Bioinformatics, 31, 2015, i311–i319
doi: 10.1093/bioinformatics/btv255
ISMB/ECCB 2015

OXFORD

FERAL: network-based classifier with application to breast cancer outcome prediction

Amin Allahyar and Jeroen de Ridder*

Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

*To whom correspondence should be addressed.

Abstract

Motivation: Breast cancer outcome prediction based on gene expression profiles is an important strategy for personalize patient care. To improve performance and consistency of discovered markers of the initial molecular classifiers, network-based outcome prediction methods (NOPs) have been proposed. In spite of the initial claims, recent studies revealed that neither performance nor consistency can be improved using these methods. NOPs typically rely on the construction of meta-genes by averaging the expression of several genes connected in a network that encodes protein interactions or pathway information. In this article, we expose several fundamental issues in NOPs that impede on the prediction power, consistency of discovered markers and obscures biological interpretation.

Results: To overcome these issues, we propose FERAL, a network-based classifier that hinges upon the Sparse Group Lasso which performs simultaneous selection of marker genes and training of the prediction model. An important feature of FERAL, and a significant departure from existing NOPs, is that it uses multiple operators to summarize genes into meta-genes. This gives the classifier the opportunity to select the most relevant meta-gene for each gene set. Extensive evaluation revealed that the discovered markers are markedly more stable across independent datasets. Moreover, interpretation of the marker genes detected by FERAL reveals valuable mechanistic insight into the etiology of breast cancer.

Availability and implementation: All code is available for download at: http://homepage.tudelft.nl/53a60/resources/FERAL/FERAL.zip.

Contact: j.deridder@tudelft.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Breast cancer is the most frequently diagnosed type of cancer and one of the leading causes of death in women (Fantozzi and Christofori, 2006). The main cause of death in these patients is, however, not the primary tumor, but its metastases at distant sites (e.g. in bone, lung, liver and brain) (Weigelt *et al.*, 2005). Typical risk factors such as lymph node status and tumor size are insufficient to accurately predict the risk of metastasis in patients (Shapiro and Recht, 2001; Weigelt *et al.*, 2005). Over the last few years, substantial efforts have been made on deriving molecular classifiers to predict clinical outcome based on gene expression profiles obtained from the primary tumor (van't Veer *et al.*, 2002; Van De Vijver *et al.*, 2002; Weigelt *et al.*, 2005).

A fundamental limitation of breast cancer outcome prediction is that it has proved very difficult to obtain a robust classifier performance across different datasets. It was found that, despite properly cross-validated classifier training, prediction performance decreases dramatically when a classifier trained on one dataset is applied to another one (Lazar *et al.*, 2013; Soneson *et al.*, 2014). Moreover, the prognostic gene signatures identified using these classifiers have poor concordance across different studies (Ein-Dor *et al.*, 2005; van Vliet *et al.*, 2008). This points to a lack of a unified mechanism through which clinical outcome can be explained from gene expression profiles, which is still a major hurdle in clinical cancer biology.

Several studies ascribe the lack of classification robustness to insufficient patient sample size (Hua *et al.*, 2009). Other causes may be the inherent measurement noise in microarray experiments or heterogeneity in the samples (Ein-Dor *et al.*, 2005; Symmans *et al.*, 1995). To mitigate these issues, breast cancer datasets are often

pooled to capture the information of as many samples as possible in the predictor (Shen *et al.*, 2004; van Vliet *et al.*, 2008). It remains, however, an open question how many samples are sufficient to account for all the noise and heterogeneity.

One of the hallmarks of cancer is that it is caused by deregulation of several processes or cellular pathways through multiple somatic mutations (Hanahan and Weinberg, 2000, 2011). More recent efforts of outcome prediction aim to exploit this hallmark by taking existing knowledge on relations between genes and pathways into account in the classifier. A common approach is to aggregate several functionally related genes to produce discriminative meta-genes or subnetworks (Babaei *et al.*, 2011; Dao *et al.*, 2010; Pujana *et al.*, 2007; Taylor *et al.*, 2009; Van den Akker *et al.*, 2011). Often, functional relationships between genes are determined based on the topology of a pre-defined biological network such as a co-expression network (Park *et al.*, 2007), cellular pathway map (Lee *et al.*, 2008) or protein–protein interaction (PPI) network (Chuang *et al.*, 2007). Therefore, we refer to such approaches as network-based outcome prediction methods (NOPs).

The approach proposed by Park *et al.* (2007) is among the first NOPs. Initially, the co-expression network is partitioned into gene sets using a linkage algorithm. Next, meta-genes are formed by taking the average expression of the genes in each gene set. Consequently, highly correlated genes will be aggregated which reduces the number of features and co-linearity among genes. The appropriate number of clusters, which determines the scale at which meta-genes are assembled, is determined by cross-validation.

Chuang *et al.* (2007) exploit the PPI network to identify predictive gene sets (called sub-networks in their work). Gene sets are constructed by a greedy procedure which starts with a gene (i.e. seed gene) and extends iteratively by adding the neighboring gene that provides the highest mutual information between corresponding average meta-gene and target label.

Taylor *et al.* (2009) exploit the topology of the PPI network. In this method, predictive hub genes (i.e. genes with more than five connections) are ranked based on the absolute difference in within-class correlation between the hub and its neighbors. The corresponding meta-genes are constructed by taking the difference of expression between the hub and its neighbors.

Unfortunately, contrary to previous claims, recent studies reported that many NOPs do not outperform a model trained over single gene features (Cun and Frohlich, 2012; Staiger *et al.*, 2012, 2013). Notably, in the analysis carried out by Staiger *et al.* (2013), neither significant improvement of classification performance nor an improvement of gene signature stability was observed, despite the fact that these authors examined many different methods and experimented with several biological networks. Perhaps even more striking is the finding that utilizing random networks (Staiger *et al.*, 2012) or integrating random genes as markers (Venet *et al.*, 2011) performs on par with complex NOPs. Taken together, it appears that current NOPs have produced very limited progress on solving the issue of robust classification performance and robust prognostic gene signature selection. This also casts doubt on the potential to extract useful insights from the derived prognostic gene signatures into the mechanisms underlying the disease.

