Proceedings of the 10th
World Congress on Intelligent Control and Automation
July 6-8, 2012, Beijing, China

Core Module Network Construction for Breast Cancer Metastasis*

Ruoting Yang[1], Bernie J. Daigle Jr[2], Linda R. Petzold[1,2,3], and Francis J. Doyle III[1,4]†
[1] *Institute for Collaborative Biotechnologies*
[2] *Department of Computer Science,* [3] *Mechanical Engineering*
[4] *Chemical Engineering*
University of California, Santa Barbara
Santa Barbara, CA 93106-5080, USA
{ruoting & petzold & doyle}@engr.ucsb.edu, bdaigle@gmail.com
†Corresponding author

***Abstract* - For prognostic and diagnostic purposes, it is crucial to be able to separate the group of "driver" genes and their first-degree neighbours, (i.e. "core module") from the general "disease module". To facilitate this task, we developed a novel computational framework COMBINER: COre Module Biomarker Identification with Network ExploRation. We applied COMBINER to three benchmark breast cancer datasets for identifying prognostic biomarkers. We generated a list of "driver genes" by finding the common core modules between two sets of COMBINER markers identified with different module inference protocols. Overlaying the markers on the map of "the hallmarks of cancer" and constructing a weighted regulatory network with sensitivity analysis, we validated 29 driver genes. Our results show the COMBINER framework to be a promising approach for identifying and characterizing core modules and driver genes of many complex diseases.**

***Index Terms* - Biomarker, Microarray, Network, Sensitivity.**

I. Introduction

DNA microarray technology has been widely used to uncover global gene expression signatures for complex diseases. In recent years, several gene signatures have been identified for predicting the risk of breast cancer metastases, including Wang et al.'s 76 gene signatures [1] and Agendia's MammaPrint chip of 70 gene signatures [2]. Typically, the gene signatures are chosen from the top differentially expressed genes (DEGs) in a single dataset. The large heterogeneity of different datasets often results in poor reproducibility of DEGs and thus even worse reproducibility for gene signatures. For example, there are only 3 overlaps between MammaPrint's 70-gene and Wang's 76-gene signatures.

To enhance reproducibility, biological pathways have been used to group DEGs into functional clusters. These pathways often contain many genes with various sub-functions, while only a few genes associated with a particular sub-function, (i.e. a functional submodule) is typically differentially expressed. Thus, these small submodules are often overlooked statistically and an overlapped pathway may have different submodules.

To specifically focus on relevant submodules, we propose a new pathway inference method that extracts the most important subset of each pathway, typically consisting of a few DEGs. The significances of the submodules can be compared using their Pathway Activities (PAs), which are vectors aggregating the information of all genes expressed in a pathway [3-7]. However, there are two challenges for comparing a PA in different datasets. First, different submodules often result from applying the same pathway inference approach to different datasets. Second, even if comparing the same submodules in different datasets, it is difficult to identify an appropriate statistical significance threshold for PAs.

In light of these challenges, we proposed a multilevel validation framework entitled COre Module Biomarker Identification with Network ExploRation (COMBINER)[8]. COMBINER uses a pathway inference method to find candidate submodules in a designated inference dataset, and it cross-validates the submodules using supervised classification with consensus feature elimination (Fig. 1). If a PA also scores highly in most of the other cohorts, we consider it to be consistently differentially expressed in the disease of interest. Finally, we collect the validated submodules together as a "core module", and use them to construct regulatory networks based on various disease phenotypes. By this way, the changes of the disease can be characterized using different properties of the networks.

Essentially, the "core module" consists of "driver" genes [9] and their first-degree neighbors [10] (Fig.2). These genes are the most invariant part of a general disease module [9], while the remainder are considered downstream passenger genes [9].

To illustrate its utility, we apply COMBINER to three benchmark breast cancer datasets. We generated a list of driver genes by finding the common core modules between two sets of COMBINER markers identified using different module inference protocols. We then explore the roles of the driver genes in the hallmarks of cancer, and we reconstruct a weighted regulatory network composed of functionally coherent modules. Finally, we validate the importance of these driver genes using network and sensitivity analysis.

II. Methods

* This work is supported by U.S. Army Research Office (PTSD Grant W911NF-10-2-011).

978-1-4673-1398-8/12/$31.00 ©2012 IEEE

A. *Gene, pathways, interactome and cancer databases*

We used three large breast cancer metastasis datasets from different countries of origin: Netherlands [11], USA [1], and Belgium [12], to evaluate our method. The Netherlands, USA, and Belgium datasets contain 295, 286, and 198 microarrays, respectively, with 78, 107, and 35 metastatic samples. These datasets contained both lymph-node negative and positive disease patients with differing estrogen receptor (ER) types, as well as patients receiving chemotherapy and hormonal therapy. We performed a two-tailed t-test on the gene expression values of each dataset to distinguish between metastatic and non-metastatic patients, using a significance threshold of p-value ≤ 0.05. Because of the high heterogeneity of the datasets, no false discovery rate adjustment was applied.

