Integrating Protein-Protein Interaction Networks with Gene-Gene Co-Expression Networks improves Gene Signatures for Classifying Breast Cancer Metastasis

Erik van den Akker[1,2,*], Bas Verbruggen[2], Bas Heijmans[1,3], Marian Beekman[1,3], Joost Kok[1,4,5], Eline Slagboom[1,3], Marcel Reinders[2,5]

[1]Molecular Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands

[2]The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[3]Netherlands Consortium of Healthy Ageing

[4]Algorithms, Leiden Institute of Advanced Computer Science, University Leiden, Leiden, The Netherlands

[5]Netherlands Bioinformatics Centre

Summary

Multiple studies have illustrated that gene expression profiling of primary breast cancers throughout the final stages of tumor development can provide valuable markers for risk prediction of metastasis and disease sub typing. However, the identification of a biologically interpretable and universally shared set of markers proved to be difficult. Here, we propose a method for *de novo* grouping of genes by dissecting the protein-protein interaction network into disjoint sub networks using pair wise gene expression correlation measures. We show that the obtained sub networks are functionally coherent and are consistently identified when applied on a compendium composed of six different breast cancer studies. Application of the proposed method using different integration approaches underlines the robustness of the identified sub network related to cell cycle and identifies putative new sub network markers for metastasis related to cell-cell adhesion, the proteasome complex and JUN-FOS signalling. Although gene selection with the proposed method does not directly improve upon previously reported cross study classification performances, it shows great promises for applications in data integration and result interpretation.

1 Introduction

A crucial step in breast cancer diagnosis and subsequent therapy is the assessment of the tumor's capacity to metastasize. An erroneous diagnosis can either lead to overtreatment or could potentially allow already spread tumors to develop in distant tissues. Since the first leads to a significant amount of unnecessary burden for the patient, while the latter is the predominant cause of death in breast cancer patients [1], a lot of effort has been invested to improve personalized risk profile predictions by employing gene expression assays. However, as whole genome assays are delivering an increasing list of transcriptomic disease markers, the low mutual overlap between different studies becomes apparent. More importantly, obtained sets of prognostic markers from one study show a significant drop in prediction performance when applied to another study [2]. Current methods for gene set grouping may be less successful when performed on a single gene basis, due to the underlying heterogeneity

[*] To whom correspondence should be addressed. Email: erikvandenakker@gmail.com

doi:10.2390/biecoll-jib-2011-188

of the disease as well as the fact that due to secondary effects many genes seem to correlate with the phenotype [2]. Consequently, resulting gene sets purely selected on single gene ranking are often uninformative from a biological point of view.

In response, several types of analyses were developed, which incorporated prior biological knowledge to ensure the biological interpretability of the selected gene set [3, 4, 5]. Genes can for instance be grouped on similar function, localization or pathway membership. However, as many genes are still not assigned to relevant groupings and moreover, all relevant groupings themselves might still not be known, the effectiveness of such an approach might be severely compromised [6].

To deal with the low coverage of predefined functional groupings, several methods have been developed to create groupings *de novo*, by, for instance, exploiting data on physical interactions between proteins. Over the last few years, this type of data has consistently been gathered and integrated with other types of interactions [7], like lethal-lethal [8], co-citations, or cellular co-localization interactions to produce large interaction networks. These so called Protein-Protein Interaction (PPI) networks contain modules that can be linked to cellular functions [9]. The use of these networks for the simultaneous task of relevant gene set discovery and prediction optimization was popularized by the work of Chuang et al. [6]. In this method, sub networks are seeded once at every node in the network and are iteratively grown by greedily adding the best neighbour, until a certain gene set summary statistic no longer improves. Resulting sub networks have been used as input for classification showing an improvement in cross study classification compared to single gene based signatures as well as providing hypothetical biological mechanisms underlying the studied phenotype [6].

Although this is clearly an improvement over previously published methods, we fear that the capacity to generalize over studies is compromised by the greedy aspect with which the seeded sub networks are grown. Given the fact that many genes seem to correlate with the studied phenotype as they are most probably co-expressed due to downstream effects, a considerable part of the data may be viewed as intrinsic biological replicates independently assessing the state of a select number of ongoing cellular processes. In view of this, we would rather like to use *all* informative genes involved in such a process to robustly characterize the cell's transcriptomic state instead of using the genes from a local greedy search only.

A second drawback of greedy network approaches becomes apparent with the growing amount of protein-protein interactions that becomes available. New data predominantly interconnects genes within existing networks, rather than that it connects previously unlinked genes to existing networks. This contributes to the 'small-world' phenomenon [10], referring to a situation where almost every gene in a network is only a few connections away from any other gene. As a consequence, the informative property of localized network sub selection is lost to global and thus less interpretable sub network solutions. A proper biological interpretation is even further compromised if overlap between identified sub networks is allowed. Under these circumstances, numerous highly similar and equally likely solutions will be produced, biasing the selection towards a select set of predictive network hubs, thereby basically reducing the algorithm to a computationally inefficient global ranking method.

Anticipating the previously described problems in selecting genes, we here propose a non-greedy method for dissecting the interaction network in a set of disjoint sub networks. We expect that by incorporating both pair wise gene expression correlation measures, as protein-protein interactions functionally more coherent sub networks will be selected. We hypothesize that building such sub networks will not only generalize better across datasets in predicting the risk of metastasis as they exploit the available information maximally but as well be more informative about the involved biological processes.

2 Materials and Methods

2.1 Materials

In this study six publicly available microarray data sets of breast cancer samples measured on the HG U133A platform (Affymetrix) were employed to test our hypothesis. Raw expression data was downloaded at the NCBI's ftp server [11] under the accessions: GSE7390 [12], GSE3494 [13], GSE6532 [14], GSE1456 [15], GSE2034 [16] and GSE11121 [17]. Data was normalized, log2 transformed and summarized per probe set using the RMA procedure in the Affy package [18] of R [19] at default settings. Replicate and duplicate samples were removed. See Table 1 for an overview of the employed studies.

A recent annotation was downloaded from the Affymetrix website [20] to map all "_at", "_s_at" and "_x_at" probe sets to Ensembl Transcript IDs. Mappings to Ensembl gene IDs and protein IDs obtained from the Ensembl site [21] and protein-protein interactions obtained from STRING [22] were used to map probe sets to the protein-protein interaction network. Probe sets missing annotations to genes, transcripts or proteins, as well as probe sets mapping to multiple genes or probe sets not associated with any interaction data were excluded for further analysis. When multiple probe sets were annotated to the same gene, "_at" probes were preferred over "_s_at" probes and "_s_at" probes over "_x_at" probes. When this did not enforce a decision the probe set with the highest standard deviation was selected. Pre-processing resulted in a mapping of 9.290 probe sets representing 9290 unique genes to a network of 169.566 undirected interactions.

"POOR" and "GOOD" prognosis of samples was assessed using metadata obtained from the NCBI's ftp server [11] as well. "POOR" refers to the occurrence of a distant metastatic event or a relapse within five years after surgery. Subjects were selected for the "GOOD" prognosis subgroup when an event free survival of at least five years was reported. Whereas some studies contained information on distant metastatic events, others reported relapses of breast cancer. When both were available, the reports on distant metastatic events were used.

