breiman L, Friedman J, Stone C, Olshen R: *Classification and regression trees.* Wadsworth International Group, California; 1984.

[77] Dudoit S, Fridlyand J, Speed TP: **Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.** *J Am Stat Assoc* 2002, **97**(457):77–87.

[78] Diaz-Uriarte R, Alvarez de Andres S: **Gene selection and classification of microarray data using random forest.** *BMC Bioinformatics* 2006, **7:**3. [http://www.biomedcentral.com/1471-2105/7/3]

[79] Pirooznia M, Yang J, Yang MQ, Deng Y: **A comparative study of different machine learning methods on microarray gene expression data.** *BMC Genomics* 2008, **9**(Suppl 1):S13. [http://www.biomedcentral.com/1471-2164/9/S1/S13]

[80] Caruana R, Niculescu-Mizil A: **An empirical comparison of supervised learning algorithms.** In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, New York, NY, USA: ACM 2006:161–168. [http://doi.acm.org/10.1145/1143844.1143865]

[81] Statnikov A, Wang L, Aliferis C: **A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.** *BMC Bioinformatics* 2008, **9**(1):319, [http://www.biomedcentral.com/1471-2105/9/319]

[82] Caruana R, Karampatziakis N, Yessenalina A: **An empirical evaluation of supervised learning in high dimensions.** In *Proceedings of the 25th international conference on Machine learning*, ICML '08, New York, NY, USA: ACM 2008:96–103. [http://doi.acm.org/10.1145/1390156.1390169]

[83] Breiman L: **Bagging Predictors.** *Machine Learning* 1996, **24:**123–140.

[84] Derksen S, Keselman HJ: **Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables.** *British J Mathematical Stat Psychology* 1992, **45**(2):265–282. [http://dx.doi.org/10.1111/j.2044-8317.1992.tb00992.x]

[85] Harrell FJ, Lee K, Mark D: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.** *Stat med* 1996, **15:**361–387.

[86] Breiman L: **Random Forests.** *Machine Learning* 2001, **45:**5–32.

[87] Svetnik V, Liaw A, Tong C, Wang T: **Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules.** In *Multiple Classifier Systems*, *Volume 3077 of* Lecture Notes in Computer Science. Edited by Roli F, Kittler J, Windeatt T, Springer Berlin / Heidelberg 2004:334–343.

[88] Shi T, Horvath S: **Unsupervised Learning With Random Forest Predictors.** *J Comput Graphical Stat* 2006, **15:**118–138. [http://dx.doi.org/10.1198/106186006X94072]

[89] McCullagh P, Nelder J: *Generalized Linear Models.* second edition, London: Chapman and Hall/CRC, 1989.

[90] Ho TK: **The random subspace method for constructing decision forests.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998, **20**(8):832–844. [http://dx.doi.org/10.1109/34.709601]

[91] Prinzie A, den Poel DV: **Random Forests for multiclass classification: Random MultiNomial Logit.** *Expert Syst Appl* 2008, **34**(3):1721 – 1732. [http://www.sciencedirect.com/science/article/pii/S0957417407000498]

[92] Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL: **Classification by ensembles from random partitions of high-dimensional data.** *Comput Stat Data Anal* 2007, **51**(12):6166–6179. [http://dx.doi.org/10.1016/j.csda.2006.12.043]

[93] Moon H, Ahn H, Kodell RL, Baek S, Lin CJ, Chen JJ: **Ensemble methods for classification of patients for personalized medicine with high-dimensional data.** *Artif Intelligence Med* 2007, **41**(3):197–207. [http://www.sciencedirect.com/science/article/pii/S0933365707000863]

[94] Panov P, Džeroski S: **Combining bagging and random subspaces to create better ensembles.** In *Proceedings of the 7th international conference on Intelligent data analysis*, IDA'07, Berlin, Heidelberg: Springer-Verlag 2007:118–129. [http://dl.acm.org/citation.cfm?id=1771622.1771637]

[95] Venables W, Ripley B: *Modern Applied Statistics with S.* fourth edition, New York: Springer, 2002.

[96] Ripley B: *Pattern Recognition and Neural Networks.* Cambridge, UK: Cambridge University Press 1996.

[97] Dettling M, Bühlmann P: **Supervised clustering of genes.** *Genome Biology* 2002, **3**(12):research0069.1–research0069.15. [http://genomebiology.com/2002/3/12/research/0069]

[98] Chang C, Lin C: **LIBSVM: a library for Support Vector Machines.** [http://www.csie.ntu.edu.tw/~cjlin/libsvm]

[99] Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci U S A* 2002, **99**(10):6567–6572.