We obtained pathway information from the MsigDB v3.0 Canonical Pathways subset [13], which contains 880 pathways collected from seven hand-curated pathway databases including KEGG, Reactome, and Biocarta. To decrease redundancy, we applied pathway filtering to remove bulky pathways such as KEGG Pathways of Cancer. This resulted in a pathway dataset containing 624 pathways with 5,155 genes assayed in all three benchmark datasets. Among all pathway associated genes, only 83 DEGs overlapped between all three breast cancer datasets.

We compared three gene signatures to our identified core module markers: Subnetwork markers (1162 genes) ([3], www.cellcircuits.com); MammaPrint's 70-gene signature (G70) (70 genes) [2]; and Wang's 76-gene signature (G76) (76 genes) [1]. The reference cancer genes for enrichment analysis were collected from datasets including NetPath [14], Atlas of Cancer Genes [15], Census Genes [16], CANgenes [17], G2SBC [18], and KEGG Pathways of Cancer [19].

B. *Core Module Inference*

As illustrated in Fig. 3, given a pathway consisting of n genes with their normalized expression values $\{z(g_1),\dots, z(g_n)\}$, we rank them in descending order of their absolute t-scores. The DEGs are those genes with p-value ≤ 0.05 in a two-tailed t-test. The resulting ordered DEGs $\{g_1,\dots g_i, \dots, g_n\}$ with normalized expression $\{z(g_1),\dots,z(g_n)\}$ are then used to construct pathway activity. If there is no DEG in a pathway, the pathway activity is set to zero. The activity of $\{g_1,\dots g_j\}$ is defined as the weighted sum of their gene expressions

$$P_j = \frac{\sum_{i=1}^{j} z(g_i)\,\mathrm{sign}(t_{score}(g_i))}{\sqrt{j}} \qquad (1)$$

where $1 \le j \le \min(n, 20)$. We limited the largest marker size to 20. Then the pathway activity P_K with its module $\{g_1,\dots, g_K\}$ is determined by the maximum activity

$$K = \arg\max(t_{score}(P_j)), \qquad (2)$$

C. *Reproducibility power*

The reproducibility power of a pathway inference method in an inference-validation pair datasets can be measured by Cscore

$$C_{score}(N) = \frac{1}{N}\sum_{i=1}^{N} t_{score}(P_I^i)\cdot t_{score}(P_V^i), \qquad (3)$$

where P_I^i is the i^{th} PA in descending order in the inference dataset, and P_V^i is its corresponding PA in the validation

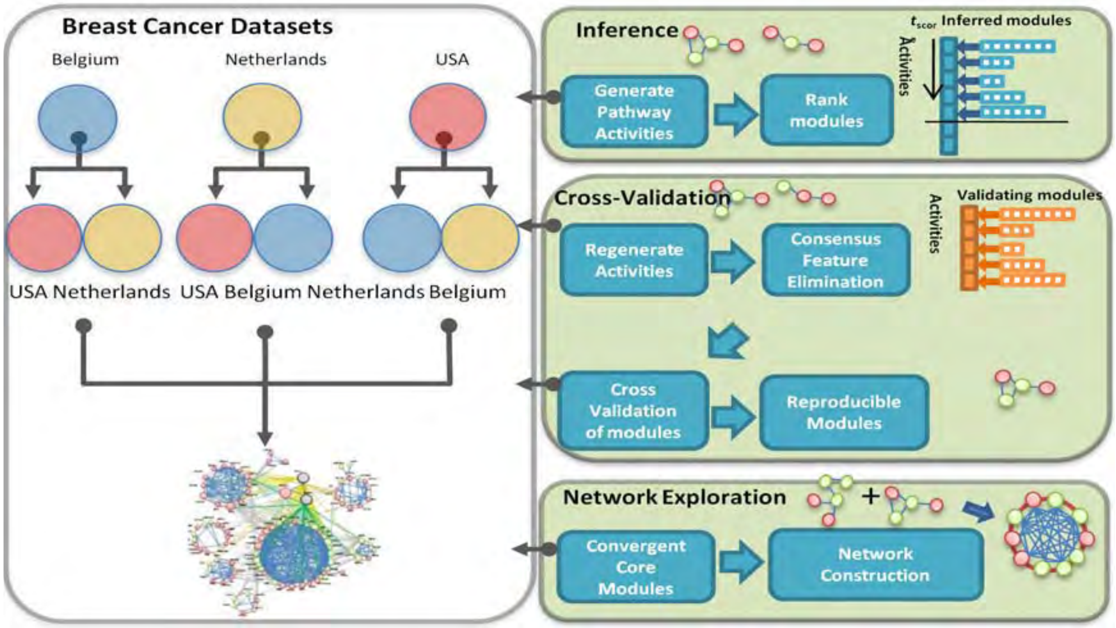

Fig. 1 COMBINER infers candidate submodules from biological pathways in any designated inference dataset, and cross-validates the submodules using supervised classification with consensus feature elimination. Finally, the validated submodules make up the "core module". To identify the "driver" genes, we reassemble the resulting core module markers into a regulatory network reflecting interactions between pathways.

dataset. A pathway inference method is more reproducible; if the identified pathway activities provide similar discriminative power for all independent datasets (i.e. they return higher average Cscores over all inference-validation pairs). For the breast cancer datasets, the overall reproducibility is given by the average Cscore of the inferred pathways over all six inference-validation pairs.

