OPEN ACCESS Freely available online

PLOS ONE

Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics

Yupeng Cun, Holger Fröhlich*

Algorithmic Bioinformatics, Bonn-Aachen International Center for IT, Bonn, Germany

Abstract

Predictive, stable and interpretable gene signatures are generally seen as an important step towards a better personalized medicine. During the last decade various methods have been proposed for that purpose. However, one important obstacle for making gene signatures a standard tool in clinics is the typical low reproducibility of signatures combined with the difficulty to achieve a clear biological interpretation. For that purpose in the last years there has been a growing interest in approaches that try to integrate information from molecular interaction networks. We here propose a technique that integrates network information as well as different kinds of experimental data (here exemplified by mRNA and miRNA expression) into one classifier. This is done by smoothing t-statistics of individual genes or miRNAs over the structure of a combined protein-protein interaction (PPI) and miRNA-target gene network. A permutation test is conducted to select features in a highly consistent manner, and subsequently a Support Vector Machine (SVM) classifier is trained. Compared to several other competing methods our algorithm reveals an overall better prediction performance for early versus late disease relapse and a higher signature stability. Moreover, obtained gene lists can be clearly associated to biological knowledge, such as known disease genes and KEGG pathways. We demonstrate that our data integration strategy can improve classification performance compared to using a single data source only. Our method, called stSVM, is available in R-package netClass on CRAN (http://cran.r-project.org).

Citation: Cun Y, Fröhlich H (2013) Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics. PLoS ONE 8(9): e73074. doi:10.1371/journal.pone.0073074

Editor: Stefano Boccaletti, Technical University of Madrid, Italy

Received June 4, 2013; **Accepted** July 16, 2013; **Published** September 3, 2013

Funding: YC was supported by the state of NRW via the B-IT research school. HF is a member of the excellence cluster ImmunoSensation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: frohlich@bit.uni-bonn.de

Introduction

One of the major goals of personalized medicine is to identify reliable molecular biomarkers that predict relevant clinical characteristics for an individual patient, like disease sub-type, his/her response to a certain therapy or survival time. Prognostic and diagnostic biomarker signatures can nowadays be constructed on the basis of multiple molecular data, such as gene expression data, miRNA, methylation or copy number alterations [1].

A common approach to obtain a signature for diagnostic or prognostic purposes is to put patients into distinct groups and then construct a classifier that can discriminative patients in the training set and is able to predict well unseen patients. Frequently applied algorithms are Prediction Analysis for Microarrays (PAM) [2], Support Vector Machine Recursive Feature Elimination (SVM-RFE) [3], Random Forests [4] or statistical tests, like Significant Analysis for Microarrays (SAM) [5], combined with machine learning techniques (SVM, k-NN, linear discriminant analysis, logistic regression,…) [6,7]. However, a commonly encountered problem is that molecular signatures are often not reproducible in the sense that inclusion or exclusion of a few patients can lead to quite different sets of selected features. Moreover, these sets are often difficult to interpret in a biological way [8]. Both issues currently prevent molecular signatures to become a standard tool in clinical practice [9]. For that reason, various network based approaches have been proposed to integrate prior knowledge on canonical pathways, Gene Ontology (GO) annotation or protein-protein interactions into feature selection algorithms [10–17]. A recent review on such approaches can be found in [18]. The general hope of these approaches is that biological knowledge can lead to better interpretable and more stable signatures. Whether network based classification methods automatically also lead to higher prediction accuracies is still a matter of debate [19,20].

Another line of research focuses on the integration of different entities of experimental data for the same patient, e.g. mRNA and miRNA expression [21–24]. The increasing amount of different kinds of molecular data from the same patient, for instance within the TCGA database (www.cancergenome.nih.gov), now opens the door to a broader disease understanding [25–27]. Moreover, the integration of data capturing different molecular mechanisms could also lead to improved molecular signatures.

In this paper we propose a filter based feature selection approach, which integrates network information by smoothing gene-wise t-statistics over the graph structure using a random walk kernel. Our approach allows for a straight forward integration of different data entities, like mRNA and miRNA expression. Comparisons of our smoothed t-statistic SVM (stSVM) with several competing approaches on a breast cancer, two prostate cancer and an ovarian cancer dataset demonstrate a favorable prediction performance of early versus late relapse and a high signature stability. Moreover, obtained gene lists are highly enriched with known disease genes and KEGG pathways.

Materials and Methods

Datasets

We retrieved one breast cancer [28], two prostate cancer [29,30] and one ovarian cancer [26] dataset from different data repositories. The breast cancer [28] and one of the prostate cancer datasets [29] were measured on Affymetrix HGU133 microarrays. The second prostate cancer dataset (MSKCC, [30]) and the ovarian cancer dataset (TCGA, [26]) were measured on Affymetrix HuEx 1.0 ST microarrays. The breast and first prostate cancer dataset were normalized via FARMS [31]. The ovarian cancer and MSKCC datasets were downloaded as ready normalized and gene-wise aggregated data from the TCGA and MSKCC homepage, respectively. As clinical end points we considered metastasis free (breast and prostate cancer) and relapse free (ovarian cancer) survival time after initial clinical treatment. For ovarian cancer only tumors with stages IIA - IV and grades G2 and G3 were considered, which after resection revealed at most 10 mm residual cancer tissue and responded completely to initial chemotherapy.

Survival time information was dichotomized into two classes according whether or not patients suffered from a reported relapse/metastasis event within 5 years (breast, prostate dataset 1), 3 years (MSKCC prostate cancer dataset) and 1 year (ovarian), respectively. Patients with a survival time shorter than 5/3/1 year(s) without any reported event were not considered and removed from our datasets. This was done, because these patients can neither reliably be put into the early nor into the late relapse class. A summary of our datasets can be found in Table 1.

Network Information

Protein-Protein Interactions (PPI). A comprehensive protein interaction network was compiled from the Pathway Commons database [32], which was downloaded in tab-delimited format (September 2012). All SIF interactions INTERACTS_WITH and STATE_CHANGE were taken into account (http://www.pathwaycommons.org/pc/sif_interaction_rules.do) and self loops removed, resulting in a large network with 11,361 nodes and 610,185 edges. Nodes in this network were identified with Entrez gene IDs. Expression values for probesets on the microarray that mapped to the same gene in the network were averaged. In order to consider genes with available probesets on the array but no corresponding network information we added for all these genes unconnected nodes to our initial network, resulting in 12,611 nodes for breast and the Sun et al. prostate cancer dataset; 11,356 nodes for ovarian cancer and 11,322 nodes for the MSKCC prostate cancer dataset. The reason for these differences is that not all dataset contain the same number of mappable transcripts.