The main goal of this article is to identify and alleviate several fundamental issues in current NOPs that impede on reaching robust prediction performance and identify a stable prognostic gene signature. We find that the main bottleneck in current NOPs is that the frequently used average operator is a poor choice to integrate the expression of functionally related genes. Moreover, the use of a single operator may not be sufficient to capture and summarize the aberration of higher level functions in cell. In addition, we conclude that decoupling the training of the classifier from the selection of genes to be used in meta-genes or the selection of the meta-genes themselves hampers the stability of gene signature identification.

To address these issues, we propose FERAL (DelFt nEtwoRk-bAsed cLassifier), a new NOP that is based on the Sparse Group Lasso (SGL) (Simon *et al.*, 2013; Yuan and Lin, 2006). SGL exploits groups of features (i.e. gene sets) and yields sparsity at both group (i.e. gene set) and feature (i.e. gene/meta-genes) levels (Friedman *et al.*, 2010). In this way, simultaneous selection of features and training of the prediction model is achieved (see Supplementary Section S1 for explanation of Lasso and its variants). Furthermore, instead of using a single operator to integrate gene-expression into meta-genes, FERAL exploits a wide range of such operators, including a previously unexplored supervised integration strategy.

We present extensive experiments using a compendium dataset called ACES (Amsterdam Classification Evaluation Suite), which was recently used for NOP model evaluation (Staiger *et al.*, 2013). FERAL achieves statistically significant performance improvement, owing to the regularization of the SGL and inclusion of multiple integration operators. We moreover find substantially improved stability of the selected prognostic gene sets. Taken together, these feats enable biological interpretation of the trained classifier, which, we find, results in highly relevant mechanistic insights.

2 Method

To motivate the design choices of FERAL we start by outlining the basic properties of existing NOPs. We focus on three well-known models proposed for network-based outcome prediction. Nonetheless, there are numerous network-based methods, which we do not take into consideration. A closer look at these methods reveals that in fact they all take two main steps to incorporate network information: gene set selection and integration (Fig. 1a). The selection step should result in gene sets that represent (part of) a cellular process or pathway that collectively exhibit aberrant behavior. In the integration step, the selected genes are summarized to produce a meta-gene capable of representing the aberrant behavior in the corresponding cellular process. Typically, this is followed by an additional round of selection and integration in which meta-genes are selected and integrated to produce a final prediction.

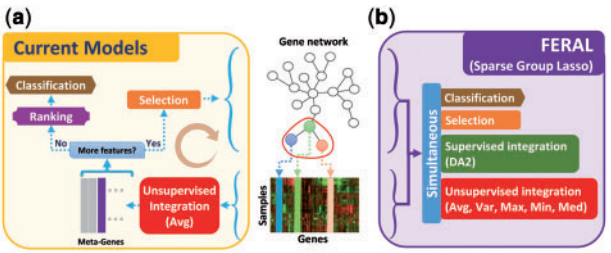

Fig. 1. Overview of the proposed model (FERAL). (**a**) Current models follow a similar path in which several nearby genes (according to a given network) are selected and then integrated using an average operator resulting in a meta-gene. These meta-genes are then ranked based on a pre-defined scoring function and top candidates are presented to the final classifier. (**b**) Instead of being limited to average-based meta-genes, FERAL computes several meta-genes using different operators and employs the SGL to select the most appropriate meta-gene for each specific gene set while simultaneously performing selection, integration and classification

2.1 Integration of gene sets into meta-genes

Most NOPs use the average operator to summarize gene expression into meta-gene expression. However, other biologically inspired operations, such as the max/min (to model AND/OR relations) or the variance (to capture variability of expression levels among genes close in the network) might also be suitable for representing higher level functions in cell. The assumption in many NOPs is that the directionality of the aberrant activity is the same (i.e. over/under expression) for nearby genes in the network. This may be inappropriate, for instance when genes exhibit opposite association with respect to the class label. In such cases, the average operator can even cancel out their predictive contribution. By assessing the expression correlation of PPIs, we established that this is a frequent event (Fig. 2a and Supplementary Fig. S3).

This problem arises because the aforementioned operators are unsupervised, i.e. an identical meta-gene would be produced using shuffled sample labels. This can be resolved by using a linear or non-linear regressor that considers the labels for achieving the best performance. In spite of their superior performance (see Fig. 2b; Supplementary Fig. S4), supervised integration operators may promote overfitting. This issue is apparent when linear operators are compared with non-linear ones (e.g. Decision Tree and support vector machine). Hence, in the integration procedure, a trade-off exists between performance and complexity.

To alleviate this issue, we propose the Direction Aware Average (DA2) operator, which adjusts the direction of genes before taking the average (see also Supplementary Section S13). DA2 is defined as:

$$\mathrm{DA2}_g = \frac{1}{|\Psi_g|}\sum_{j\in\psi_g} \mathrm{sgn}(C_j) \times E_j,$$

where Ψ_g is the gene set of seed gene g and E_j and C_j contain the expression and correlation values with the class label of gene j, respectively. Just like all supervised meta-gene constructors, DA2 only uses training samples for calculating C_j. The DA2 provides a balance between stability of unsupervised operators (owing to its simplicity) and performance of supervised operators. It suffers less from overfitting due to the fact that labels are only employed to detect the direction of genes which is more stable compared with their individual predictive power. This is also apparent from our experiment (Fig. 2b), as the DA2 provided a comparable performance to top integrating operators (e.g. regression and the Lasso).

It is worth noting that different integration operators offer different representations of higher level cellular functions. The proper operator for each gene set is not known *a priori*. It might be beneficial to use multiple of such operators and allow the classifier to select the appropriate operator to describe a gene set or allow a single gene set to be described using multiple operators. In addition to potentially achieving better performance, it provides insights into the underlying aberrant behavior of each gene set. To the best of our knowledge, there are no NOPs that use multiple integration operators.

In FERAL, gene sets are formed by the individual gene expression profiles extended with several meta-genes produced by aggregating gene expression of these genes. We included the following unsupervised aggregations. The average operator, to model the overall expression level of the gene set in a fully unsupervised way. The median operator, similar to average but with reduced sensitivity to outliers. The variance operator, to measure the fluctuation in expression of interacting genes as this may point to a loss of regulation due to rewiring. Min and max, to model the AND/OR relationship between genes. In addition to these unsupervised operators, the linear integration (which is implicitly provided by the SGL) and DA2 were also included as supervised operators. However, the supervised non-linear meta-genes, which are presented in the analysis in Figure 2b were not included, since it was observed that they were prone to overfitting (data not shown).