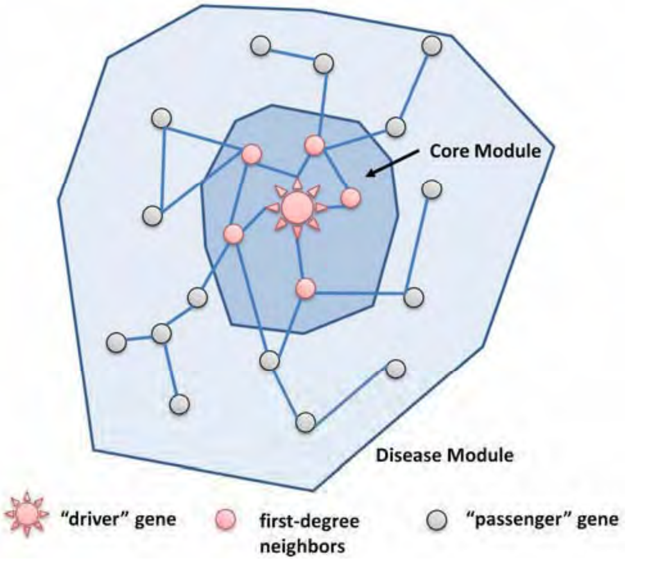

Fig. 2 The core module is the most invariant part of a general disease module, consisting of "driver" genes and their first-degree neighbors The remaining members of the disease module are considered passenger genes.

D. Consensus Feature Elimination (CFE)

To improve stability in feature selection, supervised classification methods with Consensus Feature Elimination (CFE) [20, 21] were used to rank pathway activities. As illustrated in Fig. 4, starting with 100 features, we generate 100 alternative 5-fold random splits of samples, upon which we construct 500 classifiers and compute their mean AUCs (Area Under Receiver Operating Characteristic Curve). The features were ranked by average square weight $\bar{\mathbf{w}}=\sum_{j=1}^{500}\left(\mathbf{w}^{j}\right)^{2}/500$. The lowest ranking AUC was removed recursively until the maximum mean AUC was reached. The above procedure was repeated 100 times, selecting the most frequently occurring best features. Seven methods were compared in this work, including CMI, CORG[6], Mean [5], Median[5], PCA (Principal Component Analysis) [4], LLR(Log likelihood Ratio, [7]), and Individual Gene.

E. Cancer gene enrichment analysis

The cancer gene enrichment analysis assesses statistical significance of cancer genes in a gene signature. Assuming the total number of genes N, cancer genes M, and signature genes J, the probability of having more than K cancer genes in a signature follows a hypergeometric distribution:

$$P(\#\text{ of cancer genes} > K) = 1-\sum_{i=0}^{K}\frac{C_J^i C_{N-J}^{M-i}}{C_N^M} \qquad (4)$$

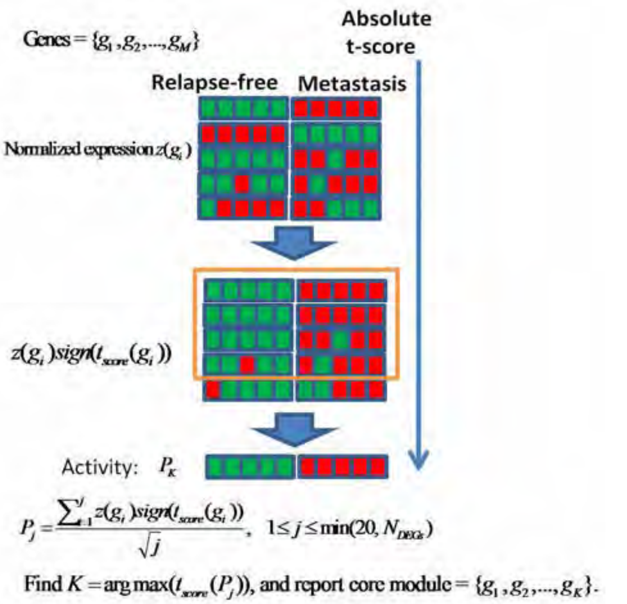
$Genes = \{g_1, g_2, ..., g_M\}$
Absolute t-score
Relapse-free
Metastasis
Normalized expression $z(g_i)$
$z(g_i)sign(t_{score}(g_i))$
Activity: P_K
$P_j = \frac{\sum_{i=1}^{j} z(g_i)sign(t_{score}(g_i))}{\sqrt{j}}, \quad 1 \le j \le \min(20, N_{DEGs})$
Find $K = \arg\max(t_{score}(P_j))$, and report core module $= \{g_1, g_2, ..., g_K\}$.

Fig. 2 An illustration of Core Module Inference. The CMI method combines both up- and down-regulated subset of genes in a pathway by reversing the sign of downregulated gene expression values.