KEGG pathways. As an alternative network information we computed a merger of all non-metabolic KEGG pathways [33]. For retrieval and merger of KEGG pathways, we employed the R-package KEGGgraph [34]. Only gene-gene interactions were considered, which resulted in an initial network with 3,087 nodes and 17,518 edges. As before this initial network was extended to contain all genes available on the array, resulting in an overall network with the same number of nodes as described above for the PPI network but a different number of edges.

miRNA-Target gene network. In addition to PPI or KEGG pathway information we optionally included predicted miRNA-target gene interactions. Target predictions were obtained from the MicroCosm database (version 5) [35] (FDR cutoff 1%). This increased the number of edges in the PPI network to 11,892 nodes for MSKCC's prostate cancer and 11,839 nodes for ovarian cancer.

Prediction Performance, Signature Stability and Biological Interpretability

In order to assess the prediction performance of all tested methods we performed a 10 times repeated 10-fold cross-validation on each dataset. That means the whole data was randomly split into 10 fold, and each fold sequentially left out once for testing, while the rest of the data was used for training and optimizing the classifier (including selection of relevant genes, hyper-parameter tuning, standardization of expression values for each gene to mean 0 and standard deviation 1, etc.). The whole process was repeated 10 times. It should be noted extra that also standardization of gene expression data was only done on each training set separately and the corresponding scaling parameters then applied to the test data.

The area under receiver operator characteristic curve (AUC) was used to measure the prediction accuracy via the R-package ROCR [36]. To assess the stability of gene selection, we computed the selection frequency of each gene within the 10 times repeated 10-fold cross-validation procedure. That means a particular gene could be selected at most 100 times In order to summarize the selection frequencies for all genes we defined a so-called stability index (SI) as

$$SI = \frac{1}{|P|}\sum_{s\in P} h(s) \tag{1}$$

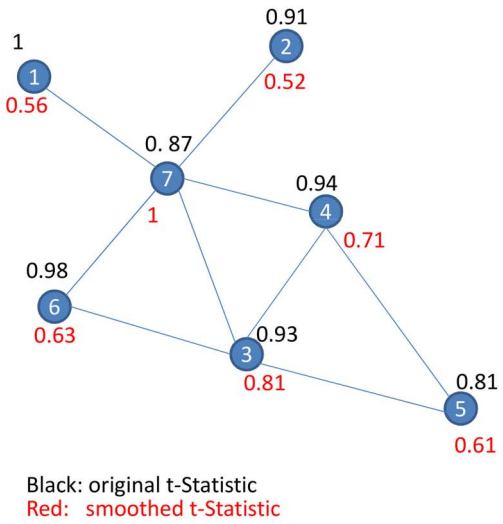
where P is the set of selected genes that had been selected at least once and $h(s)$ is the actual number of times that s was selected. SI represents a weighted histogram count of selection frequencies. Obviously, the larger SI the more stable the algorithm is. In the optimal case $SI = 100$. The SI has to be seen together with the size of gene signature, because trivially a classifier selecting all genes would always achieve $SI = 100$.

In order to check in how far signatures obtained by training the classifier on the whole dataset could be related to existing biological knowledge, we looked for enrichment of disease related genes via the tool FunDO [37] (hypergeometric test; multiple testing correction: Bonferroni's method). Moreover, we calculated

Table 1. Overview about employed datasets.

ID/source	patients	cancer type	classification	positive class
GSE4922	228	breast	metastasis free survival >5 y	69
TCGA	135	ovarian	relapse free survival >1 y	35
GSE21032(MSKCC)	79	prostate	relapse free survival >3 y	29
GSE25136	79	prostate	recurrent vs. non-recurrent	40

doi:10.1371/journal.pone.0073074.t001



One step random walk kernel matrix

node	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0.45
2	0	1	0	0	0	0	0.45
3	0	0	1	0.29	0.35	0.35	0.22
4	0	0	0.29	1	0.41	0	0.26
5	0	0	0.35	0.41	1	0	0
6	0	0	0.35	0	0	1	0.32
7	0.45	0.45	0.22	0.26	0	0.32	1

Figure 1. Toy example to demonstrate the network smoothed t-statistic.
doi:10.1371/journal.pone.0073074.g001