2.2 Selection of genes in gene sets

To determine which genes will be summarized in a meta-gene, Park selects all genes in a correlation cluster whereas Taylor uses all genes that are connected to the same hub gene in the PPI network. Both of these methods are likely to produce a highly skewed cluster size distribution, with a few very large clusters and many smaller ones (Albert, 2005; Chen *et al.*, 2002). These large clusters will contain a substantial number of irrelevant genes that may not only hamper the performance but also limit the interpretability of the meta-gene as it is difficult to identify the driver genes amongst all genes in the gene set (Cheng *et al.*, 2014). Moreover, in case of Taylor, only genes connected to hub genes can appear in a meta-gene, which *a priori* greatly limits the repertoire of genes that can be used in the final predictor.

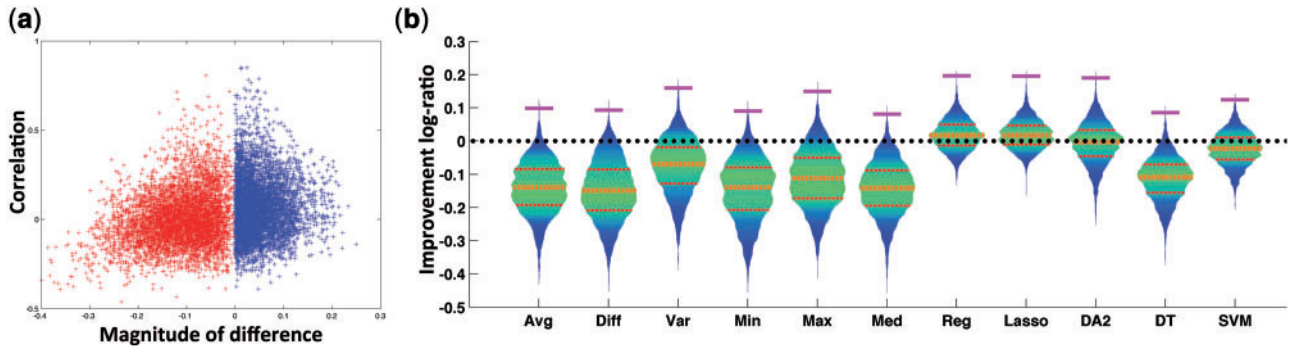

Fig. 2. Evaluation of different integration operators. (**a**) Visualization of the consistency in the direction of association with the target label for connected gene pairs in the I2D network. The *x*-axis represents the magnitude of difference, defined as $\mathrm{abs}(C_a - C_b) \times \mathrm{Sgn}(C_a \times C_b)$, where C_x denotes the correlation between gene *x* and the target label and Sgn is sign function. The *y*-axis is the correlation between two genes (see Supplementary Section S3 for details). (**b**) Performance comparison between 11 operators including (from left to right): average, average of differences between seed gene and its interactors (implemented in Taylor), variance, minimum, maximum, median, regression, lasso, DA2, Decision Tree (DT) and support vector machine with an RBF kernel. To generate each violin plot, 5000 randomly selected seed genes and their 9 closest neighbors according to the I2D network were integrated into a meta-gene using one of the operators, and the predictive performance (AUC) is determined. The *y*-axis represents the improvement log ratio of the AUC obtained with the meta-gene with the highest AUC of the individual genes. This comparison shows that other operators are able to provide similar or even better performance compared with average operator. Interestingly, adjusting the direction of genes before taking the average can improve the performance considerably

Instead, in FERAL, the gene set size is kept constant. This is achieved by defining gene sets as groups of k genes—a seed gene with $k-1$ of its closest neighbors. If performance is the only goal, the setting of k is determined by an additional inner cross-validation. By varying k we found, however, that a tradeoff exists between performance and relevance of the marker genes to cancer and set $k=10$ to provide a balance between them (see Supplementary Section S11).

To ensure each gene is included in at least one gene set, all genes were considered as seed genes, resulting in a total of N gene sets. In case a seed gene has more than $k-1$ neighbors, the gene set is reduced to a total of k genes by randomly removing genes. This random selection did not result in large performance variation (see Supplementary Section S12 for details). In case a seed gene has less than $k-1$ neighbors, the neighbors of the neighbors are considered in a similar fashion. When a weighted network is used, the edge weights are taken into account while determining the closest neighbors.

Chuang employs a greedy search to define subnetworks. This is done by iteratively extending the network from a seed gene guided by a supervised performance criterion. Because label information is used to guide the network growing, this increases the risk of overfitting and thereby it reduces the performance and the stability of selected gene sets. Moreover, this procedure also critically depends on the accuracy of gene–gene interactions, which may be problematic as concerns exist about the reliability of individual interactions in these networks (Cusick *et al.*, 2008; Von Mering *et al.*, 2002).

Instead of including all genes in a group (Park and Taylor) or using a greedy search in a noisy network (Chuang), FERAL leverages the fact that the SGL performs embedded feature selection. This is realized because SGL provides regularization both at the level of the individual genes and the gene set level. As a result, selection of the most relevant genes will be performed if sufficiently large gene sets are provided. Because feature selection and classifier training are performed simultaneously, classifiers that offer embedded feature selection often provide improved performance and select more relevant features (Guyon *et al.*, 2006). Moreover, embedded feature selection techniques prevent the need of additional cross-validation loops that are required to prevent overfitting.

2.3 Pre-ranking and integration of meta-genes

After producing the meta-genes, most NOPs employ a ranking step. This step can be considered as a second selection step at the meta-gene level. Typically, each meta-gene is assessed based on a predefined ranking function (e.g. mutual information, t-test or permutation test) and the top candidates will be used in the final prediction step (akin to so-called individual feature selection). Evaluation of meta-genes in the methods of Chuang and Taylor is performed one at a time. Hence, the ranking procedure cannot identify multiple synergistic meta-genes when they have poor individual performance nor can it determine whether several meta-genes contain the same information and are therefore redundant (see Supplementary Fig. S2.2 for an example of such cases in Chuang's method).