F. Weighted network

Our resulting core module can be used to construct a large scale network by incorporating protein-protein interaction data. However, the task of identifying interaction strength at the transcriptional level still remains. In general, the dynamics of genes $x=[\, x_1, x_2, ..., x_n]^{\mathrm{T}}$ can be represented by a simple linear model[22],

$$\dot{x} = A_I x + Bu \qquad (5)$$

where A_I is an asymmetric network structure matrix characterising the interactions of genes

$$A_I = \begin{bmatrix} -a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & -a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & -a_{nn} \end{bmatrix}, \; B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix} \qquad (6)$$

where the coefficient a_{ij} stands for the influence of gene j on i, and a_{ii} is the degradation rate of gene i. a_{ij} is non-zero if and only if a direct protein-protein interaction exists between gene i and j. $u=[\, u_1, u_2, ..., u_m]^{\mathrm{T}}$ represents the disturbances of the disease applied to the network. The input matrix B collects all disturbances on the genes.

In this work, we assume that disturbances only occur in the early stage of the disease. Thus, the gene interaction dynamics eventually converge to a non-zero steady state, i.e., $A_I x \approx 0$ [22]. We rescale the interaction strengths a_{ij}, $i \neq j$ with the degradation rate, i.e., $\omega_{ij} = a_{ij}/a_{ii}$, which then represent the scaled strength of an edge of the network. In this case, without transient disturbances, the system (5) can be rewritten as

$$\dot{x} = \hat{A}_I x \overset{\Delta}{=} f(x) \tag{7}$$

with $\hat{A}_I = \begin{bmatrix} -1 & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & -1 & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & -1 \end{bmatrix}$, $\omega_{ij} \neq 0$ if and only if an edge exists between gene i and j.

Because $\hat{A}_I x \approx 0$, the scaled strength $\omega_{ij} \neq 0$ can be solved by minimizing the following cost functions, for any

$$\hat{\omega}_{ij} = \arg\min_{\omega}(\| x_i^D - \sum_{i \neq j} \omega_{ij} x_j^D \|^2), \quad i, j = 1, ..., n. \tag{8}$$

where x_i^D, i = 1, ..., n, is the gene expression in the disease state.

G. Network sensitivity analysis

With the estimated scaled strength in Eq (8), we can investigate the sensitivity of these strengths in the neighborhood of the disease state x^D, thus determining the influence of the strengths to the entire network.

Define sensitivity matrix $S(t) = \frac{dx_i(t)}{d\omega_{ij}}$ [23] for all $\omega_{ij} \neq 0$, $i, j = 1, ..., n$. Combining Eq (8), the sensitivity matrix can be solved as follows,

$$\dot{S} = \frac{\partial f}{\partial x} S + \frac{\partial f}{\partial \omega}; \; S(0) = 0. \tag{9}$$

Integrating Eq. (9) for a short interval [0, t_1] in the neighborhood of the mean gene expression for each disease phenotype, we can compute the sensitivity matrix $S(\omega_{ij}, x(t_1))$ of the disease state. The sensitivity matrix is then normalized to $\overline{S}(t) = \frac{dx_i(t)}{d\omega_{ij}} \frac{\omega_{ij}}{x_i(t)}$, which measures the deviation of $x_i(t)$ caused by a unit change of the coefficient ω_{ij}.

We define two measures to score the overall sensitivity of a node (gene) and an edge (protein-protein interaction) with respect to the entire network.

1. Overall gene sensitivity Γ_1 is the sum of absolute normalized sensitivities $\overline{S}(\omega_{ij}, x_i(t_1))$ of node i over all edges.

$$\Gamma_1(x_i) = \sum_{\omega_{ij} \neq 0} |\, \overline{S}(\omega_{ij}, x_i(t_1))\,|, \text{ for all } \omega_{ij} \neq 0 \tag{10}$$

2. Overall interaction sensitivity Γ_2 is the sum of absolute normalized sensitivities $\overline{S}(\omega_{ij}, x(t_1))$ over all nodes from an edge.

$$\Gamma_2(\omega_{ij}) = \sum\nolimits_{i=1}^{n} |\, \overline{S}(\omega_{ij}, x_i(t_1))\,| \tag{11}$$

Larger overall sensitivity values imply a greater influence of a gene or an interaction upon the entire network.

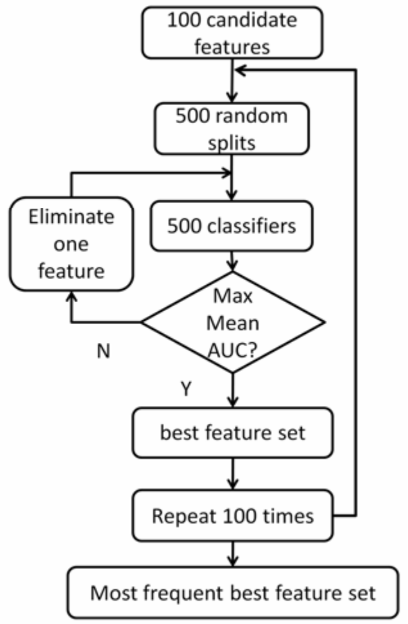

Fig. 3 We first generated 100 alternative 5-fold random splits of samples, upon which we construct 500 classifiers with their AUCs and weight vectors. Each feature is then ranked by its average square weight. The lowest ranking feature was removed recursively until the maximum average AUC was achieved. This procedure is repeated 100 times, and the most frequently occurring marker set was regarded as the final set of markers.