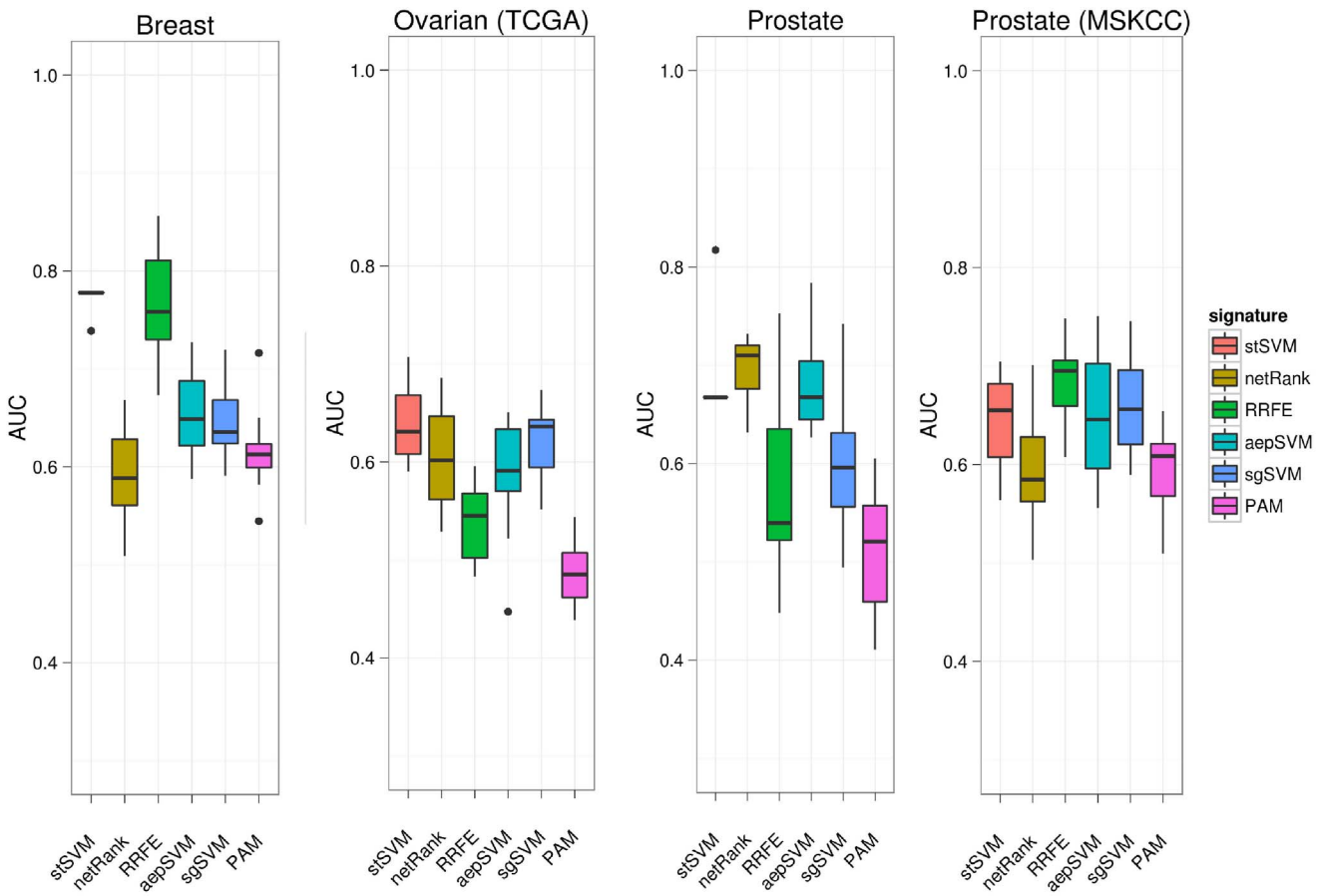


Figure 2. Prediction performance of stSVM in comparison to other methods in terms of area under ROC curve (AUC). Breast = GSE11121, Ovarian (TCGA) = GSE25136, Prostate = GSE25136, Prostate (MSKCC) = GSE21032.
doi:10.1371/journal.pone.0073074.g002

the enrichment with KEGG pathways [33] via a hyper-geometric test.

Network Smoothed T-Statistic SVMs (stSVMs)

Network Smoothed T-Statistics. Given a simple, undirected graph $G=(V,E)$ with adjacency matrix A the graph Laplacian L is defined as $L:=D-A$, where $D=diag(deg(v_1),...,deg(v_n))$ is a diagonal matrix of node degrees for nodes $v_1,...,v_n$ [38]. The graph Laplacian can be viewed as a discrete approximation of the negative Laplace operator for functions.

One way of characterizing the degree of relatedness of two nodes (e.g. proteins) v and w in a graph (e.g. a PPI network) can be obtained via the notion of random walks. The p-step random walk kernel is one particular similarity measure, which can be derived from this notion [39] and is defined as:

$$K=(\alpha I-L^{norm})^p=((\alpha-1)I+D^{-1/2}AD^{-1/2})^p \qquad (2)$$

Here $L^{norm}:=D^{-1/2}LD^{-1/2}=I-D^{-1/2}AD^{-1/2}$ is the normalized graph Laplacian matrix, α is constant, and p is the number of random walk steps (here: $a=1,p=2$). The p-step random walk kernel gives rise to a symmetric, positive semi-definite similarity matrix between network nodes, capturing their degree of topological relatedness. The advantage compared to shortest path distance based measures is that alternative routes between pairs of nodes are considered. That means, if v and w are connected via many alternative paths of the same length this marks a higher similarity than if there exists only one such path.

Suppose for each network gene we assess its differential expression on the training dataset via a t-test. This results in an absolute t-statistic $|t_i|$ for network node i. We summarize the $|t_i|,i=1,...,|V|$ into a vector $\mathbf{t}$ and consider the score vector

$$\tilde{\mathbf{t}}=\mathbf{t}^T K \qquad (3)$$

Please note that $\tilde{t}_i=\sum_j |t_j|K_{ij}$. Hence, $\tilde{t}_i$ is a network smoothed version of $|t_i|$ (Figure 1), but does not follow a t-distribution any more. We thus conduct a permutation test (here: 1000 times) to obtain a p-value for each gene. For reasons of computation time we restrict this to the 10% genes, which are highest ranked according to the network smoothed t-score (Eq. 3). Multiple testing correction is then performed using the FDR approach by [40].

Table 2. Ranking of different algorithms with respect to the median AUC in a 10 times repeated 10-fold cross-validation procedure.

	breast	ovarian	prostate	prostate MSKCC	consensus
stSVM	1	2	3	3	1
netRank	6	3	1	6	4
RRFE	2	5	5	1	3
aepSVM	3	4	2	4	5
sgSVM	4	1	4	2	2
PAM	5	6	6	5	6

doi:10.1371/journal.pone.0073074.t002

It is worth mentioning that the smoothing of absolute t-statistics particularly affects nodes with a high number of interaction partners. On one hand our procedure aggregates the scores of neighboring nodes to increase the score for these central proteins. On the other hand there is also a reverse effect, which increases the relevance of proteins in close proximity to hubs.

SVM training. We only select genes with FDR $<5\%$. Subsequently a Support Vector Machine (SVM) is trained using the optimal parameter C from $\{0.0001,0.001,...,10000\}$. To evaluate each candidate parameter C we here used the span rule, which provides a theoretical upper bound for the leave-one-out cross-validation error, but can be computed much more efficiently for datasets with few samples [41]. It has been demonstrated theoreticaly as well as empircally that the span-rule provides an excellent mechanism for parameter selection in SVMs [41]. An implementation of this procedure can be found in R-packages pathClass [42] and netClass, which is a supplement to this paper.

Integration of different experimental data. Besides network information our approach allows for a straight forward integrating on of different experimental data, e.g. mRNA and miRNA expression, into one classifier. This can be achieved by extending adjacency matrix A to miRNA-mRNA interactions and vector $\mathbf{t}$ to absolute t-statistics for miRNAs. Accordingly, network smoothing is now performed over protein-protein as well as miRNA-target gene interactions.

Results

stSVM Shows Overall Best Prediction Performance

We initial considered our proposed stSVM method using only gene expression data and PPI network information. We compared the prediction performance to a number of competing methods, namely PAM [2], a SVM trained with significant differentially expressed genes (FDR cutoff 5%) selected by SAM [5] (**sgSVM**), average gene expression of KEGG pathways (**aepSVM** [10]), pathway activity classification (**PAC** [13]), reweighted recursive feature elimination (**RRFE** [17]) and the **netRank** algorithm [43]. NetRank, similar to RRFE, uses a modification of Google's PageRank method to rank genes according to both, expression and network centrality [44]. The optimal number of selected genes in both cases was determined via the span-rule inside the cross-validation procedure [41].