As FERAL employs the SGL, which performs embedded feature selection at the gene set level, the need of meta-gene selection is circumvented altogether. This greatly improves gene set stability.

2.4 Improvements on standard NOPs

To compare against, we use the methods from Park, Chuang and Taylor, henceforth referred to as standard methods. Based on our discussion so far, it seems reasonable to change a few parts of these standard methods that evidently impede their performance. The original version of each method (prefixed by o) is implemented by strictly following the procedure described in the authors' paper. Additionally, we implemented an improved version (prefixed by i), which includes obvious improvements beneficial for their performance and stability (see Supplementary Section S2 for details). More specifically for Park's method, instead of training individual Lasso over the meta-genes produced in each level of hierarchical tree, single Lasso was trained over all meta-genes collected from levels of hierarchical tree. For Taylors method, similar to Staiger *et al.*, we took the average of differences between hub and its interactor for corresponding meta-gene. Finally, we removed the ranking procedure in Taylor and Chuang methods and, similar to Park, used the Lasso to achieve a simultaneous selection and integration of the meta-genes. To assess the utility of biological networks in the outcome prediction problem, we also included a Lasso trained on the individual genes, i.e. without exploiting network information.

2.5 Ranking and scoring of marker genes

One of the main objectives in NOPs is to detect marker genes that play a role in driving this complex disease. This can be achieved by ranking them on a pre-defined score that captures the contribution of the genes on the final prediction performance. In the Chuang method, gene sets (i.e. sub-networks) are ranked based on P value that is obtained using a permutation test. In Taylor, the average difference of the correlation coefficient between classes is used. Finally in Park, the coefficients provided by lasso are used as gene sets score, which are subsequently propagated to the genes in the cluster. In FERAL, genes are scored based on the coefficients of the SGL. To take into account the contribution of the meta-genes, gene scores are supplemented by the largest coefficient of the meta-genes in which it occurs. If a gene receives multiple scores, which is possible due to overlapping gene sets, the scores are averaged (see Supplementary Section S5 for more details).

2.6 Implementation of FERAL

The following steps are taken to train FERAL (Fig. 3). Initially, for all genes, nine of its closest neighbors are selected based on a gene network. After z-score normalization of the expression data, meta-genes are computed. Next, the SGL classifier is trained using the training samples. Implementation of the SGL in this work is based on SLEP (Liu *et al.*, 2009). We further added a wrapper around this package to implement sample weighting to mitigate unbalanced classes along with a search for estimating the optimal parameters using an inner cross-validation. The parameters λ_1 and λ_2, which control the sparsity at the group level and within the groups, respectively, are determined by the inner cross-validation. Finally, the performance of the current fold is determined using the area under the ROC curve (AUC) measure.

3 Results and discussion

For evaluation of FERAL, we use the ACES (Staiger *et al.*, 2012), a cohort of 1606 breast cancer samples collected from 12 studies in NCBIs Gene Expression Omnibus (see Supplementary Section S7 for details). The label for each patient corresponds to recurrence free survival time with respect to a 5-year threshold (good versus poor outcome). Three different networks are used in the evaluation: I2D, a PPI network that is also employed in Staiger *et al.*, a co-expression network and a random network. The co-expression network was defined on training data only and thresholded at a correlation of 0.6. To produce the random network, we shuffle the nodes in the

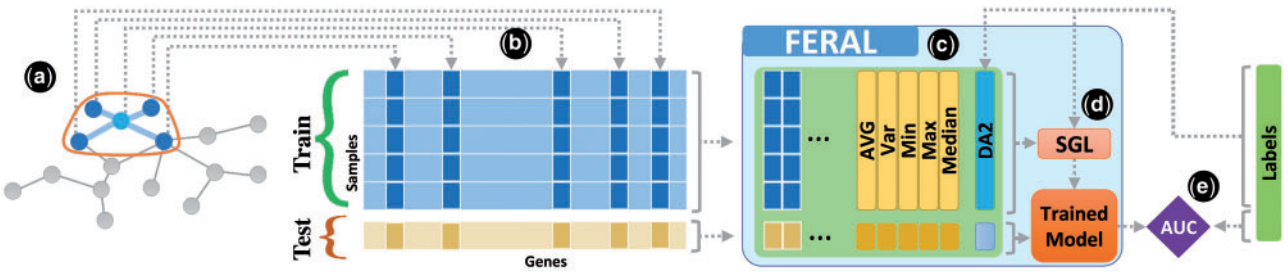

Fig. 3. Schematic of the training and testing procedures of FERAL. (**a**) In the first step, 10 genes are selected using given network. (**b**) Corresponding genes in expression dataset are selected and normalized using *z*-score. (**c**) Meta-genes are computed using the expression profiles of the gene set and target label (in case of a supervised integration). The expression of the individual genes is retained within the gene set. (**d**) The SGL is trained using training samples. (**e**) Test samples are used to assess the prediction performance (in terms of AUC) in the current fold