III. Main results

A. Core Module Inference improves reproducibility and classification accuracy

Enhanced reproducibility between independent datasets is one of the most important benefits of performing pathway inference. We compared our CMI approach with five other inference methods [4-6] as well as individual genes using C-scores (See Method section C). CMI showed two-fold increased reproducibility over the related CORG method [6] and about a 10-fold improvement over other methods [8]. Furthermore, CMI also exhibited better overall accuracy when compared to the other methods [4-7] coupled with COMBINER and Linear Discriminant Analysis (LDA) or Support Vector Machine (SVM) with CFE [8].

B. Core module markers enrich cancer-related genes

Both COMBINER run using PA vectors identified by CMI (CMI-COMBINER) and by CORG (CORG-COMBINER) showed much higher enrichment of cancer-related genes in their biomarker signatures [8]. Specifically, CMI- and CORG-COMBINER showed up to 4-fold increased enrichment over subnetwork markers [3] and up to 30-fold enrichment over other gene signatures [1, 2]. For known breast cancer genes, they exhibited up to 4 fold enrichment over the other methods [8].

C. Core module markers highlight the hallmarks of cancer

As shown in Fig. 5, the COMBINER-discovered biomarkers are overlaid on the hallmarks of cancer [24, 25], which integrate the common intracellular signalling pathways of cancer subtypes. The core module markers from CMI and CORG are listed in normal and italic fonts, respectively, while the common markers are in bold. Red/green color denotes up-/down-regulation. The remaining proteins in the pathways are abstracted as unlabeled nodes. Fig. 5 shows that the identified core modules cover all of the hallmarks, demonstrating the high specificity of COMBINER. CMI-COMBINER uniquely identified anti-apoptosis and JAK-STAT cascades, while CORG-COMBINER found anti-growth factors and death factors. Moreover, among 35 common markers between CMI- and CORG-COMBINER core modules (Table I), we 18 of the common markers are directly involved in the hallmarks of cancer. These genes include growth factors, survival factors, and members of the cell cycle and extracellular matrix. It is also notable that a few well-known mutant genes, including cyclin D1 and p53, may play an important role in connecting other signatures [3], but they showed insignificant gene expression profiles in all three breast cancer datasets.

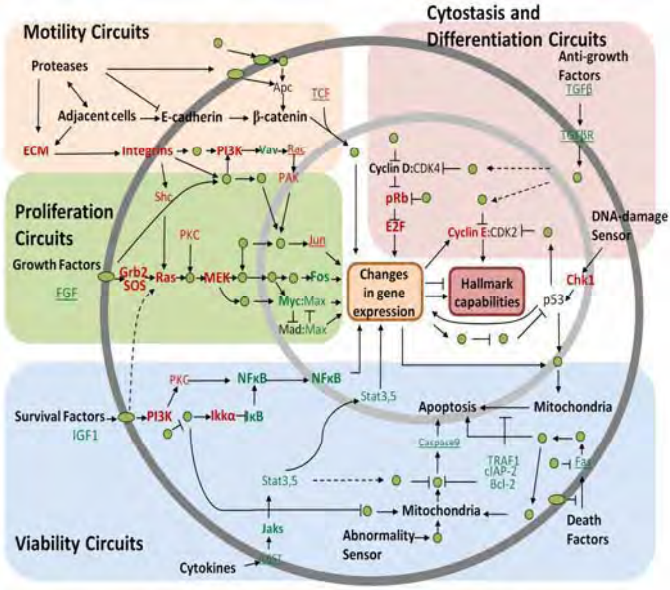

Fig. 5 COMBINER biomarkers overlap with well-known cancer-related signalling pathways. The core module markers from CMI and CORG are listed in normal and italic fonts, respectively, while the common markers are in bold. Red/green color denotes up-/down-regulation. The remaining proteins in the circuit are abstracted as unlabeled nodes.