Table 1: Overview of studies. Statistics on the six studies employed. Accession, #, # - Rep – Dup, Missing, "POOR" and "GOOD" refer to the accession code and the number of samples available at GEO, the number of samples after removal of replicates and duplicate samples, the number of samples with incomplete metadata or prematurely ended censoring, the number of "POOR" prognosis samples and the number of "GOOD" samples respectively. 1) Replicates (Desmedt/Loi) were removed from Desmedt. 2) Replicates (Desmedt/Miller) were removed from Miller. 3) Duplicates (Miller/Loi) were removed from Loi.

Study	Accession	#	# - Rep - Dup	Missing	"POOR"	"GOOD"
Desmedt	GSE7390	198	174[1]	24	31	119
Miller	GSE3494	251	232[2]	37	37	158
Loi	GSE6532	327	186[3]	47	32	107
Pawitan	GSE1456	159	156	6	35	115
Wang	GSE2034	286	286	11	95	180
Schmidt	GSE11121	199	199	19	27	153

2.2 Methods

2.2.1 Proposed method for dissecting the protein-protein interaction network in disjoint co-regulated sub networks

Sub networks are created through evidence-based filtering of edges between genes using two types of evidence: physical interaction data and expression correlations between any pair of genes. Let $\boldsymbol{E}_{ij}$ be the gene expression matrix with probe set i and subject j, where i = 1 to M and j = 1 to N. An $M \times M$ correlation matrix $\boldsymbol{C}$ is computed, where $\boldsymbol{C}_{pq}$ is defined to be the correlation between gene p and gene q over all N samples. Threshold T_{cor} is applied on $\boldsymbol{C}$ to obtain a binary matrix $\boldsymbol{C}^T$, where $\boldsymbol{C}^T_{pq}$ = 1 indicates sufficient and $\boldsymbol{C}^{\boldsymbol{T}}_{pq}$ = 0 indicates insufficient correlation between genes p and q.

Based on a distance matrix equal to 1-abs($\boldsymbol{C}$), the genes are hierarchically clustered (average linkage). The clustering dendogram is thresholded at 1-T_{cor}, creating a grouping matrix $\boldsymbol{G}$ with dimensions $M \times M$, where $\boldsymbol{G}_{pg}$ = 1 indicates co-membership of a gene cluster, and $\boldsymbol{G}_{pg}$ = 0 indicates an assignment to different clusters of gene p and q.

Let matrix $\boldsymbol{P}$ contain the protein-protein interactions, with $\boldsymbol{P}_{pq}$ ranging from 1 to 999 indicating the confidence level associated in case an interaction is reported and $\boldsymbol{P}_{pq}$ = 0 if no interactions are known. Threshold T_{ppi} is applied to $\boldsymbol{P}$ to obtain a binary matrix $\boldsymbol{P}^T$, where $\boldsymbol{P}^T$=1 indicates a presence and $\boldsymbol{P}^T$=0 indicates an absence of known interactions with a sufficient confidence level. The binary correlation matrix $\boldsymbol{C}^T$ is overlayed with the grouping matrix $\boldsymbol{G}$ and the binary protein-protein interaction matrix $\boldsymbol{P}^T$ to yield sub network matrix $\boldsymbol{S}$:

$$\boldsymbol{S}_{pq} = \boldsymbol{G}_{pq}\ \boldsymbol{C}^T_{pq}\ \boldsymbol{P}^{\boldsymbol{T}}_{pq} \qquad \forall_{pq}$$

where S_{pq} = 1 indicates an absolute correlation equal to or exceeding T_{cor} between genes p and q, they are assigned to the same cluster and a physical interaction with a confidence level exceeding T_{ppi} between the proteins of these genes has been reported. S_{pq} = 0 indicates that at least one of these conditions is not met.

Correlations between the breast cancer outcome status and gene-expression data per sub network were evaluated using the global test as summary statistic [5]. This test uses ridge regression to model the relation between breast cancer outcome (response variable) and a set of gene expressions (input variables), while correcting for the mutual correlation structure between the input variables. Obtained sub networks were filtered on significance by applying threshold T_s. Genes within significant sub networks rendered the gene sets used to determine cross study prediction performances and similarities in feature selection.

Since the thresholded gene expression (GE) network ($\boldsymbol{C}^T$) is overlaid with the thresholded PPI network ($\boldsymbol{P}^T$), both thresholds, T_{ppi} and T_{cor} are crucial in determining the connectivity of the resulting network. To balance the influence of both sources of information, T_{ppi} and T_{cor} are chosen such that roughly equal amounts of interactions are obtained for the thresholded GE and PPI networks. As the overlay network rapidly becomes sparser at PPI quality scores exceeding 500 ('medium confidence score' in STRING), T_{ppi} was set to 500 and consequently T_{cor} was set to 0.6.

2.2.2 Competing methods for gene selection

Forward filters were trained as described by van Vliet et al. [23]. In short, a double cross fold loop procedure [24] was employed splitting the data in a validation and a training set (5 folds). The latter is split in an inner training set and an inner test set (10 folds). The additional cross fold setting within the training set implements a strict separation between data used for optimizing the predictor and its evaluation. The optimal number of genes is determined within the inner set by training and evaluating a classifier for up to 200 top ranking genes. Gene ranking was done using absolute Welch's t-statistic. Once the optimal signature size is determined a classifier is trained on the ranked outer training set, which in turn is evaluated in the left out validation set. This procedure is repeated 20 times, thus producing 20*10*5 = 1000 unbiased estimates of the optimal signature size. A final predictive gene set was produced by thresholding the ranked gene list learned on the whole study with the mean over all optimal signature sizes.

Greedy network signatures were obtained by re-implementing the work by Chuang et al. [6] in R [19] using identical settings for all parameters, with the exception that sub network performances were evaluated using a Welch's t-statistic instead of the Mutual Information. Gene sets were obtained by enlisting all unique genes within significant sub networks.

2.2.3 Measures of gene set similarity

The Jaccard index [25] and odds ratio [26] were used to assess the similarity in gene selection between two different studies. The Jaccard index is used to asses the overlap in gene selection and equals the probability for a gene being implicated by both studies, given that it was implicated by at least one study. The odds ratio is used to indicate the consistency in gene selection and is a relative measure of risk representing the increase in likelihood for a gene to be selected, when also selected in another study, compared to a gene being selected, when not selected in another.

2.2.4 Evaluation of Cross Study Prediction Performances

All prediction performances were determined by employing a Nearest Mean Classifier using the cosine-correlation as a distance measure and the Area Under the Curve (AUC) of the Receiver Operator Curve (ROC) as an evaluation measure. Cross study evaluation of the prediction performance was done using two different settings. In the first setting, denoted as *"passing GeneSet"*, a classifier was trained in a five cross fold setting on the gene set indicated by the first study while employing data of the second study. This procedure was repeated 100 times and the mean classification performance over 100*5 folds was reported as the final performance. In the second setting, denoted as *"passing Classifier"*, a classifier was trained on data of the first study and was evaluated using data of a second study. Prediction performances of integration approaches were determined by using five studies as input while evaluating on the sixth. In the "early" integration approach, data integration occurs at the beginning as five studies are jointly analysed to select the genes. The "late" integration approach creates a consensus gene set by intersecting the results of selected genes per study.

2.2.5 Sub network visualization

Sub networks were visualized using the RCytoscape package [27] in R [19] to connect to Cytoscape [28] version 2.8.1. Nodes were colored according to the sign and magnitude of respectively the calculated Welch's t-test statistic and the accompanying p-value (green: higher expressed in POOR outcome compared to GOOD and red vice versa).