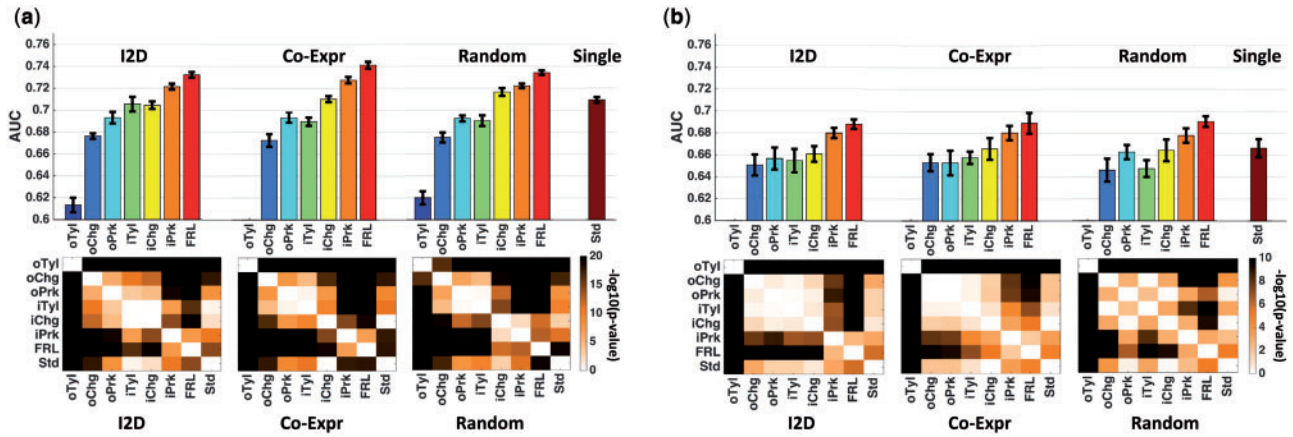

Fig. 4. Performance evaluation (AUC). Performance of the methods under study for the PPI network (I2D), a co-expression network (Co-Expr) and a random network (Random). We also added the result when a classical Lasso is employed (Single). Error bars denote the 95% confidence interval. The heatmaps indicate the *P* value of the paired *t*-test between pairwise comparison of the AUCs of the individual CV folds. (**a**) Sub-type stratified CV. (**b**) Sampled leave-one-study-out CV

I2D network to destroy any biological knowledge while keeping its structure.

We used AUC as the main measure of performance throughout the article. Two types of cross-validation are considered. In the first type (sub-type stratified CV), the ratio of breast cancer sub-types is kept constant in the training and test set. In the second type (sampled leave-one-study-out CV), half of the samples in each study is randomly selected (with replacement), while all samples from one study are excluded from selection and kept hidden as a test set. This configuration forms 12-folds, equal to the number of studies available in ACES. For both cross-validations, the indices of training and testing samples in each fold are kept identical across all methods.

3.1 Performance comparison

Figure 4a shows the obtained average AUCs for 10 repeats of the subtype stratified CV. As a first observation, we note that the improved versions of the standard methods offer better performance. This improvement is most notable for Park's method, which achieves this improved performance despite the fact that the clearly suboptimal average operation was used to construct meta-genes. This points to a relatively small impact of the integration step on the performance of NOPs, an observation that can also be drawn from Figure 2b.

Secondly, as a general trend, all methods produced a similar performance using the random network compared with the case where biologically relevant networks are used. The only exception is the oPrk method, which performs slightly better when a random network is used. The negligible positive contribution of biologically relevant networks on performance of NOPs has been previously observed (Ein-Dor *et al.*, 2005; Staiger *et al.*, 2012, 2013). The most likely explanation for this is the presence of large number of genes that are correlated with the target label which, in turn, makes it possible to construct many alternative features with comparable performance (Ein-Dor *et al.*, 2005; Venet *et al.*, 2011). This points to a limited influence of the selection step on predictive power of NOPs.

Based on these two observations, it can be concluded that the most important factor in the performance improvement is the simultaneous selection and integration achieved by the Lasso. This is most clearly demonstrated by comparing oChg with iChg in Figure 4, as employment of the Lasso is the only difference between these methods.

Even though existing methods can easily be improved by including a simultaneous selection and integration step, we observe that FERAL still offers superior performance across all three networks considered. This performance improvement is very significant (P value $< 7 \times 10^{-8}$; paired t-test with the best other method iPrk using the co-expression network). This demonstrates that, on top of the SGL approach, it is beneficial to provide the classifier with a rich collection of meta-genes based on different aggregation strategies.

Figure 4b shows the results for 10 repeats of the sampled leave-one-study-out CV. As expected, all classifiers showed performance reduction, but the general trends remain the same, that is the

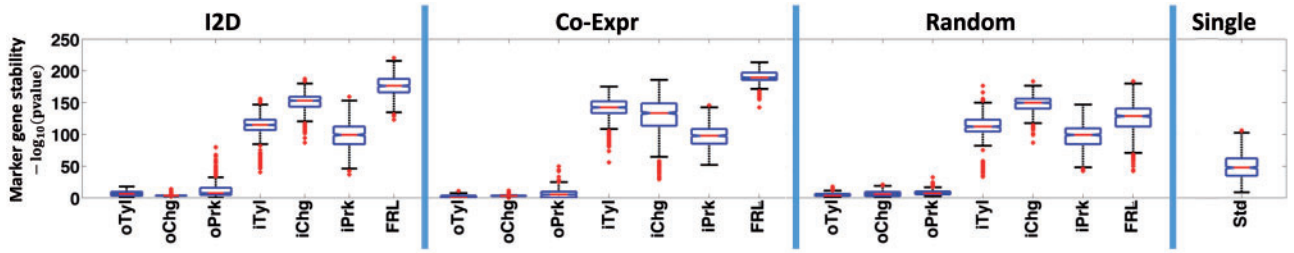

Fig. 5. Stability measurement (using Fisher's exact test) for three different networks including I2D, Co-Expr and random network. The original version of the standard methods produced a much a lower overlap between folds due to pre-ranking of meta-genes. Similarly, Lasso produced a low overlap due to random selection of correlated features. FERAL obtained a higher gene set stability across folds for the I2D and Co-Expr network

standard methods performed poorly compared with their improved counterparts and FERAL significantly outperforms all other classifiers. It should be noted that, although FERAL achieves a better overall performance, the overall classification performance improvement stay relatively modest. It is likely that there is a limit on the maximum performance that can be achieved for the problem at hand (∼70% AUC). This is in line with previous observations (Staiger *et al*., 2013; van Vliet *et al*., 2008; Venet *et al*., 2011).

3.2 Stability of marker genes

Finding robust marker genes is one of the key challenges in breast cancer research as prognostic gene signatures identified in independent datasets often show little to no overlap. To assess how FERAL and the (improved) standard methods perform in terms of signature stability, we follow Staiger *et al.* and assess the stability of selected gene across folds by means of a Fishers exact test. To this end, we measured the overlap between the top 100 genes selected by each of the methods in every fold (see Supplementary Section S5 for details on these score functions). The leave-one-study-out CV was used without subsampling, resulting in a 12-fold cross-validation and the same initial genesets were used in each fold.