For stSVM, netRank and RRFE, the same large PPI network was used as biological background information. The aepSVM and PAC methods use KEGG pathways. PAC relies on a so-called activity score, which is calculated per individual pathway and then taken as as a feature for classification purposes. For aepSVM we first conducted a global test [45] to select pathways being significantly associated with the class label (FDR cutoff 1%) on the training data and then calculated the mean expression of each selected pathway as a feature for SVM based classification. The prediction of all methods was assessed via a 10 times repeated 10-fold cross-validation procedure, as described in the Materials and Methods part of this paper.

Generally we observed a large variability of prediction performances of most tested algorithms across different datasets, which is in agreement with our previous observations [19]. However, our proposed stSVM approach showed on all of our four gene expression datasets a consistently high prediction performance with respect to the area under ROC curve (AUC, Figure 2) and significantly outperformed several competing methods (Tables S5, S6, S7, S8). Notably on two datasets (breast, prostate dataset 1) the AUC was extremely stable and showed only a very small variance across the cross-validation procedure.

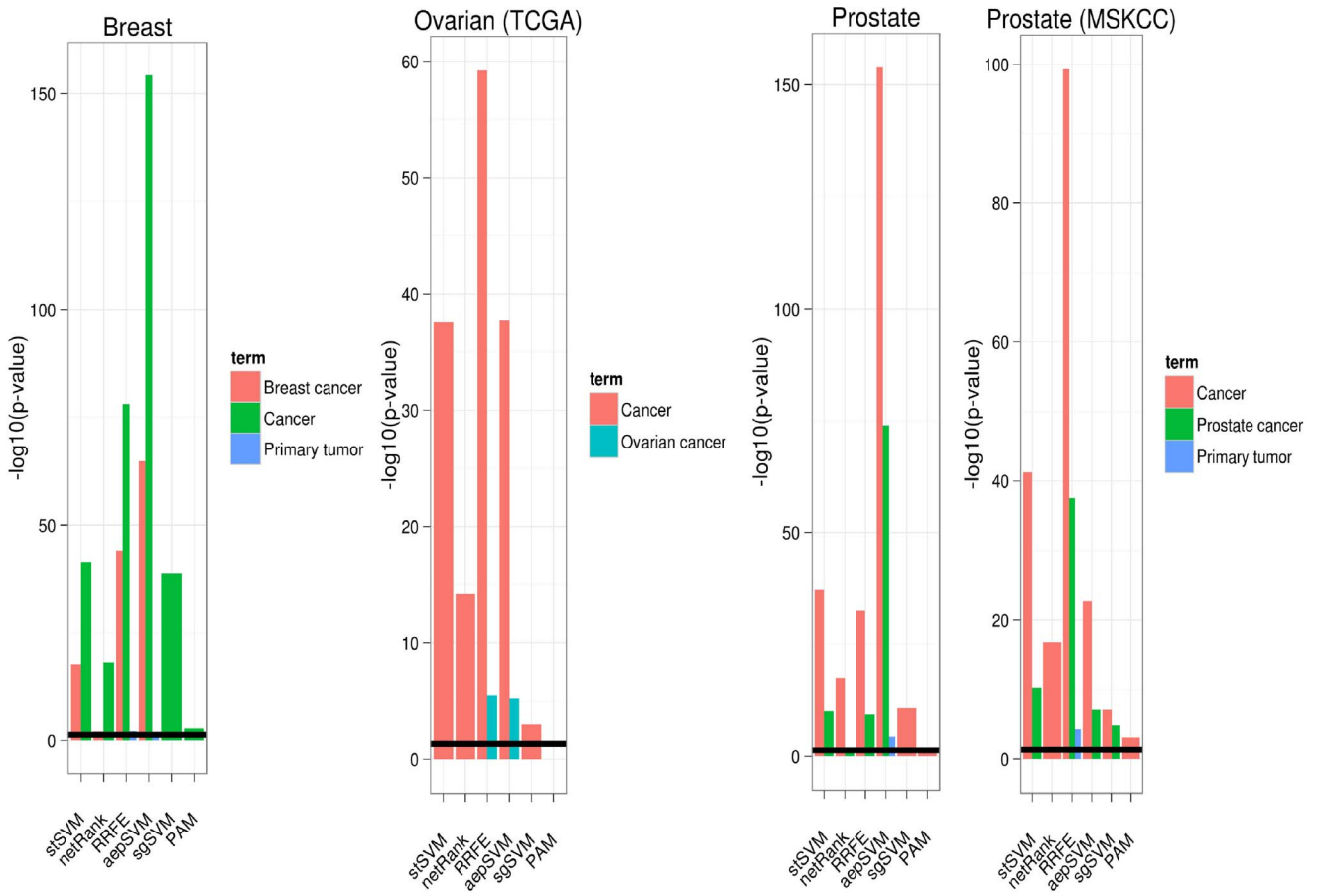


Figure 3. Enrichment of signatures with disease related genes. The y-axis shows -$\log_{10}$ p-values computed via a hypergeometric test (Bonferroni correction for multiple testing). Black horizontal line = 5% significance cutoff.
doi:10.1371/journal.pone.0073074.g003

In order to get a more objective and comprehensive view we conducted a ranking of all methods in each dataset according to the median cross-validated AUC value. We then calculated a consensus ranking using Kendall's τ distance method [46] (Table 2). This confirmed our impression that stSVM was the overall best performing method. Interestingly enough, sgSVM was ranked second highest here, which is in agreement with our earlier finding that network based approaches do not consistently outperform classical ones [19].

stSVM Yields Highly Stable Classification

We investigated the stability of signatures obtained during the 10 times repeated 10-fold cross-validation procedure using the concept of the stability index (Eq. 1), showing for stSVM an extremely robust behavior (Figure S1). Most of the signature probesets were selected consistently during the cross-validation procedure. Interestingly enough, at the same time the number of selected probesets was comparably high for stSVM, which may be attributed to the fact that the network smoothing enforces the selection of correlated genes. Tables S1, S2, S3, S4 show 10 consistently selected genes in each dataset. As expected these genes typically reveal a high node degree in the PPI network. Many of these hub genes are well known to play a role in the disease pathology, e.g. BRCA1 for all tumors [47–49] and AR for prostate cancer [50]. Other disease related and consistently selected genes include p53 (all datasets), EGFR (breast and prostate cancer [51,52]), RB1 (breast and ovarian tumors [53–55]) and EP300 (prostate cancer [56]).

stSVM Shows Clear Association to Biological Knowledge

In order to test the association with existing biological knowledge more systematically we trained each of our tested methods on complete datasets and subsequently tested the resulting signatures (Tables S9, S10, S11, S12 for stSVM, Tables S13 and S14 for stSVM(mi-mRNA)) for enrichment of disease related genes and KEGG pathways (Figures 3, S2). For testing the association with disease related genes we used the FunDO tool [37], which is based on a hyper-geometric test.