3 Results

3.1 Data is dissected in functionally coherent sub networks

Using the proposed methodology, disjoint sub networks were created for six well studied publically available breast cancer studies [12-17] using T_{cor} = 0.6, T_{ppi} = 500 and T_s = 0.05.

Resulting sub networks were visualized using Cytoscape [28] (Fig. 1). Obtained sub networks varied in sizes ranging from 2 up to 192 genes and were either enriched (e.g. Fig 1: B) or depleted (e.g. Fig 1: A) of predictive markers. Furthermore, genes within resulting sub networks showed a preference to be either jointly down or up-regulated, leading to the observation that hardly any significant sub network (sub networks with a red bounding box in Fig. 1) contained oppositely correlating gene expressions with respect to the studied phenotype.

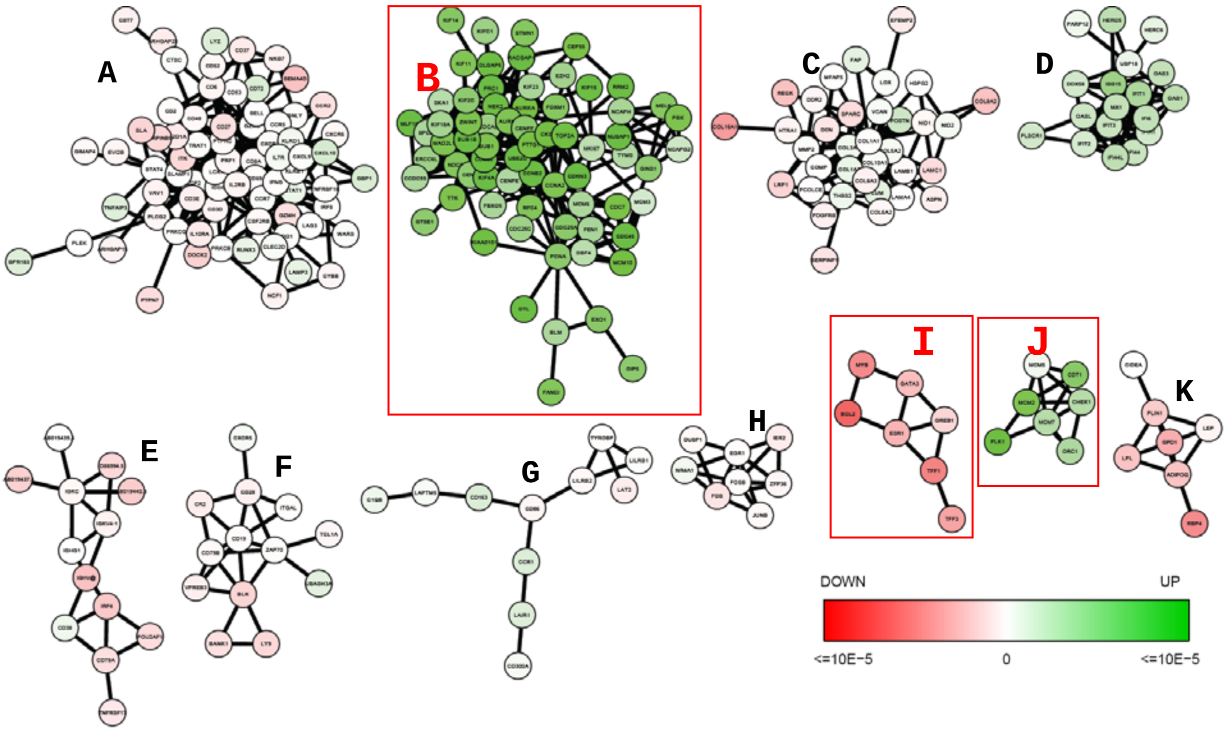

Figure 1: Disjoint sub networks of varying sizes were obtained from the Desmedt study. An overview of the largest sub networks is displayed here. Genes are colored according to the p-value of the Welch-t test on the expression between POOR and GOOD outcome subjects (green is higher expressed in POOR). A red bounding box around a sub network indicates a significant sub network score obtained with the global test on the gene set indicated by the sub network.

In order to assess whether application of the method led to a biologically meaningful dissection of the data, DAVID [29] was used to test for enrichments in functional gene annotations using GO FAT categories. GeneRIF descriptions were inspected for common denominators in case the enrichment analysis returned a-specific or no functional annotations. Sub networks that showed significant associations with respect to the studied phenotype often also showed significant GO enrichments for hallmark processes of breast cancer. For example, for the Desmedt study in Fig. 1: B is enriched for cell cycle phase; I for response to estrogen stimulus; and J for DNA replication. When not related to breast cancer, sub networks could be attributed to processes in lymphocytes or fat tissue. Sub networks enriched for the terms cell cycle phase (GO:0022403), leukocyte activation (GO:0045321) and proteinaceous extracellular matrix (GO:0005578) were seen in all six studies (In Fig. 1 these are sub networks A, B and C respectively).

3.2 Eight sub networks are consistently identified

To get a more thorough view whether the observed dissection in functionally coherent sub networks was consistent between studies, we extended our analyses beyond overlaps in Gene Ontology terms by employing pair wise similarity. For this analysis we calculated Jaccard

doi:10.2390/biecoll-jib-2011-188

indices [25] between sub networks extracted from the six studies and clustered the obtained similarity matrix. The analysis was limited to sub networks with a minimal size of 7 genes yielding 9 to 16 sub networks per study and a total of 83 sub networks (see also Fig. 3). Cluster analysis shows groupings of six sub networks each derived in a different study implicating a high degree of consistency of detected sub networks between the studies (Fig. 2). Besides the previously consistently identified functionalities: leukocyte activation, proteinaceous extracellular matrix and cell cycle phase (Fig. 2, clusters VII, VI and V respectively), five other sub networks with a-specific or no GO enrichments were consistently identified. Common denominators extracted from GeneRIF indicated functionalities related to JUN / FOS signalling for cluster I, interferon induced proteins including ubiquitins for cluster II, Adiponectin / lipid storage for cluster III, Chains of immunoglobulin for cluster IV and immune related genes for cluster VIII.