Figure 5 shows boxplots of the marker gene stability for all pairwise comparisons between the 12-folds. It is striking to see that FERAL and the improved standard methods clearly have better marker gene stability compared with the standard methods (least significant *P* value: 1.7×10^{-52}), which perform poorly, irrespective of the network employed. For the oChg and oTyl methods, this can be explained by the fact that only very few meta-genes are used in the classifier, which apparently vary substantially between folds. The poor consistency for the oPrk method is caused by a combination of variability of the linkage tree and unstable regression coefficients resulting from the Lasso.

The concordance is highest for FERAL, which even has significantly improved marker gene stability compared with the improved standard methods (least significant *P* value: 1.8×10^{-10}). This demonstrates that FERAL's approach to refrain from a pre-filtering of top genes or gene sets and providing the embedded feature selection of SGL with all genes and many meta-genes using different operators is beneficial for marker gene stability.

Marker gene stability is also improved compared with the single gene classifier. This method performs a Lasso using all genes as predictors and therefore also no pre-filtering is applied in this method. Nevertheless, the overlap of marker genes between folds is still much lower than that obtained with FERAL (*P* value: 5.3×10^{-53}). One explanation is that Lasso randomly selects features if they are highly correlated (Grave *et al.*, 2011). Another reason is that in different samples, separate—yet functionally related—genes play the strongest role in predicting the outcome. As a result, in any subset of the data, different marker genes will be selected. FERAL (and to some extent also the improved standard methods) are able to mitigate this by exploiting network information and summarize functionally related or interacting genes into meta-genes. This is supported by the observation that marker gene stability is significantly reduced when the random network is used (*P* value: 1.6×10^{-29}). For the improved standard measures there is no significance different in case the random network is used. Thus, although utilizing network information does not improve performance substantially, it is helpful in producing more stable sets of marker genes.

3.3 Functional enrichment of marker genes

If an NOP attains reasonable and robust performance and the marker genes selected across the folds are stable, the selected genes may be amenable to interpretation. This facilitates improved understanding of the underlying aberrant processes that play a role in this complex disease. To assess whether the methods under study are capable of detecting relevant genes, we evaluate the concordance of sets of known cancer-related genes with the ranked set of genes produced by each methods under study using the AUC measure. For this purpose, all genes are ranked based on the average score across all folds and repeats of the leave-one-study-out cross-validation (see Supplementary Section S5 for details). For a comprehensive evaluation, we included a ranking based on the individual predictive power of genes (indicated by Ind*) and further a random ranking of genes (indicated by Rnd*). We performed functional enrichment based on a collection of nine cancer-related gene sets, including six cancer-related GO terms (see Supplementary Section S8 for a complete list of these genes within each set).

The observed enrichments obtained using the I2D network are depicted in Figure 6a. The results show that all methods have very modest enrichments not exceeding 0.6 for all but one cancer-related gene set. The notable exception is the enrichment obtained with FERAL, which is vastly superior and close to 0.7 for most cancer-related gene sets and 0.75 for two of them. The enrichment obtained using the Ind* ranking is generally poor, which confirms that differential expression analysis is unsuitable for finding genes involved in the disease. Surprisingly, we observed a severe reduction of gene enrichment using the co-expression network for all methods (see Supplementary Section S6). This corroborates previous findings that PPI networks capture regulatory interaction and functional relations (Kelley and Ideker, 2005).

Taken together, these observations support those made in Sections 3.1 and 3.2, that is incorporating network information does not greatly improve performance, but it does contribute to stabilizing the marker gene sets and finding the biologically relevant genes.

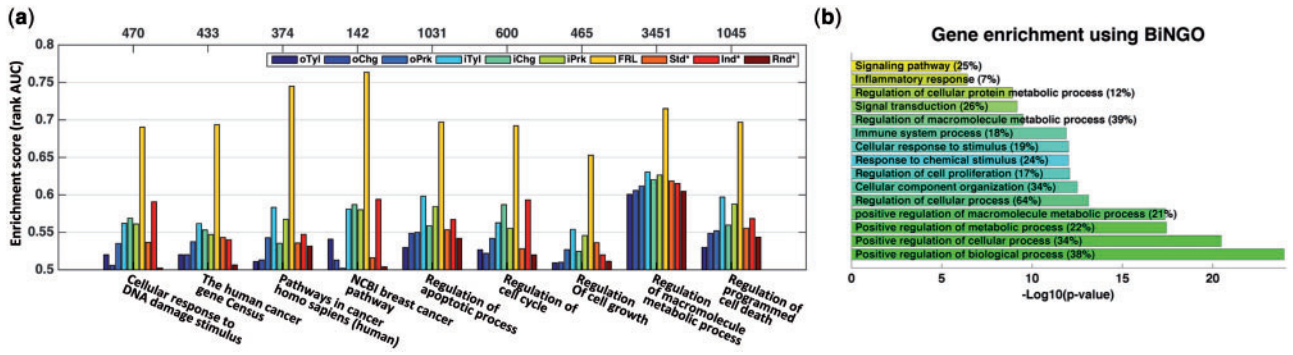

Fig. 6. Gene enrichment. (**a**) Gene enrichment of top genes for each method when the I2D network is employed. The values on top of each group represent the number of genes in each gene set. A notably increased enrichment is obtained using the gene sets produced by FERAL. (**b**) Result of top 15 gene enrichments by BiNGO applied to top 400 genes provided by FERAL