TABLE I
COMMON MARKERS BETWEEN CMI- AND CORG-COMBINER

Symbol	Entrez	Description
BRCA1	672	breast cancer 1, early onset
FOS	2353	v-fos FBJ murine osteosarcoma viral oncogene homolog
CCNA2	890	cyclin A2
MYC	4609	v-myc myelocytomatosis viral oncogene homolog (avian)
CCNB2	9133	cyclin B2
CCNE2	9134	cyclin E2
CDC45	8318	cell division cycle 45 homolog
GRB2	2885	growth factor receptor-bound protein 2
JAK1	3716	Janus kinase 1
VAV3	10451	vav 3 guanine nucleotide exchange factor
NFKB1	4790	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
NFKBIA	4792	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha
PIK3CA	5290	phosphoinositide-3-kinase, catalytic, alpha polypeptide
PIK3CG	5294	phosphoinositide-3-kinase, catalytic, gamma polypeptide
GNG12	55970	guanine nucleotide binding protein (G protein), gamma 12
CHEK1	1111	CHK1 checkpoint homolog
BCL2	596	apoptosis regulator Bcl-2
CFL1	1072	cofilin 1 (non-muscle)
MCM10	55388	minichromosome maintenance complex component 10
SOS1	6654	son of sevenless homolog 1 (Drosophila)
MCM2	4171	minichromosome maintenance complex component 2
MAP2K1	5604	mitogen-activated protein kinase kinase 1
ORC6L	23594	origin recognition complex, subunit 6
E2F1	1869	E2F transcription factor 1
E2F2	1870	E2F transcription factor 2
SHC1	6464	SHC (Src homology 2 domain containing) transforming protein 1
PKMYT1	9088	protein kinase, membrane associated tyrosine/threonine 1
PPIA	5478	peptidylprolyl isomerase A (cyclophilin A)
RB1	5925	retinoblastoma 1
PLK1	5347	polo-like kinase 1
PSMA7	5688	proteasome (prosome, macropain) subunit, alpha type, 7
PSMD2	5708	proteasome (prosome, macropain) 26S subunit, non-ATPase, 2
PSMD7	5713	proteasome (prosome, macropain) 26S subunit, non-ATPase,7
CSNK2A1	1457	casein kinase 2, alpha 1 polypeptide
DPYD	1806	dihydropyrimidine dehydrogenase

Bold: Validated driver genes

D. Core module markers in predicted protein-protein interaction networks underpin functional modules

As shown in Fig. 6, we used known protein-protein interaction between the core module markers to construct a regulatory network, consisting of 96 nodes and 485 edges. The protein information was obtained from STRING 9[26]. The biomarkers neatly clustered into a few interconnected functional modules, including JAK-STAT, cell cycle and ECM. The gene nodes with many connections were considered the most important nodes in the network. Here we regard the 20 most highly connected genes as "hub genes" (larger pink/green nodes), which interconnected the eight functional modules. In particular, 13 of the hub genes overlapped with common genes (highlighted in Table II).

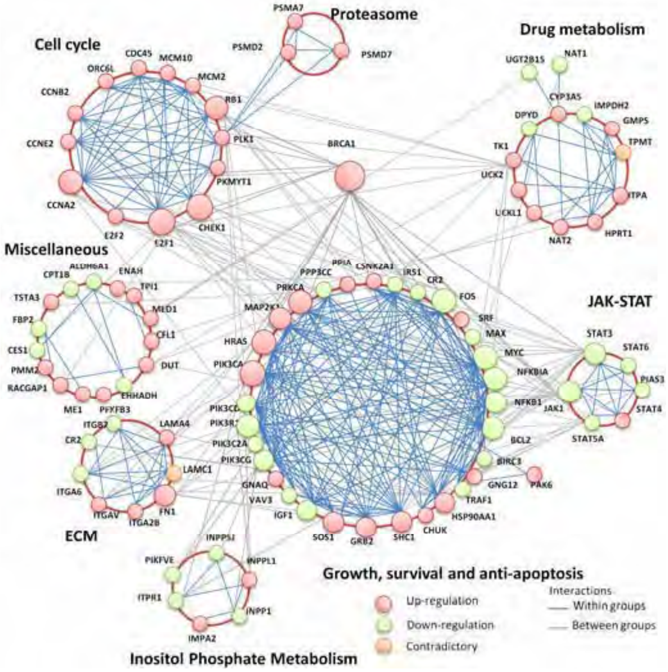

Fig. 6 Regulatory networks of CMI-COMBINER biomarkers The pink/green nodes denote up-/down-regulation of gene expression. The orange nodes indicate contradictory regulation in different datasets. Larger nodes are highly connected in the network; most are overlaps between CMI- and CORG-COMBINER.

E. Weight difference between non-metastatic and metastatic network

We consider the metastatic and non-metastatic breast cancer networks as two disease networks with the same structure but differing interaction strengths Fig. 6. When fit with either the 559 non-metastatic samples x^C and 220 metastatic samples x^D from three breast cancer datasets, we obtained two sets of scaled strengths $\hat{\omega}_{ij}^{C}$ and $\hat{\omega}_{ij}^{D}$, each of which represents 970 directed interactions. The difference between the two networks can be measured by the differential ratio $R_{ij} = (\hat{\omega}_{ij}^{D} - \hat{\omega}_{ij}^{C})/|\hat{\omega}_{ij}^{C}|$. As shown in Fig. 7, the colored edges represent the significant differential ratios such that $|R_{ij}|>10$, with red/green edges denoting increasing/decreasing interaction strenth, i.e. $R_{ij}>10$ or $R_{ij}<-10$. Notably, 29 of the 35 common genes in Table I and 22 hub genes were identified as differentially connected. Significant differences occurred inside the cell cycle module, as well as in interactions between the Growth, survival, and anti-apoptosis module and the ECM module.