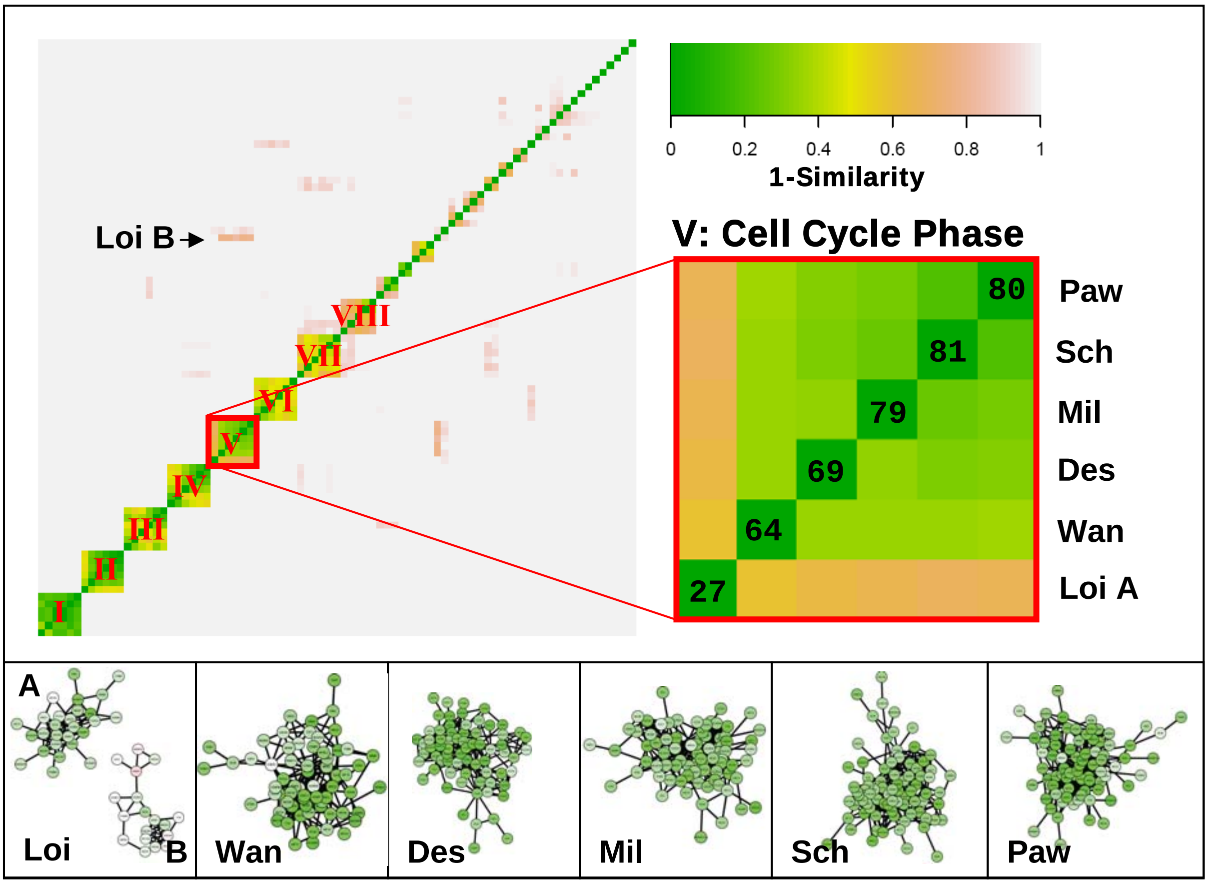

Figure 2: Pair wise similarities were calculated between sub networks obtained from the six studies using Jaccard indices. The resulting similarity matrix was hierarchically clustered and was depicted as a heat map in the upper left corner. The heatmap is symmetric along the diagonal and each row or column represents a unique sub network identified in one of the studies. The grouping belonging to cluster V (Cell Cycle Phase) is blown up to the right. Numbers on the diagonal indicate the number of genes within the identified sub networks. Extensive similarities are observed between sub networks from the six studies except for comparisons involving Loi, caused by the low number of genes found in the Loi study. Icons of sub networks at the bottom represent the sub networks for the different studies that were clustered together in cluster V, which are all also enriched for Cell Cycle Phase. Note that whereas for the Loi study two small sub networks were identified, others studies only returned a single large sub network.

3.3 A "late" integration approach reveals a functionally coherent set of consensus genes putatively involved in metastasis

A consensus gene set of 29 interconnected proteins was retrieved by selecting the genes that were part of a significant sub network throughout *all* six studies ("late" integration), see also Fig. 3. Closer inspection revealed that the majority of these genes have already been implicated as potential therapeutic targets in the treatment of either breast cancer or other types of cancer. This consensus gene set appears to play a pivotal role in the regulation of the cell cycle as not only a considerable enrichment for terms involving the cell cycle (p = 2.6e-16), but as well an enrichment for proteins with known activating capacities was found (5 out of 29 are protein kinases, p = 0.0033). Interestingly, all genes are on average higher expressed within the "POOR" labelled samples compared to the "GOOD" labelled samples, fitting the cancer's hallmark of a shortened cell cycle time. Moreover, all these genes are connected to each other by at least one (predicted) physical interaction exceeding T_{ppi} = 500, thereby suggesting a plausible molecular mechanism how primary breast tumors acquire or maintain their metastatic capacities.

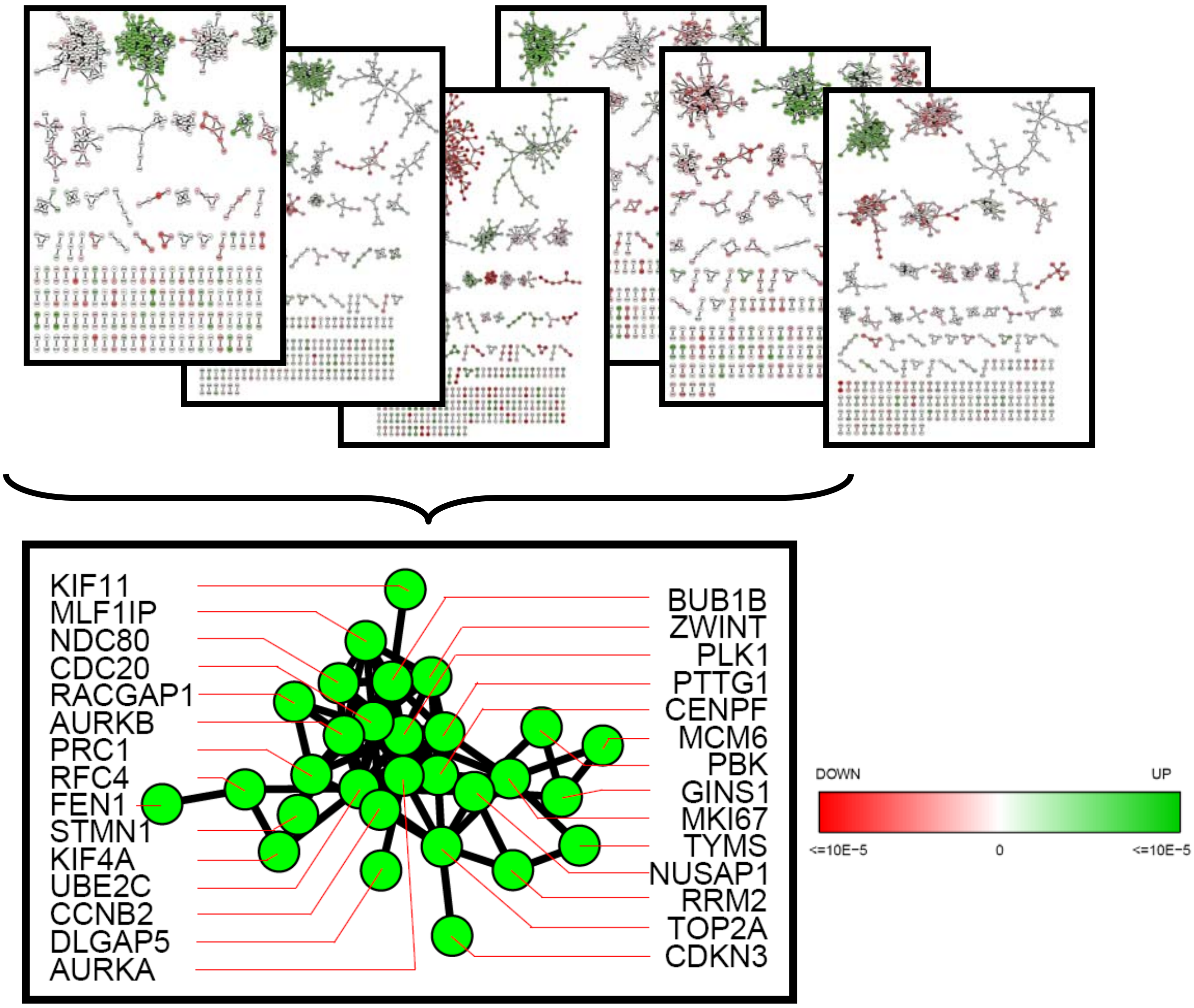

Figure 3: Schematic overview of the construction of a consensus network. Detected sub networks are depicted at the top from left to right for the Desmedt, Muller, Loi, Pawitan, Wang and Schmidt study respectively. A consensus sub network was constructed with genes present in significant sub networks (α=0.05) in all six studies and is depicted at the bottom. Edges in the consensus sub network are drawn when confidence values of reported PPI interactions exceed TPPI.

doi:10.2390/biecoll-jib-2011-188

3.4 An “early” integration approach reveals new sub network markers

We showed that application of the proposed method to six different data sets studying an identical phenotype led to a highly reproducible dissection of the data in at least eight distinct processes. Besides these eight broadly picked up processes, additional smaller clusters are visible along the diagonal in Fig. 2, suggesting that there might be more ongoing processes in primary breast tumor tissue that are harder to detect. By applying the proposed method the data from the six studies concatenated (“early integration”), three new putative sub network markers for metastasis were identified in addition to the eight previously established sub network markers (Fig. 4). These three new putative sub network markers for metastasis (Fig. 4: A to C) could be related to: unfolded protein binding (GO:0051082), cell-cell adhesion (GO:0016337) and proteasome complex (GO:0000502). All previously established sub network markers now dropped below the set significance threshold T_s <= 0.05 and showed a significant enrichment for at least a single GO term. The newly established sub networks B (cell-cell adhesion) and C (proteasome complex) and the previously established sub network markers I (JUN & FOS signalling) and V (Cell Cycle Phase) remained significant even after a Bonferroni correction for multiple testing (sub networks with red bounding box in Fig. 4). All genes identified by the “late” integration approach were again part of significant sub networks found in the “early” approach, predominantly sub network V (26 out of 29), except for the gene *STMN1*. We therefore can view cluster V in Fig. 4 as an extension of the consensus sub network in Fig. 3, containing 22 more candidate genes.

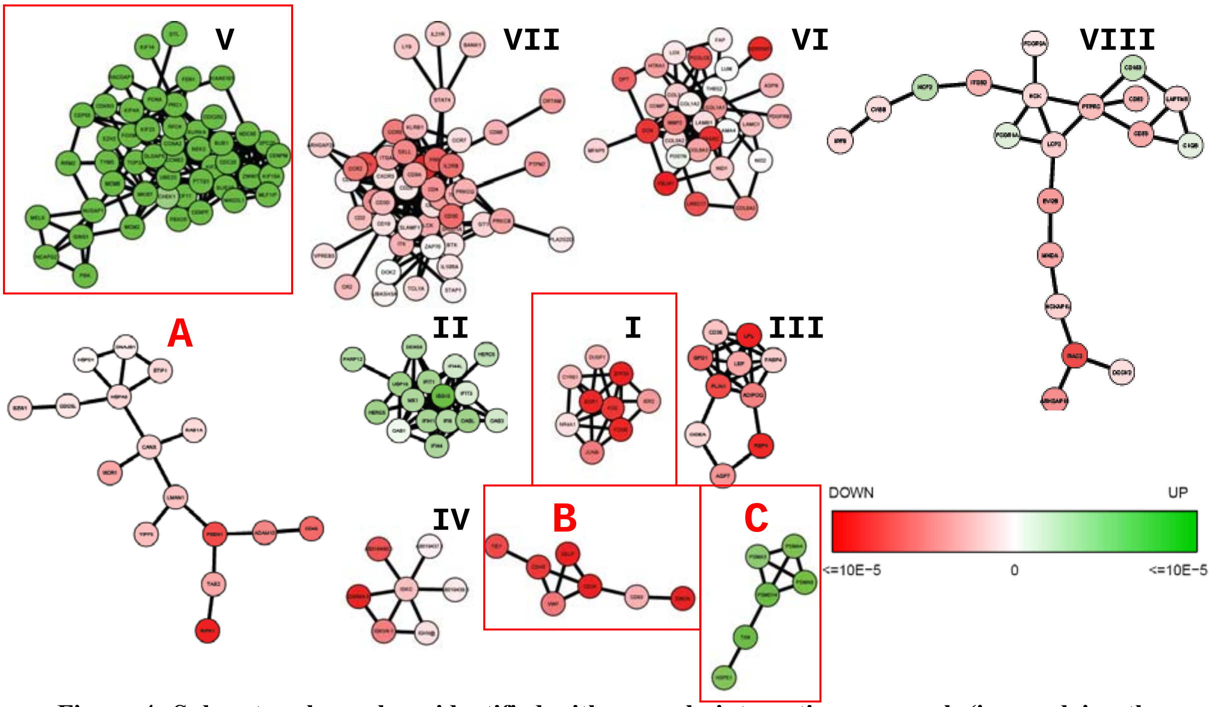

Figure 4: Sub network markers identified with an early integration approach (i.e. applying the procedure to find sub networks on the data of all six studies concatenated). Sub networks marked with black roman numerals correspond to the reported eight consistently identified sub networks, also indicated in Fig 2. Sub networks A, B and C were newly identified and were enriched for the GO terms: unfolded protein binding (GO:0051082), cell-cell adhesion (GO:0016337) and proteasome complex (GO:0000502) respectively. Significant sub networks (T_s <= 0.05) showing a functional enrichment for at least one GO category were reported for this analysis only. Sub networks marked by red bounding boxes remained significant after correction for multiple testing.

3.5 A more consistent gene selection is performed compared to other methods

Consistency in gene selection by the proposed method was compared to a classical gene ranking approach known as forward filtering, as described by van Vliet et al. [23] (Methods 2.2.2) and a greedy network approach, as described by Chuang et al. [6]. Forward filters were used to find optimal predicting gene sets using either all available probes on the array (Table 2: FWD, n = 22,283) or all genes mapped to the protein-protein interaction network (Table 2: FWDNetw, n = 9,290). When starting with a reduced set of initial genes (FWDNetw), only a few additional genes were required for obtaining predictors with very similar prediction performances than when started with the set of all genes (FWD). Both network approaches selected considerably more genes as compared to both settings in which the forward filter was employed. This observation was most extreme for the greedy network approach of Chuang et al. (Table 2: ChuangNetw) for which from 11.6% to 23.0% of the genes mapped to the PPI network (n = 9,290) were selected in hundreds of overlapping sub networks. Application of the proposed method (Table 2: CoRegNetw) resulted in the identification of comprehensible numbers of disjoint co-regulated sub networks and implicating only 1.4% to 5.5% of the genes mapped to the PPI network.

Table 2: Results of selecting predictive genes using different methods on six breast cancer studies. Forward filters (following van Vliet et al. [23]) were used to extract the optimal number of predictive genes (columns # genes (%) refer to the number and percentage of selected genes) when initially starting with all genes on the array (FWD) or all genes mapped to the PPI network (FWDNetw). The method proposed in this article (CoRegNetw) was also compared to the network approach of Chuang et al. [6] (ChuangNetw) and for methods the number of sub networks (# netw.) and average sub network sizes (μ) were reported also.

	FWD	FWDNetw	ChuangNetw		CoRegNetw	
	# genes (%)	# genes (%)	# genes (%)	# netw. (μ)	# genes (%)	# netw. (μ)
Des	49 (0.22)	51 (0.55)	1437 (15.5)	356 (14.4)	130 (1.4)	25 (5.2)
Mil	21 (0.09)	28 (0.30)	2137 (23.0)	662 (14.2)	240 (2.6)	35 (6.9)
Loi	59 (0.26)	75 (0.80)	1098 (11.8)	317 (13.0)	515 (5.5)	80 (6.4)
Paw	48 (0.22)	44 (0.47)	1237 (13.3)	293 (13.7)	290 (3.1)	52 (5.8)
Wan	55 (0.25)	60 (0.65)	1004 (10.8)	423 (11.6)	184 (2.0)	38 (4.8)
Sch	65 (0.29)	56 (0.60)	1696 (18.3)	331 (14.8)	172 (1.9)	22 (7.8)

Consistency of selected genes across different studies using the four previously introduced methods was assessed by calculating (1) Jaccard indices indicating gene set similarities and (2) odds ratios indicating the increase in risk for genes of being selected as a result of a previous selection in another study. The proposed network approach (CoRegNetw) considerably outperformed both ranking settings (FWD and FWDNetw) for all pair wise comparisons between studies for both criteria (Table 3). Whereas the mean Jaccard index was 2.5% and 2.7% for the ranking approaches, respectively, our method showed a mean Jaccard index of 25.9%. Chuang's greedy network approach was outperformed for all odds ratios (Table 3, panel C and D below diagonal), but not for all Jaccard indices (Table 3, panel C and D above diagonal). Although pair wise comparisons involving the Loi study (marked in red) showed lower similarities for our method compared to those observed when employing the method proposed by Chuang et al., the mean Jaccard index of our method still substantially outperformed the means calculated on all other methods (21.9% for CoRegNetw versus 2.7%, 2.5%, and 16.7% and 21.9% for respectively FWD, FWDNetw and ChuangNetw).

Table 3: Gene set similarities calculated between gene sets obtained from significant gene lists (FWD and FWDNetw) or significant sub networks (the method of Chuang et al. ChuangNetw and the method proposed in this paper CoRegNetw) within each single study. Shown similarity measures are the Jaccard index (above diagonal, italic) or odds ratio (below diagonal, not italic) grouped per method (Panels A to D). Pair wise comparisons depicted in green are outperforming all competing methods, the comparisons depicted in red are outperformed by at least one other method.

A: FWD:

OR \ JI	Des	Mil	Loi	Paw	Wan	Sch
Des		*0,01*	*0,00*	*0,08*	*0,09*	*0,05*
Mil	9,6		*0,00*	*0,03*	*0,00*	*0,02*
Loi	1,0	1,0		*0,00*	*0,01*	*0,00*
Paw	37,3	21,1	1,0		*0,04*	*0,05*
Wan	44,8	1,0	2,9	16,4		*0,02*
Sch	17,4	15,4	1,0	17,8	5,5	

B: FWDNetw:

OR \ JI	Des	Mil	Loi	Paw	Wan	Sch
Des		*0,04*	*0,00*	*0,09*	*0,09*	*0,02*
Mil	23,0		*0,00*	*0,01*	*0,00*	*0,04*
Loi	1,0	1,0		*0,01*	*0,01*	*0,00*
Paw	47,4	7,9	2,9		*0,03*	*0,02*
Wan	38,5	1,0	2,1	11,8		*0,02*
Sch	6,9	20,8	1,0	8,1	5,9	

C: ChuangNetw:

OR \ JI	Des	Mil	Loi	Paw	Wan	Sch
Des		*0,21*	*0,16*	*0,20*	*0,15*	*0,19*
Mil	3,3		*0,16*	*0,19*	*0,17*	*0,22*
Loi	3,0	2,6		*0,14*	*0,12*	*0,15*
Paw	3,9	3,2	2,6		*0,13*	*0,18*
Wan	3,0	3,1	2,5	2,5		*0,14*
Sch	3,0	2,9	2,5	3,0	2,4	

D: CoRegNetw:

OR \ JI	Des	Mil	Loi	Paw	Wan	Sch
Des		*0,25*	*0,07*	*0,25*	*0,32*	*0,33*
Mil	71,3		*0,07*	*0,42*	*0,20*	*0,38*
Loi	8,8	4,7		*0,08*	*0,08*	*0,05*
Paw	76,2	123,1	4,7		*0,22*	*0,35*
Wan	117,6	32,2	6,9	39,3		*0,22*
Sch	126,7	139,1	5,6	120,6	44,3	

3.6 Network approaches do not outperform classical ranking approaches in a cross study prediction evaluation

We next were interested whether our method for a highly reproducible dissection in functionally coherent sub networks would improve the robustness of cross study prediction performances. We evaluated the prediction performances in two settings. In both settings a gene set is derived from a first study. In the first setting, denoted "*passing GeneSet*", this gene set is than passed to a second study, where the actual prediction rule is build and evaluated using a proper cross validation. In the second setting, this gene set is used to train a prediction rule with the first study and is evaluated only on the second study. This setting is denoted as "*passing Classifier*" (Fig. 5 and 6).

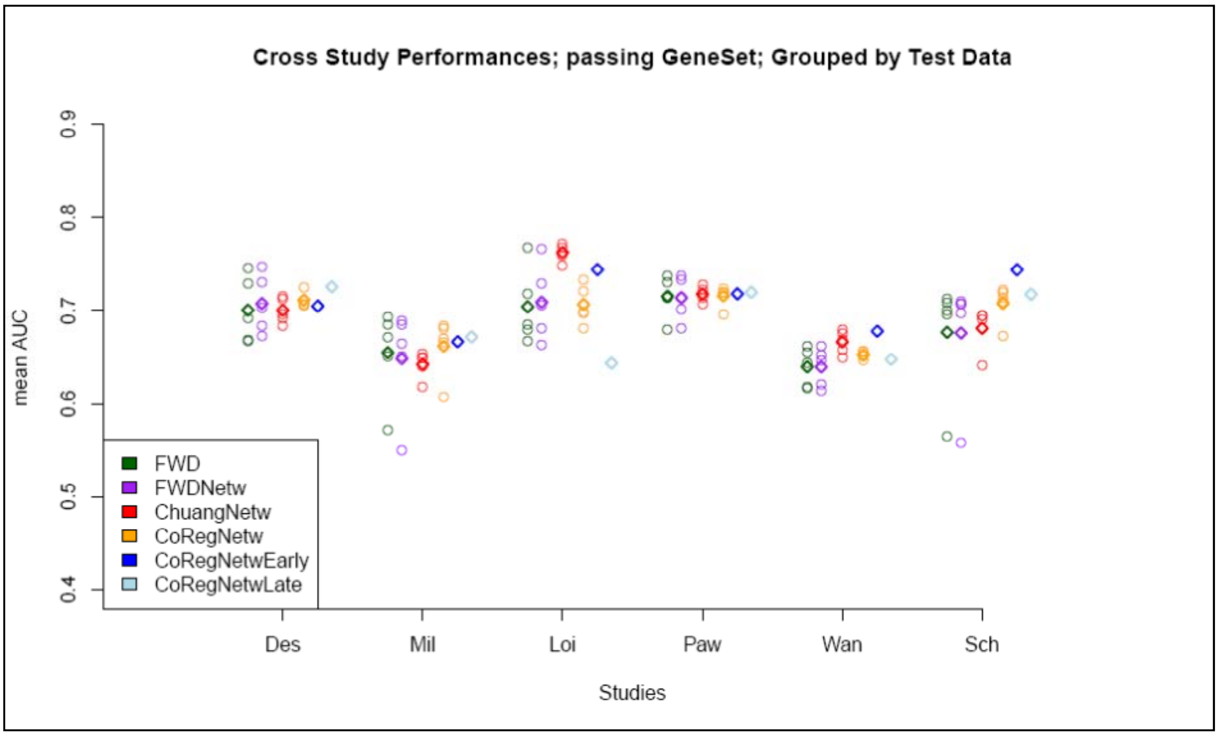

Figure 5: Cross study prediction performances of several methods grouped per evaluation study in the "*passing GeneSet*" setting. Circles indicate results of cross study prediction performances involving a single study for training, diamonds show results involving five studies for training. The latter can either be a summarization statistic (mean) or be the result of an integration approach (CoRegNetwEarly and CoRegNetwLate).

In the "*passing GeneSet*" setting (Fig. 5), network approaches either outperform or show comparable classification performances as compared to classical rankings. Notably, when evaluating on the Loi study Chuang's approach shows a considerable improvement compared to the other methods and when evaluating on the Schmidt study our method considerably outperforms other methods. Prediction performances of the two integration approaches "early" and "late" were evaluated as well. Whereas the "early" integration approach (dark blue diamonds) improves or at least not significantly worsens the prediction performances upon the mean single study approaches (yellow diamonds), the "late" integration approach shows an adverse effect. Especially for the Loi study, the "late" integration approach seems to fail.

In the clinically more relevant "*passing Classifier*" setting (Fig. 6), variations in prediction performances have increased, as expected, compared to the "*passing GeneSet*". Now, classical ranking approaches consistently outperform network approaches. Notably, Chuang's method applied to the Loi study now shows the worst overall performance. The "early" and "late" integration approaches now show a correlated behaviour across data sets, improving upon mean single study performances (yellow diamond) in four out of six times and improving upon both ranking approaches in three out of six evaluations. The integration approaches especially seem to deteriorate prediction performances for the Loi and Wang Study.

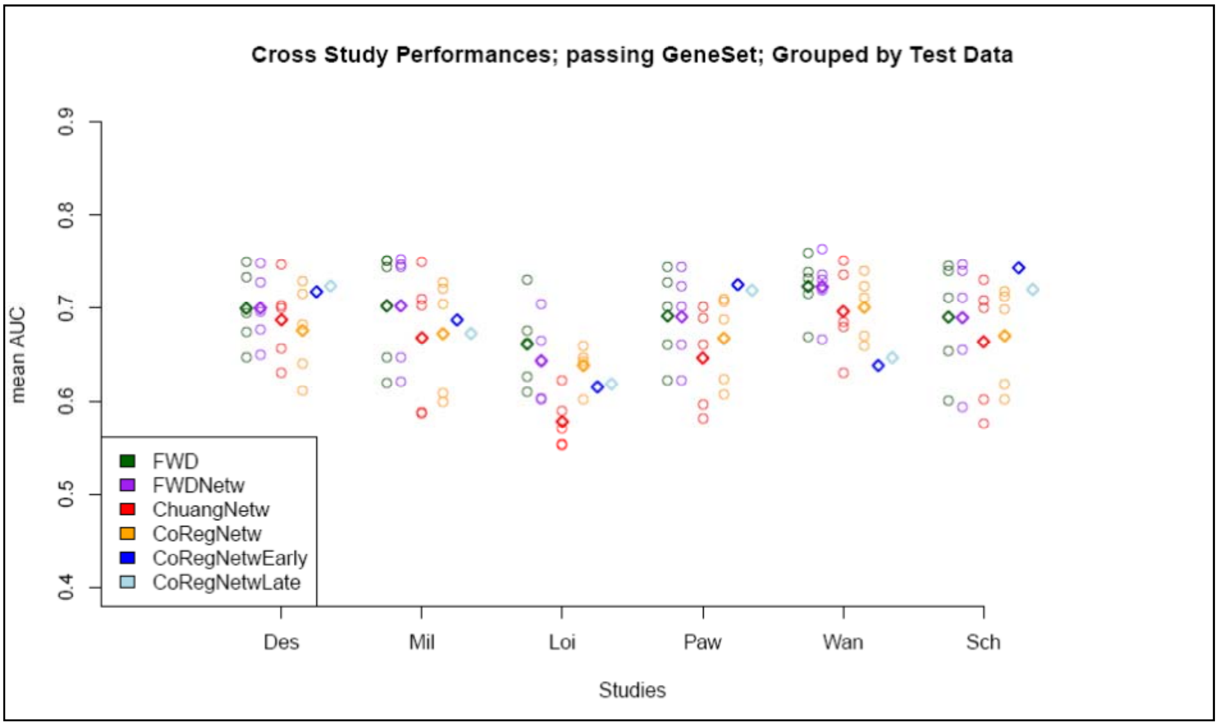

Figure 6: Cross study prediction performances of several methods grouped per evaluation study in the "*passing Classifier*" setting. Circles indicate results of cross study prediction performances involving a single study for training, diamonds show results involving five studies for training. The latter can either be a summarization statistic (mean) or be the result of an integration approach (CoRegNetwEarly and CoRegNetwLate).

4 Discussion & Conclusion

We proposed a method for *de novo* grouping of genes by dissecting the protein-protein interaction network into disjoint sub networks using pair wise gene expression correlation measures. By selecting sub networks significantly correlated with phenotypic outcome, we expected that this would result in a functionally more coherent gene selection as compared to competing risk profile predictors. We verified this by applying the proposed method and two competing methods to a breast cancer compendium composed of six different studies. Furthermore, we investigated whether the expected consistency in gene selection would have benefits for risk prediction of metastasis.

Experiments on the breast cancer compendium have shown that the proposed methodology leads to a *functionally coherent* dissection of genes into sub networks. Furthermore, similarity analyses showed that a considerable amount of these sub networks are picked up *consistently* across studies, suggesting that previously reported low overlaps in predictive gene sets can not be attributed to differences in ongoing basal processes picked up by the different studies. The observation that sub networks were consistently identified underlines the weaknesses of previous methods that purely rely on pre-defined functional groupings for their analyses and interpretation.

Quite contrary to classical gene ranking approaches, extensive overlap between predictive gene sets derived from different studies is observed when employing the proposed method. A consensus gene set that consisted of genes that were part of significant sub networks in *all* six studies was predominantly composed of genes previously implicated in a wide variety of cancers, and was heavily enriched for both the GO term "cell cycle phase" as for the presence of proteins with known regulatory capacities (kinases). This so called "late" integration approach improves robustness in gene selection but at the cost of power to detect potential candidate genes. This was clearly illustrated by the fact that the Loi study alone was most decisive for the gene composition of the consensus gene set, due to its relatively small significant sub network representing cell cycle phase.

A consistent overlap between studies also cleared the way for an “early” integration approach where the data of all studies is concatenated before detecting sub networks. This approach confirmed and extended the consensus sub network found by the late integration approach and identified potential new sub network markers involved in JUN & FOS signalling, cell-cell adhesion and the proteasome complex.

When comparing consistency in gene set selection across studies over different methods, the proposed method always significantly outperforms classical ranking approaches. Chuang’s greedy network approach [6] is outperformed as well except for comparisons involving the Loi study. However, on average Chuang’s method is outperformed using this metric. Moreover when odds ratios for the risk of reselection over the risk of no reselection were compared, our method substantially outperforms Chuang’s method for all pair wise comparisons. This suggests that once a gene is implicated by our method in one study, the chance that it will be implicated again in another study is much higher.

Despite the observed consistency in selection of gene sets, no improvements in classification performance were observed when compared to competing methods in the clinically most relevant evaluation setting (“*passing Classifier*”). Moreover, when no integration approach was employed to exploit the presence of multiple studies, all network approaches were outperformed by the classical gene ranking approaches, suggesting that the higher interpretability comes at the expense of predictive power. In the work of Chuang et al. an evaluation setting similar to the one denoted as “*passing GeneSet*” was used. Indeed we confirmed that in such a setting, network approaches either outperform or show comparable classification performances as compared to classical rankings. However, we would like to issue a word of caution when interpreting the classification results while employing the “*passing GeneSet*” setting. The results with the overall highest prediction performance in the “*passing GeneSet*” setting were created by applying Chuang’s feature selection on the Loi study. Meanwhile, these results also show the largest discrepancy with the setting denoted as “*passing Classifier*”, where it shows the overall lowest prediction performance. We hypothesize that other studies might be particularly uninformative about the Loi study, as this study is the only one in which the majority is treated with tamoxifen, thereby negating or possibly reversing previously observed relations between gene expressions and outcome.

When considering the “*passing Classifier*” setting, integration approaches seem to deteriorate prediction performances especially for two studies: Loi en Wang. In case of the Loi study, integration approaches are expected to be even more sensitive for the previously described disruptive effects of tamoxifen on relations between gene expressions and outcome. Due to the larger amounts of training data, more specific predictors are obtained, which are less capable to generalize when underlying processes are differing. The drop in prediction performance can be explained by the fact that the Wang study is the only one with a balanced number of POOR and GOOD outcomes. Other studies have a much lower incidence of POOR outcome class and therefore training on these studies will focus the classifier mainly on recognizing the more heterogeneous subset of GOOD outcome subjects.

Whereas integration approaches showed some adverse effects in the “*passing GeneSet*” evaluation setting, correlated prediction performances were observed in the “*passing Classifier*” setting. When ignoring the Loi and Wang study, “early” integration approaches seem to only slightly outperform “late” integration approaches. This observation is especially relevant when considering integration of data measured on different platforms in which an “early” integration approach is not feasible.

Employing several analytic strategies, we consistently found a gene sub network involved in an established hallmark of cancer, cell cycle phase, which is persistent over-expressed in all six breast cancer studies in the “POOR” labelled samples compared to the “GOOD” labelled samples. Moreover, application of the proposed method in an “early” integration approach revealed new putative sub network markers, implicating molecular mechanisms involved in cell-cell adhesion, proteasome complex and JUN & FOS signalling to be involved in metastasis. Although not directly improving previously reported cross study classification performances, knowledge-based decomposition of measured gene expression data into co-regulated modules seems to result in a consistent and biologically relevant feature selection and might therefore have a general applicability beyond the field of breast cancer.

Acknowledgements

This work was supported by a grant from the Medical Delta (http://www.medicaldelta.nl).

References

[1] B. Weigelt, J. L. Peterse, and L. J. van ’t Veer. Breast Cancer Metastatis: Markers and Models. *Nature Reviews Cancer*, 5, 591-602, 2005.

[2] L. Ein-Dor, I. Kela, G. Getz, D. Givol and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171-8, 2005.

[3] L. Tian, S.A. Greenberg, S.W. Kong, J. Altschuler, I.S. Kohane, and P.J. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102(38):13544–9, 2005.

[4] A Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[5] J.J. Goeman, S.A. van de Geer, F. de Kort and H.C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93-9, 2004.

[6] H-Y. Chuang, E. Lee, Y-T. Liu, D. Lee and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(140):1-10, 2007.

[7] B. Snel, G. Lehmann, P. Bork, M.A. Huynen. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, 28(18):3442-4, 2000.

[8] M.Michaut, A. Baryshnikova, M. Constanzo, C.L. Myers, B.J. Andrews, C. Boone and G.D. Bader. Protein complexes are central in the yeast genetic landscape. *PLoS Comput Biol*, 7(2):1-11, 2011.

[9] C. Zhang, S. Liu and Y. Zhou. Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. *J Proteome Res*, 5(4):801-807, 2006.

[10] A-L Barabási and Z. Oltvai Network biology: understanding the cell's functional organization. *Nature Reviews*, 5:101-113, 2004.

[11] GEO: ftp://ftp.ncbi.nih.gov/pub/geo

[12] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M.S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A.L.Harris,J.G.M. Klijn, J.A. Foekens, F. Cardoso, M.J. Piccart, M. Buyse and C. Sotiriou. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicentre independent validation series. *Clin Cancer Res*, 1(13):3207-14, 2007.

[13] L.D. Miller, J. Smeds , J. George, V.B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E.T. Liu and J. Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA*, 102(38):13550-5, 2005.

[14] S. Loi, B. Haibe-Kains, C. Desmedt, P. Wirapati, F. Lallemand, A.M. Tutt, C. Gillet, P. Ellis, K. Ryder, J.F. Reid, M.G. Daidone, M.A. Pierotti, E.M. Berns, M.P. Jansen, J.A. Foekens, M. Delorenzi, G. Bontempi, M.J. Piccart and C. Sotiriou. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 22:9-239, 2008.

[15] Y. Pawitan, J. Bjöhle, L. Amler, A.L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S Klaar, E.T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P.M. Shaw, J. Smeds, L. Skoog, S. Wedrén and J. Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*, 7(6):R953-64, 2005.

[16] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, T. Jatkoe, E.M. Berns, D. Atkins and J.A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671-9, 2005.

[17] M. Schmidt, D. Böhm, C. von Törne, E. Steiner, A. Puhl, H. Pilch, H.A. Lehr, J.G. Hengstler, H. Kölbl and M. Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res*, 68(13):5405-13, 2008.

[18] L. Gautier, L. Cope, B.M. Bolstad, and R.A. Irizarry. Affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307-315, 2004.

[19] R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[20] Affymetrix: http://www.affymetrix.com/support

[21] Biomart: http://www.biomart.org/biomart/martview

[22] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel. STRING: a database of predicted functional associations between proteins. URL http://string-db.org/. *Nucleic Acids Res*, 31(1):258-61, 2003

[23] M.H van Vliet, F. Reyal, H.M. Horlings, M.J.T. Reinders and L.F.A. Wessels. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics*, 9(375), 2008.

[24] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, C.J. Veerman, H. Dai, Y.D. He and L.J. van ‘t Veer. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21:3755-62, 2005.

[25] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, 37:547-579, 1901.

[26] A.W.F. Edwards. The measure of association in a 2×2 table. *JSTOR*, 126(1):1-28, 1968.

[27] P. Shannon. RCytoscape: Display and manipulate graphs in Cytoscape. R package version 1.1.53. (2011). http://rcytoscape.systemsbiology.net/versions/current/

[28] M. Smoot, K. Ono, J. Ruscheinski, P-L Wang and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011

[29] D.W. Huang, B.T. Sherman and R.A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*, 4(1):44-57, 2009