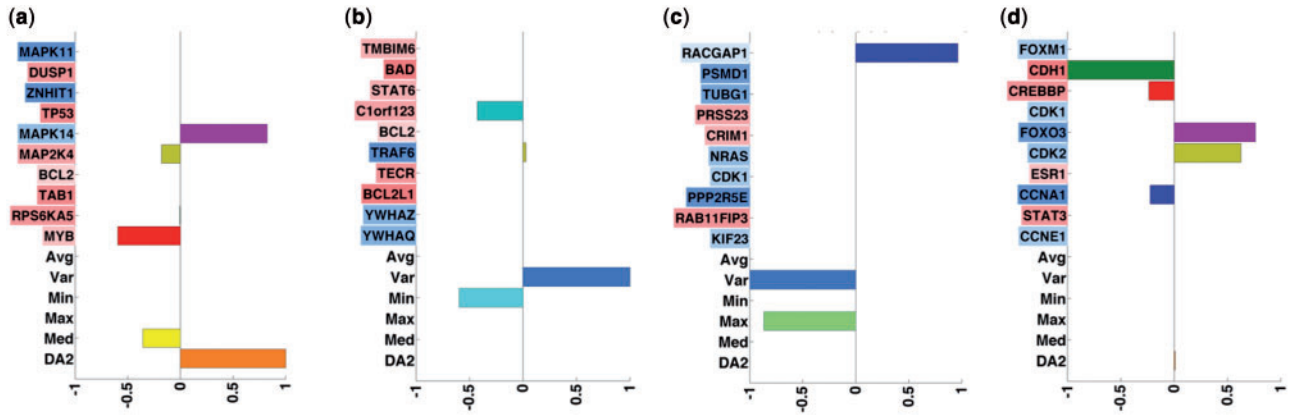

Fig. 7. Frequently identified gene sets by FERAL. The bars represent the median coefficient across folds, normalized to the range $\{-1, 1\}$. Background colors indicate the correlation with target label ranging from positive (blue) to negative (red)

Finally, we used BiNGO (Maere *et al.*, 2005) to determine enrichment across all available gene sets. The hypergeometric test with a Benjamini–Hochberg false discovery rate of 5% is performed for detecting overrepresentation of the top 400 genes in the GO_Biological_Process category. The top 15 most enrichment GO categories are summarized in Figure 7b. Most significant enrichments are observed in various functional categories related to regulation, signaling and proliferation. This finding suggests that FERAL is able to uncover a wide diversity of genes that may play a role in the processes underlying breast cancer metastasis.

3.4 Interpretation of meta-genes in frequently selected networks

Next, we investigated the selected gene sets and meta-genes by FERAL and determine whether they provide new insights into the mechanisms of breast cancer metastasis. To this end, we trained FERAL using the leave-one-study-out CV and obtained optimized λ_1 and λ_2. In this model, still about 1000 gene sets received non-zero coefficients. In an effort to reduce this further, while retaining the most essential ones, λ_2 was increased until the number of selected gene sets was less than 100 in each fold. The majority (66) of the selected gene sets were selected in at least 10 of the 12-folds, demonstrating the stability of selected gene sets across studies. These 66 gene sets were then investigated for relevance to breast cancer in general and metastasis in particular (see Supplementary Section S9 for a complete list).

We performed gene set enrichment for all 66 gene sets using BiNGO (see Supplementary Section S10 for a complete list). The majority of gene sets (94%) were enriched (hypergeometric test with a Benjamini–Hochberg false discovery rate of 5%) for key processes involved in cancer development, such as signaling of cell growth and survival, (regulation of) cell cycle, cell division, proliferation and apoptosis. This shows that FERAL is able to retrieve coherent sets of genes that are involved in cancer. We observe that for all gene sets, at least one of the genes was selected as a predictor in the final model. In the complete set of 66 gene sets, there were 11 that exclusively used expression of individual genes. This corroborates the finding that it is important to supply the classifier with the actual expression profiles of the genes (Babaei *et al.*, 2011; Van den Akker *et al.*, 2011).

Figure 7 displays four of the selected gene sets, along with their median coefficient across the folds (horizontal bars) and association of the individual genes with the survival label (shading behind the gene names). In all four gene sets (and in 83% of all gene sets), a meta-gene obtained a non-zero coefficient. In three cases (and in 62% of all gene sets), even more than one meta-gene was selected. This demonstrates the importance of including multiple summarizations of the gene expression in addition to expression profiles of the genes. Finally, we note that the simple, yet effective, DA2 operator was selected in gene set (a). This was the case in 33% of all gene sets. Taken together, we observe that the final predictor was able to exploit both the raw gene expression profiles and a number of carefully constructed meta-genes.

Next we investigated each of the gene sets in Figure 7 using ingenuity pathway analysis (IPA). Gene set (a) is strongly enriched for p38 MAPK signaling (P value $= 1.4 \times 10^{-14}$). There is ample evidence to suggest that MAPK signaling plays an important role in breast cancer, specifically through Notch regulation (Izrailit *et al.*, 2013). Interestingly, among the genes in this gene set is P53, which typically is not detected through differential expression analysis (Chuang *et al.*, 2007). In this gene set, P53 is also not directly selected but is included in the final prediction model through the meta-genes that are constructed using the DA2 and Median constructors. IPA also suggested a strong involvement of these genes in proliferation of T-lymphocytes (P value $= 1.5 \times 10^{-12}$). This is of particular interest as tumor-infiltrating lymphocytes may be a good biomarker and have recently been implicated in predicting response to neoadjuvant chemotherapy in breast cancer (Mao *et al.*, 2014).

Gene set (b) was most enriched for PI3K/AKT signaling (P value $= 8.4 \times 10^{-8}$), which is one of the major pathways directly related to proliferation and cancer and for which there exist promising therapeutic intervention possibilities (Davis *et al.*, 2014). For the genes in gene set (c), IPA revealed a strong enrichment for breast cancer regulation by stathmin1, a downstream target of CDK1, which is included in gene set (c) (P value $= 1.4 \times 10^{-6}$). This gene set also included RACGAP1, which was recently shown to have prognostic significance in high-risk early breast cancer (Pliarchopoulou *et al.*, 2013). Finally, the gene set (d) was significantly enriched for estrogen-mediated S-phase entry (P value $= 2.9 \times 10^{-13}$). Estrogen is strongly implicated in breast cancer risk due to its role in promoting division of breast cells (Foster *et al.*, 2001).

4 Conclusion

In this work, we proposed a network-based outcome prediction method FERAL that exploits network information in molecular classification of breast cancer outcome. Our method deviates from traditional NOPs in two important aspects. First, FERAL includes several different integration strategies to construct meta-genes, including a novel supervised integration strategy. Our results indicate that the final classification model frequently uses meta-genes produced by these constructors, often even multiple meta-genes based on the same gene set. This underscores the importance of extending traditional meta-genes based on a simple average. The second important improvement is that FERAL performs simultaneous selection and training of the classifier by employing the SGL. This mitigates the need for pre-ranking of genes and/or meta-genes, which is likely to severely reduce the stability of selected genes.