Our analysis revealed a high enrichment of signatures obtained via stSVM to known disease genes on all datasets. The enrichment was always higher than for non-network based methods (sgSVM, PAM) as well as for signatures obtained via the netRank algorithm. The latter might be attributed to the fact that netRank typically selects only very few genes, which thus could cause a loss of statistical power for enrichment analysis.

Besides disease related genes we also found a high enrichment of stSVM derived signatures for several KEGG pathways in all datasets (Figure S2). Examples were *Pathways in cancer* (prostate, breast cancer), *Prostate Cancer* (both prostate cancer datasets), *Wnt signaling*, *MAPK signaling* and *ERBB signaling*. The latter three were significant in breast and prostate cancer and are known to play a role in the respective disease pathologies [57–63]. In ovarian cancer we particularly detected a high enrichment of several

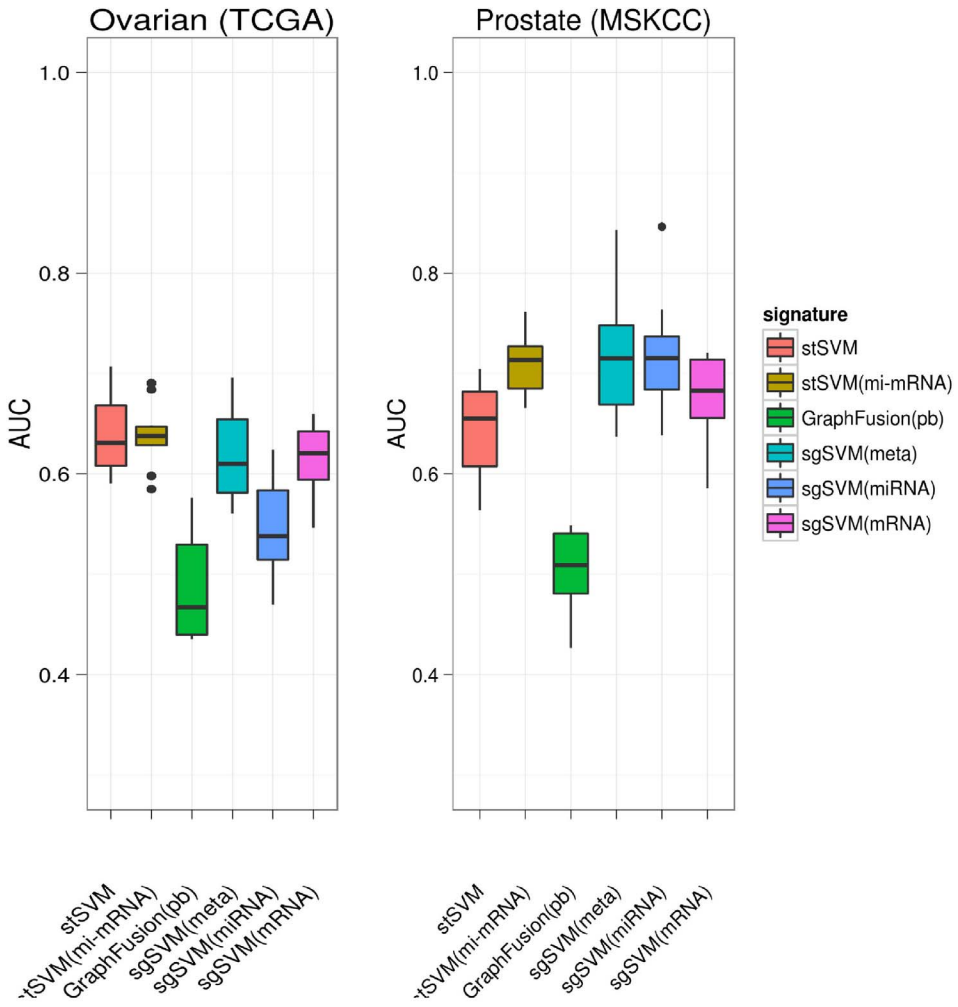


Figure 4. Prediction performance of stSVM on integrated gene and miRNA expression data compared to other approaches.
doi:10.1371/journal.pone.0073074.g004

metabolic pathways, such as *Fatty acid metabolism*. This fits to the fact that adipocytes were recently found to promote rapid tumor growth in ovarian tumors [64]. The significance of enrichment for KEGG pathways was generally higher for stSVM than for all other methods.

Taken together stSVM derived signatures showed a clear association to existing biological knowledge, which eases their biological understanding.

Influence of Network Structure

We asked the question, in how far the observed good prediction performance of stSVM was dependent on the incorporated network structure. We hence re-ran our cross-validation procedure with a different network structure, which was compiled from a merger of all non-metabolic KEGG pathways (see Materials and Methods). It is worthwhile to mention that both networks contained the same number of nodes, but different number of edges. The KEGG derived network was much sparser then the previously used PPI network.

We observed that our original PPI network in all but one case (ovarian cancer dataset) yielded significantly higher AUCs, which highlights the principle influence of the network structure (Figure S3). We can only speculate why on the ovarian cancer dataset the KEGG based network appeared to work at least as good as the PPI network. Principally KEGG pathways capture different biological aspects (canonical pathways) than large scale protein-protein interaction networks. It may be due to the nature of the disease that KEGG pathways reflect better the relevant biology for ovarian cancer than for breast and prostate tumors.

stSVM Allows for mRNA and miRNA Data Integration

Our stSVM method allows for a straight forward integration of different types of experimental data on network level (see Materials and Methods). We here exemplify this property by using gene expression together with miRNA expression data for the TCGA ovarian cancer and for the MSKCC prostate cancer datasets. Correspondingly network information now consisted of a combined PPI and miRNA-target gene network. We call the corresponding variant of our method **stSVM(mi-mRNA)**. We compared stSVM(mi-mRNA) to the graph fusion approach by Gade et al. [21] (**GraphFusion**). In their original paper Gade et al. used CoxBoost [15] to make survival risk prediction. In our classification based framework we replaced CoxBoost by the related PathBoost algorithm [15].