F. Overall Gene Sensitivity and interaction sensitivity difference between non-metastatic and metastatic network

To further investigate the influence of a node or an edge upon the entire network, we applied the sensitivity analysis and calculated overall gene sensitivities and overall interaction sensitivities (See Methods section F). Table II lists the top 20 most sensitive genes of both non-metastatic and metastatic networks. The two sensitive gene sets are similar, and both have 8 genes not included in the hub genes. Among those, GNG12, VAV3, and CFL1 belong to the common gene markers in Table I.

Fig 8 shows the top 20 most sensitive interactions for both non-metastatic and metastatic networks. The black edges denote the common sensitive interactions, while the blue/orange edges are sensitive interactions for non-metastatic only/ metastatic only. Five common markers (CCNA2, GNG12, PIK3CG, CSNK2A1, MYC) were involved in six metastatic-specific links (blue), while only four common markers appeared in 19 non-metastatic-specifc links (black and orange).

Taken together, our results involving the hallmark genes of cancer, HUB genes, weight differences, and highly sensitive genes and interactions suggest the 29 common COMBINER gene markers to be the most probable "driver" genes of breast cancer metastasis (highlighted in Table I). In particular, we note that MYC and PIK3CG were highlighted in all of the above analyses.

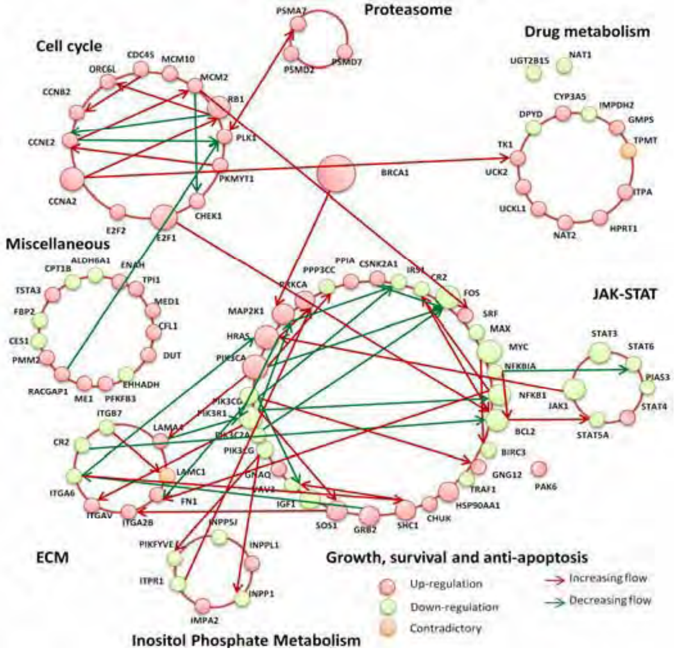

Fig. 7 Differences between metastatic and non metastatic core module network. The red/green edges denote increasing / decreasing strengths of protein-protein interaction in the metastatic network.

TABLE II
TOP 20 HUB, SENSITIVE GENES

HUB	Gene Sens. Non-Metastatic	Gene Sens. Metastatic
MYC	FN1	**MYC**
BCL2	**FOS**	NAT1
PIK3CA	NAT1	FN1
FOS	**MYC**	**FOS**
PIK3R1	STAT3	IRS1
STAT3	HSP90AA1	STAT3
IRS1	IRS1	HSP90AA1
MAP2K1	**JAK1**	PIK3R1
NFKB1	**VAV3**	**JAK1**
SHC1	**NFKBIA**	**NFKBIA**

JAK1	**PIK3CA**	**GNG12**
E2F1	PIK3R1	**VAV3**
HRAS	**GNG12**	BCL2
IGF1	IGF1	**PIK3CA**
BRCA1	BCL2	**CFL1**
CCNA2	**CFL1**	IGF1
GRB2	IMPDH2	**MAP2K1**
PIK3CD	**PIK3CG**	**PIK3CG**
PIK3CG	TPI1	IMPDH2
NFKBIA	**MAP2K1**	ITGAV

Bold: common genes between CMI- and CORG-COMBINER

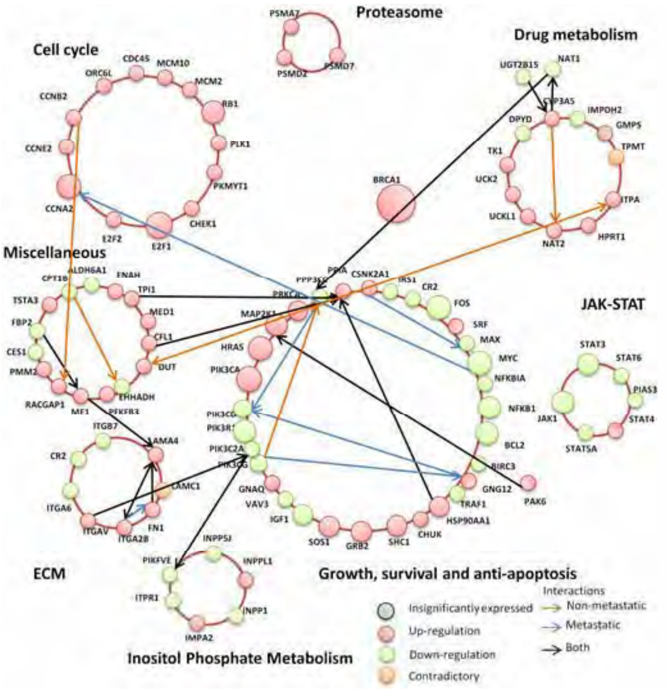

Fig. 8 Top 20 overall interaction sensitivities for both non-metastatic and metastatic network. The black edges denote the common sensitive interactions, while the blue/orange edges are sensitive interactions for non-metastatic only/ metastatic only. Five common markers were involved in metastatic specific sensitive links (blue).