[100] Draper N, Smith H, Pownell E: *Applied regression analysis. Volume 3.* New York: Wiley; 1966.

[101] Tibshirani R: **Regression shrinkage and selection via the lasso.** *J R Stat Soc. Ser B (Methodological)* 1996, :267–288.

[102] Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Stat Soc: Ser B (Statistical Methodology)* 2005, **67**(2):301–320.

[103] Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J stat software* 2010, **33:**1.

[104] Simon N, Friedman JH, Hastie T, Tibshirani R: **Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.** *J Stat Software* 2011, **39**(5):1–13. [http://www.jstatsoft.org/v39/i05]

[105] Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nat genet* 2003, **33:**49–54. [http://dx.doi.org/10.1038/ng1060]

[106] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, Mclaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415**(6870):436–442. [http://dx.doi.org/10.1038/415436a]

[107] van't Veer L, Dai H, van de Vijver M, He Y, Hart A, Mao M, Peterse H, van der kooy K, Marton M, Witteveen A, Schreiber G, Kerkhoven R, Roberts C, Linsley P, Bernards R, Friend S: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530–6.

[108] Alon U, Barkai N, Notterman DA, Gishdagger K, Ybarradagger S, Mackdagger D, Levine AJ: **Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays.** *Proc Natl Acad Sci U S A* 1999, **96:**6745–50.

[109] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**(5439):531–7.

[110] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, :503–511.

[111] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24**(3).

[112] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1**(2):203–209. [http://view.ncbi.nlm.nih.gov/pubmed/12086878]

[113] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7**(6):673–679. [http://dx.doi.org/10.1038/89044]

[114] Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, von Deimling A, Pomeroy SL, Golub TR, Louis DN: **Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification.** *Cancer Res* 2003, **63**(7):1602–1607. [http://cancerres.aacrjournals.org/content/63/7/1602.abstract].

[115] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8:**68–74. [http://dx.doi.org/10.1038/nm0102-68]

[116] Kuner R, Muley T, Meister M, Ruschhaupt M, Buness A, Xu EC, Schnabel P, Warth A, Poustka A, Sültmann H, Hoffmann H: **Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes.** *Lung Cancer* 2009, **63:**32–38.

[117] Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Fárez-Vidal M: **Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer.** *Int J Cancer* 2011, **129**(2):355–364. [http://dx.doi.org/10.1002/ijc.25704]

[118] **Clinically annotated tumor database.** [https://expo.intgen.org/geo/]

[119] Swindell WR, Johnston A, Carbajal S, Han G, Wohn C, Lu J, Xing X, Nair RP, Voorhees JJ, Elder JT, Wang XJ, Sano S, Prens EP, DiGiovanni J, Pittelkow MR, Ward NL, Gudjonsson JE: **Genome-Wide Expression Profiling of Five Mouse Models Identifies Similarities and Differences with Human Psoriasis.** *PLoS ONE* 2011, **6**(4):e18266. [http://dx.doi.org/10.1371\%2Fjournal.pone.0018266]

[120] Nair RP, Duffin KCC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJJ, Ruether A, Schreiber S, Weichenthal M, Gladman D, Rahman P, Schrodi SJ, Prahalad S, Guthery SL, Fischer J, Liao W, Kwok PYY, Menter A, Lathrop GM, Wise CA, Begovich AB, Voorhees JJ, Elder JT, Krueger GG, Bowcock AM, Abecasis GR, Collaborative Association Study of Psoriasis: **Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.** *Nat genet* 2009, **41**(2):199–204. [http://dx.doi.org/10.1038/ng.311]

[121] Yao Y, Richman L, Morehouse C, de los Reyes M, Higgs BW, Boutrin A, White B, Coyle A, Krueger J, Kiener PA, Jallal B: **Type I Interferon: Potential Therapeutic Target for Psoriasis?** *PLoS ONE* 2008, **3**(7):e2737. [http://dx.plos.org/10.1371\%2Fjournal.pone.0002737]

[122] Brynedal B, Khademi M, Wallström E, Hillert J, Olsson T, Duvefelt K: **Gene expression profiling in multiple sclerosis: A disease of the central nervous system, but with relapses triggered in the periphery?** *Neurobiology of Disease* 2010, **37**(3):613 – 621. [http://www.sciencedirect.com/science/article/pii/S0969996109003362]

[123] Kemppinen AK, Kaprio J, Palotie A, Saarela J: **Systematic review of genome-wide expression studies in multiple sclerosis**. *BMJ Open* 2011, **1**. [http://bmjopen.bmj.com/content/1/1/e000053.abstract]

[124] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.** *Bioinformatics* 2005, **21**(5):631–643. [http://bioinformatics.oxfordjournals.org/content/21/5/631.ab\stract]

[125] Li S, Harner EJ, Adjeroh D: **Random KNN feature selection - a fast and stable alternative to Random Forests.** *BMC Bioinformatics* 2011, **12:**450. [http://www.biomedcentral.com/1471-2105/12/450]

[126] Chang CC, Lin CJ: **Training v-Support Vector Classifiers: Theory and Algorithms.** *Neural Comput* 2001, **13**(9):2119–2147.

[127] Yang F, Wang Hz, Mi H, Lin Cd, Cai Ww: **Using random forest for reliable classification and cost-sensitive learning for medical diagnosis.** *BMC Bioinformatics* 2009, **10**(Suppl 1):S22. [http://www.biomedcentral.com/1471-2105/10/S1/S22]

[128] Lopes F, Martins D, Cesar R: **Feature selection environment for genomic applications.** *BMC Bioinformatics* 2008, **9**(1):451. [http://www.biomedcentral.com/1471-2105/9/451]

[129] Frank A, Asuncion A: **UCI Machine Learning Repository.** 2010, [http://archive.ics.uci.edu/ml]

[130] Meinshausen N, Bühlmann P: **Stability selection.** *J R Stat Soc: Ser B (Statistical Methodology)* 2010, **72**(4):417–473. [http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x]

[131] Furlanello C, Serafini M, Merler S, Jurman G: **An accelerated procedure for recursive feature ranking on microarray data.** *Neural Networks* 2003, **16:**641 – 648. [http://www.sciencedirect.com/science/article/pii/S0893608003001035]

[132] Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**(19):2507–2517. [http://bioinformatics.oxfordjournals.org/content/23/19/2507.abstract]

[133] Perlich C, Provost F, Simonoff JS: **Tree Induction vs. Logistic Regression: A Learning-Curve Analysis.** *J machine learning resarch* 2003, **4:**211–255.

[134] Arena V, Sussman N, Mazumdar S, Yu S, Macina O: **The Utility of Structure-Activity Relationship (SAR) Models for Prediction and Covariate Selection in Developmental Toxicity: Comparative Analysis of Logistic Regression and Decision Tree Models.** *SAR and QSAR in Environ Res* 2004, **15:**1–18. [http://www.tandfonline.com/doi/abs/10.1080/1062936032000169633]

[135] Pino-Mejias R, Carrasco-Mairena M, Pascual-Acosta A, Cubiles-De-La-Vega MD, Munoz-Garcia J: **A comparison of classification models to**

identify the Fragile X Syndrome. *J Appl Stat* 2008, **35**(3):233–244. [http://www.tandfonline.com/doi/abs/10.1080/02664760701832976]

[136] van Wezel M, Potharst R: **Improved customer choice predictions using ensemble methods.** *Eur J Operational Res* 2007, **181:**436 – 452. [http://www.sciencedirect.com/science/article/pii/S0377221706003900]

[137] Wang G, Hao J, Ma J, Jiang H: **A comparative assessment of ensemble learning for credit scoring.** *Expert Syst Appl* 2011, **38:**223–230. [http://dx.doi.org/10.1016/j.eswa.2010.06.048]

[138] Shadabi F, Sharma D: **Comparison of Artificial Neural Networks with Logistic Regression in Prediction of Kidney Transplant Outcomes.** In *Proceedings of the 2009 International Conference on Future Computer and Communication*, ICFCC '09, Washington, DC, USA: IEEE Computer Society 2009:543–547. [http://dx.doi.org/10.1109/ICFCC.2009.139]

[139] Sohn S, Shin H: **Experimental study for the comparison of classifier combination methods.** *Pattern Recognit* 2007, **40:**33 – 40. [http://www.sciencedirect.com/science/article/pii/S0031320306003116]

[140] Bühlmann P, Yu B: **Analyzing Bagging.** *Ann Stat* 2002, **30:**927–961.

[141] Freund Y, Schapire RE: **A decision-theoretic generalization of on-line learning and an application to boosting.** In *Proceedings of the Second European Conference on Computational Learning Theory*, EuroCOLT '95, London, UK, UK: Springer-Verlag 1995:23–37. [http://dl.acm.org/citation.cfm?id=646943.712093]

[142] Song L, Langfelder P, Horvath S: **Random generalized linear model: a highly accurate and interpretable ensemble predictor.** *BMC Bioinformatics* 2013, **14:**5.

[143] Meyer P, Hoeng J, Rice JJ, Norel R, Sprengel J, Stolle K, et al.: **Industrial methodology for process verification in research (IMPROVER): toward systems biology verification.** *Bioinformatics* 2012, **28:**1193-201.

[144] Halbert R, Natoli J, Gano A, Badamgarav E, Buist A, Mannino D: **Global burden of COPD: systematic review and meta-analysis.** *European Respiratory Journal* 2006, **28:**523-32.

[145] Fabbri LM, Rabe KF: **From COPD to chronic systemic inflammatory syndrome?** *Lancet* 2007, **370:**797-9.

[146] Holleman DR, Jr., Simel DL: **Does the clinical examination predict airflow limitation?** *JAMA* 1995, **273:**313-9.

[147] Ito K, Barnes PJ: **COPD as a disease of accelerated lung aging.** *Chest* 2009, **135:**173-80.

[148] Vestbo J, Hurd SS, Agusti AG, Jones PW, Vogelmeier C, Anzueto A, et al.: **Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease GOLD Executive Summary.** *American journal of respiratory and critical care medicine* 2013, **187:**347-65.

[149] Sørheim I-C, Johannessen A, Gulsvik A, Bakke PS, Silverman EK, DeMeo DL: **Gender differences in COPD: are women more susceptible to smoking effects than men?** *Thorax* 2010, **65:**480-5.

[150] Bhattacharya S, Srisuma S, DeMeo DL, Shapiro SD, Bueno R, Silverman EK, et al.: **Molecular biomarkers for quantitative and discrete COPD phenotypes.** *American journal of respiratory cell and molecular biology* 2009, **40:**359.

[151] Oldham MC, Langfelder P, Horvath S: **Network methods for describing sample relationships in genomic datasets: application to Huntington's disease.** *BMC Syst Biol* 2012, **6:**63.

[152] Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8:**118-27.

[153] Harrell FE: *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* Springer, 2001.

[154] Carolan BJ, Harvey BG, De BP, Vanni H, Crystal RG: **Decreased expression of intelectin 1 in the human airway epithelium of smokers compared to nonsmokers.** *The Journal of Immunology* 2008, **181:**5760-7.359.

[155] Vanni H, Kazeros A, Wang R, Harvey BG, Ferris B, De BP, et al.: **Cigarette smoking induces overexpression of a fat-depleting gene AZGP1 in the human.** *Chest* 2009, **135:**1197-208.

[156] Raman T, O'Connor TP, Hackett NR, Wang W, Harvey BG, Attiyeh MA, et al.: **Quality control in microarray assessment of gene expression in human airway epithelium.** *BMC Genomics* 2009, **10:**493.

[157] Hubner RH, Schwartz JD, De Bishnu P, Ferris B, Omberg L, Mezey JG, et al.: **Coordinate control of expression of Nrf2-modulated genes in the human small airway epithelium is highly responsive to cigarette smoking.** *Mol Med* 2009, **15:**203-19.

[158] Turetz ML, O'Connor TP, Tilley AE, Strulovici-Barel Y, Salit J, Dang D, et al.: **Trachea epithelium as a "canary" for cigarette smoking-induced biologic phenotype of the small airway epithelium.** *Clin Transl Sci* 2009, **2:**260-72.

[159] Strulovici-Barel Y, Omberg L, O'Mahony M, Gordon C, Hollmann C, Tilley AE, et al.: **Threshold of biologic responses of the small airway epithelium to low levels of tobacco smoke.** *Am J Respir Crit Care Med* 2010, **182:**1524-32.

[160] Shaykhiev R, Otaki F, Bonsu P, Dang DT, Teater M, Strulovici-Barel Y, et al.: **Cigarette smoking reprograms apical junctional complex molecular architecture in the human airway epithelium in vivo.** *Cell Mol Life Sci* 2011, **68:**877-92.

[161] Carolan BJ, Heguy A, Harvey BG, Leopold PL, Ferris B, Crystal RG: **Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers.** *Cancer Res* 2006, **66:**10729-40.

[162] Tilley AE, Harvey BG, Heguy A, Hackett NR, Wang R, O'Connor TP, et al.: **Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease.** *Am J Respir Crit Care Med* 2009, **179:**457-66.

[163] Ammous Z, Hackett NR, Butler MW, Raman T, Dolgalev I, O'Connor TP, et al.: **Variability in small airway epithelial gene expression among normal smokers.** *Chest* 2008, **133:**1344-53.