FERAL reached a significant performance increase compared with all standard NOPs, including those that contained significant improvements made by us. This improvement was also obtained using a random network, leading to the conclusion that the biological knowledge encoded in the network is not used to obtain these improvements. The stability of marker genes improves substantially as a result of the procedure implemented in FERAL. This improvement was not observed when the random network was used, indicating that the biological knowledge contributes to the stability of the gene signatures. This improvement was exclusively observed for the PPI network and not for the co-expression network.

Because FERAL attains robust performance and stable marker gene selection, the selected genes and gene sets might reveal insight into the underlying aberrant processes that play a role in this complex disease. We find that almost all the gene sets used in the final model were enriched for cancer related processes. The four gene sets that were studied in more detail revealed very strong suggestive evidence for their involvement in breast cancer, with clear links to MAPK, PI3K and AKT signaling and regulation by stathmin1. In summary, although classification performance of breast cancer outcome obtained with NOPs is unlikely to improve beyond ~70% AUC, we have shown that FERAL achieves much more stable marker gene selection that enables valuable mechanistic insight into the etiology of breast cancer.

Funding

This work was carried out on the Dutch national e-infrastructure with the support of the SURF Foundation. J.d.R. was supported by the Netherlands Organisation for Scientific Research (NWO-Veni: 639.021.233).

Conflict of Interest: none declared.

References

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell Sci.,* **118**, 4947–4957.

Babaei,S. *et al.* (2011) Integrating protein family sequence similarities with gene expression to find signature gene networks in breast cancer metastasis. In: Loog,M. *et al.* (eds), *6th IAPR International Conference, Pattern Recognition in Bioinformatics (PRIB)*. Springer-Verlag Berlin Heidelberg, Delft, The Netherlands, pp. 247–259.

Chen,G. *et al.* (2002) Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Stat. Sin.,* **12**, 241–262.

Cheng,W. *et al.* (2014) Graph-regularized dual lasso for robust eqtl mapping. *Bioinformatics,* **30**, i139–i148.

Chuang,H.-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.,* **3**, 140.

Cun,Y. and Frohlich,H. (2012) Prognostic gene signatures for patient stratification in breast cancer—accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics,* **13**, 69.

Cusick,M.E. *et al.* (2008) Literature-curated protein interaction datasets. *Nat. Methods,* **6**, 39–46.

Dao,P. *et al.* (2010) Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics,* **26**, i625–i631.

Davis,N.M. *et al.* (2014) Deregulation of the egfr/pi3k/pten/akt/mtorc1 pathway in breast cancer: possibilities for therapeutic intervention. *Oncotarget,* **5**, 4603–4650.

Ein-Dor,L. *et al.* (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics,* **21**, 171–178.

Fantozzi,A. and Christofori,G. (2006) Mouse models of breast cancer metastasis. *Breast Cancer Res.,* **8**, 212.

Foster,J.S. *et al.* (2001) Multifaceted regulation of cell cycle progression by estrogen: regulation of cdk inhibitors and cdc25a independent of cyclin d1-cdk4 function. *Mol. Cell. Biol.,* **21**, 794–810.

Friedman,J. *et al.* (2010) A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Grave,E. *et al.* (2011) Trace lasso: a trace norm regularization for correlated designs. In: Shawe-taylor,J. *et al.* (eds), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*, Neural Information Processing Systems. pp. 2187–2195.

Guyon,I. *et al.* (2006) *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell,* **100**, 57–70.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell,* **144**, 646–674.

Hua,J. *et al.* (2009) Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognit.,* **42**, 409–424.

Izrailit,J. *et al*. (2013) High throughput kinase inhibitor screens reveal trb3 and mapk-erk/tgf pathways as fundamental notch regulators in breast cancer. *Proc. Natl. Acad. Sci. U S A,* **110**, 1714–9.

Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.,* **23**, 561–566.

Lazar,C. *et al*. (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.,* **14**, 469–490.

Lee,E. *et al*. (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.,* **4**, e1000217.

Liu,J. *et al*. (2009) *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, http://www.public.asu.edu/~jye02/Software/SLEP.

Maere,S. *et al*. (2005) Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics,* **21**, 3448–3449.

Mao,Y. *et al*. (2014) The value of tumor infiltrating lymphocytes (tils) for predicting response to neoadjuvant chemotherapy in breast cancer: a systematic review and meta-analysis. *PLoS One,* **9**, e115103.

Park,M.Y. *et al*. (2007) Averaged gene expressions for regression. *Biostatistics,* **8**, 212–227.

Pliarchopoulou,K. *et al*. (2013) Prognostic significance of racgap1 mRNA expression in high-risk early breast cancer: a study in primary tumors of breast cancer patients participating in a randomized hellenic cooperative oncology group trial. *Cancer Chemother. Pharmacol.,* **71**, 245–55.

Pujana,M.A. *et al*. (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.,* **39**, 1338–1349.

Shapiro,C.L. and Recht,A. (2001) Side effects of adjuvant treatment of breast cancer. *N. Engl. J. Med.,* **344**, 1997–2008.

Shen,R. *et al*. (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics,* **5**, 94.

Simon,N. *et al*. (2013) A sparse-group lasso. *J. Comput. Graphical Stat.,* **22**, 231–245.

Soneson,C. *et al*. (2014) Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One,* **9**, e100335.

Staiger,C. *et al*. (2012) A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PloS One,* **7**, e34796.

Staiger,C. *et al*. (2013) Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.,* **4**, 289.

Symmans,W.F. *et al*. (1995) Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Hum. Pathol.,* **26**, 210–216.

Taylor,I.W. *et al*. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.,* **27**, 199–204.

Van De Vijver, M.J. *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.,* **347**, 1999–2009.

Van den Akker, E.B. *et al*. (2011) Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis. *J. Integr. Bioinform.,* **8**, 188.

van Vliet, M.H. *et al*. (2008) Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics,* **9**, 375.

van't Veer, L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* **415**, 530–536.

Venet,D. *et al*. (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.,* **7**, e1002240.

Von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature,* **417**, 399–403.

Weigelt,B. *et al*. (2005) Breast cancer metastasis: markers and models. *Nat. Rev. Cancer,* **5**, 591–602.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B (Stat. Methodol.),* **68**, 49–67.