Moreover, we compared stSVM(mi-mRNA) to sgSVM trained on mRNA data only, on miRNA data only and to a meta-classifier, which combines classification outputs from the mRNA and miRNA sgSVM classifiers into one consensus classifier

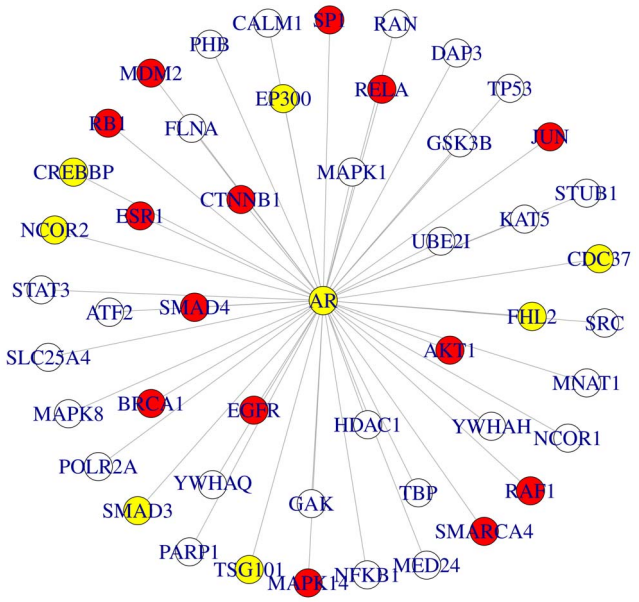


Figure 5. Sub-graph of disease related module identified by stSVM (MSKCC prostate cancer). The shown sub-graph consists of consistently selected genes in the interactome of the AR. For better visualization edges between neighbors of the AR are omitted. Red: cancer related genes; yellow: prostate cancer related genes.
doi:10.1371/journal.pone.0073074.g005

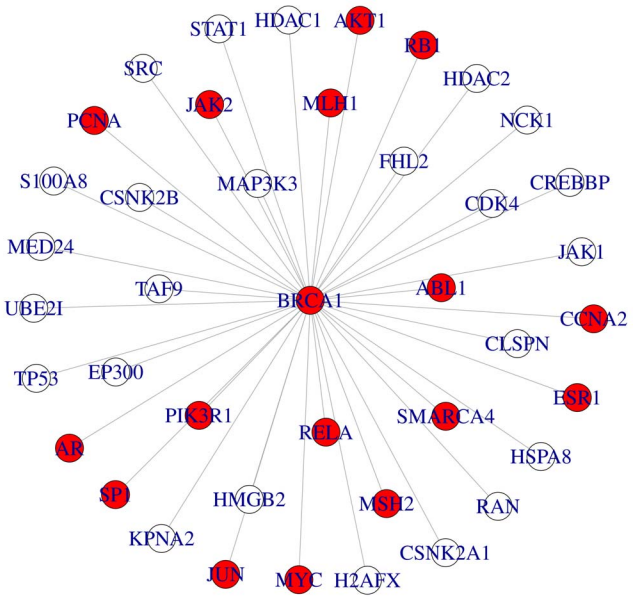


Figure 6. Sub-network of disease related module identified by stSVM (ovarian cancer). The shown sub-graph consists of consistently selected genes in the interactome of the BRCA1. For better visualization edges between neighbors of the BRCA1 are omitted. Red: cancer related genes.
doi:10.1371/journal.pone.0073074.g006

(**sgSVM(meta)**). This was done as follows: The sgSVM method was separately trained on both datasets to yield a linear SVM classifier using significant differentially expressed genes and miRNAs, respectively. Each of these SVM classifiers yields a ranking (not classification) function of the form

$$f(\mathbf{w}) = \sum_{i=1}^{n} \alpha_i y_i \mathbf{w}_i + b$$

where α_i are the fitted Lagrangian multipliers, $y_i \in \{-1,1\}$ the class labels and b the intercept [65]. Note that the corresponding classification function can be obtained by taking the sign of $f(\mathbf{w})$. Let $f_1(\mathbf{x}), f_2(\mathbf{z})$ be the SVM ranking functions for mRNA profile $\mathbf{x}$ and miRNA profile $\mathbf{z}$, respectively. Then both rankings can be combined into a meta-classifier by fitting a logistic regression function

$$\Pr(y_i = 1 | f_1(\mathbf{x}), f_2(\mathbf{z})) = \frac{1}{1 + \exp(-\theta_0 - \theta_1 f_1(\mathbf{x}) - \theta_2 f_2(\mathbf{z}))}$$

where $\theta_0, \theta_1, \theta_2$ are parameters, which can be fitted to the data.

The comparison of our stSVM(mi-mRNA) approach to the graph fusion algorithm as well as to the above described meta-classifier approach (**sgSVM(meta)**) revealed a superior performance of our method. GraphFusion was outperformed with large margin (Figure 4), and the gain compared to sgSVM(meta) was still weakly/moderately significant ($p = 0.065$ for ovarian and $p = 0.041$ for prostate cancer; Wilcoxon signed rank test). In that context it was interesting that only on the prostate cancer dataset a significant improvement by integration of mRNA and miRNA data could be observed at all: The comparison of stSVM(meta) versus stSVM yielded a p-value of 0.008 (Wilcoxon signed rank test). On the ovarian cancer dataset miRNA expression data did not appear to contribute any useful classification information. This is also highlighted by the weak performance of the sgSVM classifier trained only on miRNA expression data (**sgSVM(miRNA)**).

Consistently Selected Features Form Disease Related Network Modules

Taking the set of genes and miRNAs, which were consistently selected by stSVM in the above investigated ovarian and MSKCC prostate cancer datasets, we asked the question, whether these features were connected to each other on network level, indicating that stSVM preferentially selected network connected genes and miRNAs.

To answer this question we looked for the largest sub-network that was purely formed by consistently selected features. In case of the ovarian cancer dataset we found 368 genes and 50 miRNAs out of 377 genes and 235 miRNAs to form such a network module. In case of the MSKCC prostate cancer dataset 384 genes and 96 miRNAs out of 386 genes and 254 miRNAs were inside one network module. This demonstrates that stSVM preferentially selected features, which were connected to each other on network level. The fraction of consistently selected genes that were inside one network module was, however, higher than the corresponding fraction of miRNAs. The reason could be that differential expression of a miRNA does not automatically imply that its target genes are also differentially expressed. Consequently miRNA markers do not always (but still in a significant proportion – see prostate cancer dataset) cluster together with gene markers on network level.

For both, ovarian and prostate cancer, network modules were highly enriched for known disease genes ($p = 4.39e-11$ for *prostate cancer* in MSKCC prostate cancer case, $p = 1.18e-3$ for *ovarian cancer* in ovarian cancer case) according to FunDO. Figure 5 and Figure 6 visualize sub-networks of these modules centered at the

AR (MSKCC prostate cancer) and BRCA1 (ovarian cancer), respectively.