IV. Conclusions

Identifying core modules of complex diseases is an important challenge for gene expression analysis. To facilitate this task, we developed COMBINER, a novel computational framework that extracts the essential "core module" of disease from known biological pathways. We generated a list of "driver genes" by finding the common core modules between CMI- and CORG-COMBINER markers. Overlaying the markers on the map of "the hallmarks of cancer" and constructing a weighted regulatory network with sensitivity analysis, we validated 29 driver genes. After proving the efficiency of COMBINER using the benchmark cancer datasets, we have extended this framework to diseases that are less well-characterized, such as Post-Traumatic Stress Disorder (PTSD) and prion diseases.

References

[1] Y. Wang, *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet,* vol. 365, pp. 671-679, 2005.

[2] L. J. van 't Veer, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature,* vol. 415, pp. 530-536, 2002.

[3] H.-Y. Chuang, *et al.*, "Network-based classification of breast cancer metastasis," *Mol. Syst. Biol.,* vol. 3, 2007.

[4] A. H. Bild, *et al.*, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature,* vol. 439, pp. 353-357, 2006.

[5] Z. Guo, *et al.*, "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics,* vol. 6, p. 58, 2005.

[6] E. Lee, *et al.*, "Inferring pathway activity toward precise disease classification," *PLoS Comput. Biol.,* vol. 4, p. e1000217, 2008.

[7] J. Su, B.-J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PLoS ONE,* vol. 4, p. e8161, 2009.

[8] R. Yang, B. J. Daigle, L. R. Petzold, and F. J. D. III, "Core module biomarker identification with network exploration for breast cancer metastasis," *BMC Bioinformatics,* vol. 13, 2012.

[9] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nat. Rev. Genet.,* vol. 12, pp. 56-68, 2011.

[10] A. Beyer, S. Bandyopadhyay, and T. Ideker, "Integrating physical and genetic maps: From genomes to interaction networks," *Nat Rev Genet,* vol. 8, pp. 699-710, 2007.

[11] M. J. van de Vijver, *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *N England J Med,* vol. 347, pp. 1999-2009, 2002.

[12] C. Desmedt, *et al.*, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series," *Clin. Cancer Res.,* vol. 13, pp. 3207-3214, 2007.

[13] A. Subramanian, *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. USA,* vol. 102, pp. 15545-15550, 2005.

[14] K. Kandasamy, *et al.*, "Netpath: A public resource of curated signal transduction pathways," *Genome Biol.,* vol. 11, p. R3, 2010.

[15] J.-L. Huret, *et al.*, "Atlas of genetics and cytogenetics in oncology and haematology, an interactive database," *Nucleic Acids Res.,* vol. 28, pp. 349-351, 2000.

[16] P. A. Futreal, *et al.*, "A census of human cancer genes," *Nat. Rev. Cancer,* vol. 4, pp. 177-183, 2004.

[17] T. Sjöblom, *et al.*, "The consensus coding sequences of human breast and colorectal cancers," *Science,* vol. 314, pp. 268-274, 2006.

[18] E. Mosca, *et al.*, "A multilevel data integration resource for breast cancer study," *BMC Sys. Biol.,* vol. 4, p. 76, 2010.

[19] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.,* vol. 28, pp. 27-30, 2000.

[20] C. A. Davis, *et al.*, "Reliable gene signatures for microarray classification: Assessment of stability and performance," *Bioinformatics,* vol. 22, pp. 2356-2363, 2006.

[21] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple svm-rfe for gene selection in cancer classification with expression data," *IEEE Trans. NanoBiosci.,* vol. 4, pp. 228-234, 2005.

[22] M. Zampieri, G. Legname, D. Segrè, and C. Altafini, "A system-level approach for deciphering the transcriptional response to prion infection," *Bioinformatics,* vol. 27, pp. 3407-3414, 2011.

[23] H. Rabitz, M. Kramer, and D. Dacol, "Sensitivity analysis in chemical kinetics," *Annual Review of Physical Chemistry,* vol. 34, pp. 419-461, 1983.

[24] D. Hanahan and R. Weinberg, "The hallmarks of cancer," *Cell,* vol. 100, pp. 57 - 70, 2000.

[25] D. Hanahan and Robert A. Weinberg, "Hallmarks of cancer: The next generation," *Cell,* vol. 144, pp. 646-674, 2011.

[26] D. Szklarczyk, *et al.*, "The string database in 2011: Functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.,* vol. 39, pp. D561-D568, 2011.