Discussion and Conclusion

In this article we proposed network smoothed t-statistics as a method to integrate network information as well as different types of experimental data into one classifiers for biomarker signature discovery. Our method smoothed a widely used marginal statistic (the t-statistic) for differential expression over the graph structure of a biological network using random walk kernels. Our approach has on the technical level certain similarities with kernel based ranking methods for gene prioritization, which have been proposed e.g. by Moreau and co-workers to predict putative disease causing genes in genetic disorders [66–68]. Note, that this is a rather different problem than finding prognostic biomarker signatures.

We showed that our approach overall leads to a highly predictive, stable and biologically interpretable classifier. We exemplified the straight forward integration of different types of experimental data here by building joint classifiers of gene and miRNA expression data. Other kinds of data (e.g. methylation, copy number variations) could principally be integrated in a similar manner. This is, however, not necessarily straight forward and thus subject to future research.

Taken together we think that our method is a step towards the challenging goal to build integrative classification models, which not only make use of biological background information, but also allow to combine various kinds of molecular data in order to make accurate predictions for an individual patient. In the light of the TCGA project and other large scale efforts the time is now ripe to move into this direction.

Supporting Information

Figure S1 Stability index and signature sizes within the 10 times repeated 10-fold CV procedure. A) stability index according to Eq. (1) in main document,B) number of selected probesets.
(TIF)

Figure S2 Enrichment of signatures with KEGG pathways: Depicted is a heatmap of the -log p-value for the 10 most significant pathways.
(TIF)

Figure S3 Classification performance of stSVM using two different sources of network information.
(TIF)

Table S1 10 consistently selected genes in ovarian cancer dataset.
(XLS)

Table S2 10 consistently selected genes in breast cancer dataset.
(XLS)

Table S3 10 consistently selected genes in prostate cancer dataset (GSE25136).
(XLS)

Table S4 10 consistently selected genes in prostate cancer dataset (MSKCC).
(XLS)

Table S5 False discovery rates resulting from pairwise Wilcoxon signed rank tests to compare AUC values for different classification algorithms: breast cancer dataset.
(XLS)

Table S6 False discovery rates resulting from pairwise Wilcoxon signed rank tests to compare AUC values for different classification algorithms: overian cancer dataset.
(XLS)

Table S7 False discovery rates resulting from pairwise Wilcoxon signed rank tests to compare AUC values for different classification algorithms: prostate cancer dataset (GSE25136).
(XLS)

Table S8 False discovery rates resulting from pairwise Wilcoxon signed rank tests to compare AUC values for different classification algorithms: prostate cancer dataset (MSKCC).
(XLS)

Table S9 Final signatures obtained by stSVM in breast cancer dataset.
(XLS)

Table S10 Final signatures obtained by stSVM in ovarian cancer dataset.
(XLS)

Table S11 Final signatures obtained by stSVM in prostate cancer dataset (GSE25136).
(XLS)

Table S12 Final signatures obtained by stSVM in prostate cancer dataset (MSKCC).
(XLS)

Table S13 Final signatures obtained by stSVM(mi-mRNA) in ovarian cancer dataset.
(XLS)

Table S14 Final signatures obtained by stSVM(mi-mRNA) in in prostate cancer dataset (MSKCC).
(XLS)

Author Contributions

Conceived and designed the experiments: YC HF. Performed the experiments: YC. Analyzed the data: YC HF. Contributed reagents/materials/analysis tools: YC HF. Wrote the paper: YC HF.

References

1. Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, et al. (2012) Cancer genomics:Technology, discovery, and translation. Journal of Clinical Oncology 30: 647–660.
2. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 99: 6567–6572.
3. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46: 389–422.
4. Breiman L (2001) Random forests. Machine Learning 45: 5–32.
5. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116–5121.
6. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23: 2507–2517.
7. Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning. New York, NY, USA: Springer.
8. Gönen M (2009) Statistical aspects of gene signatures and molecular targets. Gastrointestinal cancer research: GCR 3: S19.

9. Blazadonakis ME, Zervakis ME, Kafetzopoulos D (2011) Complementary gene signature integration in multiplatform microarray experiments. Information Technology in Biomedicine, IEEE Transactions on 15: 155–163.
10. Guo Z, Zhang T, Li X, Wang Q, Xu J, et al. (2005) Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics 6: 58.
11. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.
12. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP (2007) Classification of microarray data using gene networks. BMC Bioinformatics 8: 35.
13. Lee E, Chuang HY, Kim JW, Ideker T, Lee D (2008) Inferring pathway activity toward precise disease classification. PLoS Comput Biol 4: e1000217.
14. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol 27: 199–204.
15. Binder H, Schumacher M (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinformatics 10: 18.
16. Zhu Y, Shen X, Pan W (2009) Network-based support vector machine for classification of microarray samples. BMC Bioinformatics 10 Suppl 1: S21.
17. Johannes M, Brase JC, Fröhlich H, Gade S, Gehrmann M, et al. (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics 26: 2136–2144.
18. Cun Y, Fröhlich H (2012) Biomarker gene signature discovery integrating network knowledge. Biology 1: 5–17.
19. Cun Y, Fröhlich 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.
20. Staiger C, Cadot S, Kooter R, Dittrich M, Müller T, et al. (2012) A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PloS one 7: e34796.
21. Gade S, Porzelius C, Fälth M, Brase JC, Wuttig D, et al. (2011) Graph based fusion of mirna and mrna expression data improves clinical outcome prediction in prostate cancer. BMC bioinformatics 12: 488.
22. Van der Auwera I, Limame R, Van Dam P, Vermeulen P, Dirix L, et al. (2010) Integrated mirna and mrna expression profiling of the inammatory breast cancer subtype. British journal of cancer 103: 532–541.
23. Zhu M, Yi M, Kim CH, Deng C, Li Y, et al. (2011) Integrated mirna and mrna expression profiling of mouse mammary tumor models identifies mirna signatures associated with mammary tumor lineage. Genome biology 12: R77.
24. Gutiérrez NC, Sarasquete ME, Misiewicz-Krzeminska I, Delgado M, De Las Rivas J, et al. (2010) Deregulation of microrna expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. Leukemia 24: 629–637.
25. Chin L, Hahn WC, Getz G, Meyerson M (2011) Making sense of cancer genomic data. Genes & development 25: 534–555.
26. The Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.
27. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, et al. (2010) International network of cancer genome projects. Nature 464: 993–998.
28. Schmidt M, Böhm D, von Törne C, Steiner E, Puhl A, et al. (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res 68: 5405–5413.
29. Sun Y, Goodison S (2009) Optimizing molecular signatures for predicting prostate cancer recurrence. Prostate Jul 1; 69(10): 1119–27.
30. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, et al. (2010) Integrative genomic profiling of human prostate cancer. Cancer cell 18: 11–22.
31. Hochreiter S, Clevert DA, Obermayer K (2006) A new summarization method for affymetrix probe level data. Bioinformatics 22: 943–949.
32. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, et al. (2011) Pathway commons, a web resource for biological pathway data. Nucleic Acids Res 39: D685–D690.
33. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) Kegg for linking genomes to life and the environment. Nucleic Acids Res 36: D480–D484.
34. Zhang JD, Wiemann S (2009) Kegggraph: a graph approach to kegg pathway in r and bioconductor. Bioinformatics 25: 1470–1471.
35. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) mirbase: tools for microrna genomics. Nucleic Acids Research 36: D154–D158.
36. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) Rocr: visualizing classifier performance in r. Bioinformatics 21: 3940–3941.
37. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, et al. (2009) Annotating the human genome with disease ontology. BMC Genomics 10 Suppl 1: S6.
38. Chung F (2007) The heat kernel as the pagerank of a graph. Proceedings of the National Academy of Sciences 104: 19735–19740.
39. Gao C, Dang X, Chen Y, Wilkins D (2009) Graph ranking for exploratory gene data analysis. BMC Bioinformatics 10 Suppl 11: S19.
40. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.
41. Chapelle O, Vapnik V, Bousquet O, Mukherjee S (2002) Choosing multiple parameters for support vector machines. Machine Learning 46: 131–159.
42. Johannes M, Fröhlich H, Sültmann H, Beissbarth T (2011) pathclass: an r-package for integration of pathway knowledge into support vector machines for biomarker discovery. Bioinformatics 27: 1442–1443.
43. Winter C, Kristiansen G, Kersting S, Roy J, Aust D, et al. (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Computational Biology 8: e1002511.
44. Morrison JL, Breitling R, Higham DJ, Gilbert DR (2005) Generank: using search engine technology for the analysis of microarray experiments. BMC Bioinformatics 6: 233.
45. Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93–99.
46. Pihur V, Datta S, Datta S (2009) Rankaggreg, an r package for weighted rank aggregation. BMC bioinformatics 10: 62.
47. Gonzalez R, Silva JM, Dominguez G, Garcia JM, Martinez G, et al. (1999) Detection of loss of heterozygosity at rad51, rad52, rad54 and brca1 and brca2 loci in breast cancer: pathological correlations. Br J Cancer 81: 503–509.
48. Papp J, Csokay B, Bosze P, Zalay Z, Toth J, et al. (1996) Allele loss from large regions of chromosome 17 is common only in certain histological subtypes of ovarian carcinomas. Br J Cancer 74: 1592–1597.
49. Fiorentino M, Judson G, Penney K, Flavin R, Stark J, et al. (2010) Immunohistochemical expression of brca1 and lethal prostate cancer. Cancer Res 70: 3136–3139.
50. Correa-Cerro L, Wöhr G, Häussler J, Berthon P, Drelon E, et al. (1999) (cag)ncaa and ggn repeats in the human androgen receptor gene are not associated with prostate cancer in a french-german population. Eur J Hum Genet 7: 357–362.
51. Clement JH, Sänger J, Höffken K (1999) Expression of bone morphogenetic protein 6 in normal mammary tissue and breast cancer cell lines and its regulation by epidermal growth factor. Int J Cancer 80: 250–256.
52. Brys M, Stawinska M, Foksinski M, Barecki A, Zydek C, et al. (2004) Androgen receptor versus erbb-1 and erbb-2 expression in human prostate neoplasms. Oncol Rep 11: 219–224.
53. Marsh KL, Varley JM (1998) Frequent alterations of cell cycle regulators in early-stage breast lesions as detected by immunohistochemistry. Br J Cancer 77: 1460–1468.
54. Ceccarelli C, Santini D, Chieco P, Taffurelli M, Gamberini M, et al. (1998) Retinoblastoma (rb1) gene product expression in breast carcinoma. correlation with ki-67 growth fraction and biopathological profile. J Clin Pathol 51: 818–824.
55. Terasawa K, Sagae S, Takeda T, Ishioka S, Kobayashi K, et al. (1999) Telomerase activity in malignant ovarian tumors with deregulation of cell cycle regulatory proteins. Cancer Lett 142: 207–217.
56. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, et al. (2012) Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. Nat Genet 44: 685–689.
57. Howe LR, Brown AMC (2004) Wnt signaling and breast cancer. Cancer Biol Ther 3: 36–41.
58. Yardy GW, Brewster SF (2005) Wnt signalling and prostate cancer. Prostate Cancer Prostatic Dis 8: 119–126.
59. Yardy GW, Brewster SF (2006) The wnt signalling pathway is a potential therapeutic target in prostate cancer. BJU Int 98: 719–721.
60. Sukhtankar D, Okun A, Chandramouli A, Nelson MA, Vanderah TW, et al. (2011) Inhibition of p38-mapk signaling pathway attenuates breast cancer induced bone pain and disease progression in a murine model of cancer-induced bone pain. Mol Pain 7: 81.
61. Kinkade CW, Castillo-Martin M, Puzio-Kuter A, Yan J, Foster TH, et al. (2008) Targeting akt/mtor and erk mapk signaling inhibits hormone-refractory prostate cancer in a preclinical mouse model. J Clin Invest 118: 3051–3064.
62. Shaw G, Prowse DM (2008) Inhibition of androgen-independent prostate cancer cell growth is enhanced by combination therapy targeting hedgehog and erbb signalling. Cancer Cell Int 8: 3.
63. Hardy KM, Booth BW, Hendrix MJC, Salomon DS, Strizzi L (2010) Erbb/egf signaling and emt in mammary development and breast cancer. J Mammary Gland Biol Neoplasia 15: 191–199.
64. Nieman KM, Kenny HA, Penicka CV, Ladanyi A, Buell-Gutbrod R, et al. (2011) Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth. Nat Med 17: 1498–1503.
65. Schölkopf B, Smola A (2002) Learning with kernels. Cambridge: MIT Press Schölkopf, B, Mika, S, Burges, C J, P Knirsch, K-R M, Rätsch, G, & Smola, A J : -2000-81.
66. De Bie T, Tranchevent LC, van Oeffelen LM, Moreau Y (2007) Kernel-based data fusion for gene prioritization. Bioinformatics 23: i125–i132.
67. Gonçalves JP, Francisco AP, Moreau Y, Madeira SC (2012) Interactogeneous: Disease gene prioritization using heterogeneous networks and full topology scores. PloS one 7: e49634.
68. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics.