Core Module Biomarker Identification with Network Exploration for Breast Cancer Metastasis

Ruoting Yang[1*], Bernie J. Daigle Jr[2*], Linda R. Petzold[1,2,3], Francis J. Doyle III[1,4, §]

[1]Institute for Collaborative Biotechnologies, [2] Department of Computer Science,

[3]Department of Mechanical Engineering [4]Department of Chemical Engineering,

University of California Santa Barbara, Santa Barbara, CA 93106-5080

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

Ruoting Yang: ruoting@engineering.ucsb.edu

Bernie J Daigle:bdaigle@gmail.com

Linda R Petzold: petzold@engineering.ucsb.edu

Francis J Doyle III: doyle@engineering.ucsb.edu

Abstract

Background

In a complex disease, the expression of many genes can be significantly altered, leading to the appearance of a differentially expressed "disease module". Some of these genes directly correspond to the disease phenotype, (i.e. "driver" genes), while others represent closely-related first-degree neighbours in gene interaction space. The remaining genes consist of further removed "passenger" genes, which are often not directly related to the original cause of the disease. 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".

Results

We have developed COMBINER: COre Module Biomarker Identification with Network ExploRation. COMBINER is a novel pathway-based approach for selecting highly reproducible discriminative biomarkers. We applied COMBINER to three benchmark breast cancer datasets for identifying prognostic biomarkers. COMBINER-derived biomarkers exhibited 10-fold higher reproducibility than other methods, with up to 30-fold greater enrichment for known cancer-related genes, and 4-fold enrichment for known breast cancer susceptible genes. More than 50% and 40% of the resulting biomarkers were cancer and breast cancer specific, respectively. The identified modules were overlaid onto a map of intracellular pathways that comprehensively highlighted the hallmarks of cancer. Furthermore, we constructed a global regulatory network intertwining several functional clusters and uncovered 13 confident "driver" genes of breast cancer metastasis.

Conclusions

COMBINER can efficiently and robustly identify disease core module genes and construct their associated regulatory network. In the same way, it is potentially applicable in the characterization of any disease that can be probed with microarrays.

Background

In recent years, gene expression signatures based on DNA microarray technology have proven useful for predicting the risk of breast cancer. Agendia's MammaPrint has become the first FDA-cleared breast cancer prognosis marker chip containing 70 gene signatures [1]. Many other microarray-based biomarkers, such as 76 gene signatures [2] have been derived using independent data sources. However, there are only three overlaps between MammaPrint's 70-gene and Wang's 76-gene signatures. Furthermore, many of these markers are functionally unrelated to breast cancer. In order to identify robust, functionally relevant disease biomarkers, it is crucial to find gene signatures that are consistent in various data sources.

A complex disease such as breast cancer results in many differentially expressed genes (DEGs), which together can be used to construct a "disease module" network [3]. Some of these DEGs directly correspond to the disease phenotype (i.e. "driver" genes). The expression changes enacted on the driver genes lead to a cascade of changes of other genes: initially to their first-degree interaction neighbors[4], followed by downstream effects to so-called "passenger" genes. Due to their direct relevance to the biology of the disease in question, the expression changes of the driver genes and their first-degree neighbours (i.e. members of the "core module"), should be more consistent than those of the passenger genes when compared across independent cohorts. However, it is often difficult to separate the core module from the passenger genes for a given disease[5, 6]. In this paper, we aim to isolate the core module from the more general disease module and further identify the driver genes using network analysis.

The most intuitive way of finding the disease core module is to identify the Differential Expressed Genes (DEGs) over various cohorts. Unfortunately, the typically larger number of passenger genes in each cohort will contribute to the majority of gene overlaps, due to statistical chance. A more biologically-motivated technique for identifying the core module is to find overlapping differentially expressed pathways. However, a pathway may also contain hundreds of passenger genes with respect to the disease in question. Moreover, the number of DEGs in a pathway is not always directly proportional to the relevance of that pathway in the disease.

In light of the aforementioned challenges, we propose to identify Pathway Activities (PAs) from cohorts of data and use supervised classification to isolate a consistent core module. Each PA is a vector aggregating the information of all genes expressed in a pathway [7-11]. The use of PAs for biomarker identification has been shown improve reproducibility and disease-related functional enrichment of the resulting biomarkers[7]. The main idea behind our method is to infer the most significant PAs in each data cohort, and validate these PAs using classification methods in other cohorts. If a PA also scores highly in all the other cohorts, we consider it to be consistently differentially expressed in the disease of interest. Furthermore, we would consider the genes that make up the PA to belong to the disease core module.

In this work, we develop a novel biomarker identification framework entitled COre Module Biomarker Identification with Network ExploRation (COMBINER). COMBINER identifies "core module" (Fig. 1) that are consistently differentially expressed as a whole in the data cohorts of interest. COMBINER uses a Core Module Inference (CMI) component to infer candidate PAs from pathway database, a Consensus Feature Elimination (CFE) component to filter out irreproducible PAs, and a multi-level reproducibility validation framework to find the consistent PAs, which in turn make up the complete core module. In its final step, COMBINER uses known pathways and protein networks to identify the driver genes within this core module.

To illustrate its utility, we apply COMBINER to three benchmark breast cancer datasets. We evaluate the resulting core modules for accuracy, reproducibility, and enrichment for known cancer-related genes. We then explore the roles of the COMBINER-identified core modules in the hallmarks of cancer, and we reconstruct a breast cancer-specific interaction network composed of functionally coherent modules. Finally, we summarize our analyses by identifying 13 high confidence driver genes from COMBINER markers.

Results

Overview

COMBINER is a multi-level optimization framework for identifying core module markers (Figure 1, Materials and methods). Briefly, COMBINER infers candidate submodules from known pathways, identifies the reproducible "core module" using independent cohorts, and uses intracellular signaling pathways and protein networks to identify the "driver" genes from the "core module".

We applied COMBINER to three independent breast cancer datasets to evaluate its effectiveness: Netherlands [12], USA [2], and Belgium [13]. We obtained pathway information from the MsigDB v3.0 Canonical Pathways subset [14]. 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.

Core Module Inference improves reproducibility and classification accuracy

A primary challenge of pathway inference is to find pathway subsets that are reproducible between independent datasets. We compared Core Module Inference (CMI) with five other inference methods as well as individual genes (see Materials and methods). When compared to a range of numbers of inferred Pathway Activities (PAs), CMI showed two-fold increased reproducibility over the related CORG method and about a 10-fold improvement over other methods (Figure 2).

We then compared the classification accuracy of CMI and the other inference methods using Linear Discriminant Analysis-Consensus Feature Elimination (LDA-CFE) classifiers focused on the top 100 inferred PAs (Materials and methods). As shown in Figure 3, COMBINER run using PA vectors identified by CMI (CMI-COMBINER) exhibits better overall accuracy than the other methods coupled with COMBINER. Similarly, CMI also shows good overall accuracy using the SVM classifier (Supplementary Figure S1).

Core module markers enrich cancer-related genes

We compared the enrichment of known cancer genes in the biomarkers discovered by CMI-COMBINER, (93 genes); CORG-COMBINER, (i.e. COMBINER run using CORG activity vectors), (123 genes); Subnetwork markers (1162 genes) ([7], www.cellcircuits.com); MammaPrint's 70-gene signature (G70) (70 genes) [1]; and Wang's 76-gene signature (G76) (76 genes) [2]. Seven known cancer gene datasets were compared (see Materials and methods). Both CMI-COMBINER and CORG-COMBINER showed much higher enrichment of cancer-related genes in their biomarker signatures (Table 1). Specifically, CMI- and CORG-COMBINER showed up to 4-fold increased enrichment over subnetwork markers and up to 30-fold enrichment over other gene signatures. In particular for known breast cancer genes in Census, they exhibited up to 4 fold enrichment over others. More than 50% and 40% of the resulting biomarkers are cancer and breast cancer specific, respectively. Additionally, CMI-COMBINER showed greater enrichment than CORG-COMBINER with respect to the Atlas of Cancer Genes, which is the largest cancer gene collection. Consistent to Chuang et al's results[7], we also found insignificant enrichment in CANgene dataset including 122 mutative genes from 11 breast cancer cell lines. A possible explanation is that "the cancer cell lines capture a different disease state than that found in the population of patients surveyed by microarray profiling." [7] The COMBINER core module markers with associated pathways are summarized in Supplementary Tables S1 and S3. Supplementary Table S2 lists the overlaps between CMI-/CORG-COMBINER and KEGG pathways of cancer, along with up-/down-regulation information.

Core module markers highlight the hallmarks of cancer

As shown in Figure 4, the COMBINER-discovered biomarkers are overlaid on the hallmarks of cancer [15, 16], which integrate the common intracellular signalling pathways of all subtypes of cancer. The components of the core modules from CMI and CORG along with eighteen common markers are listed in different fonts. The remaining proteins (most were not differentially expressed) in the pathways are consolidated into unlabeled nodes. Figure 4 shows that the identified core modules comprehensively highlight the hallmarks, demonstrating the high specificity of COMBINER. In particular, 18 common markers, which we regard as the most reliable predictors, describe well-characterized processes involving growth factors, survival factors, the cell cycle, and the extracellular matrix. The modules unique to CMI-COMBINER include anti-apoptosis and JAK-STAT cascades, while pathways describing anti-growth factors and death factors were unique to CORG-COMBINER. A few well-known mutant proteins, including cyclin D1 and p53, may play an important role in connecting other signatures [7], but they showed only limited predictive ability in the three breast cancer datasets.

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

Figure 5 shows how a regulatory network was constructed using the interactome of the core modules. The regulatory network was divided into a few functional modules, including cell cycle and ECM. These functional modules were interconnected by 20 "hub" genes (larger pink/green nodes), 13 of which overlapped with the hallmarks of cancer (Supplementary Table S1, Figure 4). Our results imply that these 13 "hub" markers are the essential "driver" genes of breast cancer metastasis (Table 2). For example, BRCA1 is among the most well-characterized genes whose mutation gives rise to breast cancer. In addition, low E2F1 transcript levels strongly predicted good prognosis based on quantitative RT-PCR in 317 primary breast cancer patients [17]. We further enlarged the nodes of three standard breast cancer indicators TP53, BRCA1, and ERBB2, which connect many of the surrounding hub genes. Although TP53 and ERBB2 are useful for a mechanistic understanding of breast cancer, they were not identified as discriminative gene markers. A regulatory network was also created representing CORG-COMBINER (Supplementary Figure S2), but no additional "hub" markers were found.

Conclusions

Identifying accurate and reproducible disease biomarkers is an important challenge for gene expression analysis. To facilitate this task, we developed COMBINER, a novel pathway-based biomarker identification method that extracts the essential "core module" of disease from known biological networks. Compared to existing methods, COMBINER substantially improves the reproducibility and cancer-specific enrichment of its resulting biomarkers. We examined the identified markers in intracellular signalling networks highlighting the hallmarks of cancer. Reassembling the core modules into a regulatory network, we found 13 "driver" genes connecting eight functional modules. We anticipate such molecular descriptions to prove even more useful when applied to diseases that are less well-characterized; our current work focuses on several such applications.

Methods

Gene expression, pathways, cancer gene databases, and interactome

We used three breast cancer datasets from different countries of origin to evaluate our method: Netherlands [12], USA [2], and Belgium [13]. Each dataset recorded whether the assayed patients developed metastasis within 5 years after surgery. The

Netherlands, USA, and Belgium datasets contain expression profiles for 295, 286, and 198 patients, respectively, with 78, 107, and 35 patients experiencing metastasis. All of the patients in the USA and Belgium datasets had lymph-node-negative disease, although their estrogen receptor (ER) types differed. The Netherlands data contained both lymph-node positive and negative disease patients with differing ER types, 130 of which received adjuvant systemic therapy including 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, considering genes with p-value $\leq .05$ as differentially expressed (DE).

The reference cancer genes for enrichment analysis were collected from datasets including NetPath [18] (all cancers, http://www.netpath.org/), Atlas of Cancer Genes [19] (all cancers, http://atlasgeneticsoncology.org/), Census Genes [20] (all cancers), CANgenes [21] (breast cancer), G2SBC [22] (breast cancer, http://www.itb.cnr.it/breastcancer/), and KEGG Pathways of Cancer [23] (all cancers, KEGG hsa05200 http://www.genome.jp/kegg/pathway /hsa/hsa05200.html).

Pathway information was obtained from the MsigDB v3.0 Canonical Pathways subset [14, 24]. This collection contains 880 pathways collected from seven hand-curated pathway databases including KEGG, Reactome, and Biocarta.

Predicted protein protein interaction information was obtained from STRING 9 [25].

Core Module Inference

The CMI method adopts the strategy of the CORG method [10] of finding the genes with the most discriminative power, differing in three ways: first, the CORG method collects CORGs only from the up- or downregulated subset of genes in a pathway, and some key genes can thus be discarded. In contrast, CMI considers both up- and downregulation together. Second, CMI improves the greedy search for the discriminative set of genes. Third, CMI considers only differentially expressed genes. As illustrated in Box 1, given a pathway consisting of genes {g_1,… g_i, ..., g_n} ranking by a descending order of their absolute t-scores, with their normalized expression values {$z(g_1)$,…, $z(g_n)$}, determining a core module {g_1,…, g_K} is equivalent to finding the K^{th} component, such that

$$K = \arg\max(t_{score}(P_j)), \qquad (1)$$

where

$$P_j = \begin{cases} \dfrac{\sum_{i=1}^{j} z(g_i)\mathrm{sign}(t_{score}(g_i))}{\sqrt{j}}, 1 \le j \le \min(|\, g_i \in DEGs\,|, 20), & |\, g_i \in DEGs\,| > 0, \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad , & |\, g_i \in DEGs\,| = 0. \end{cases} \qquad (2)$$

g_i is the i^{th} DEG in descending order and P_j is the PA containing from g_1 to g_j. $|\, g_i \in DEGs\,|$ denotes number of DEGs in the pathway, The DEGs by default are the genes with p-value ≤ 0.05 in a two-tailed t-test. We limit the largest marker size to 20 DEGs. In fact, most marker sets have fewer than 20 components.

Reproducibility power

We consider two pathways to be reproducible if their pathway activities provide similar discriminative power for all independent datasets. First, we rank the PAs inferred from the inference dataset in descending order by their tscores. Then, we define reproducibility by

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

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 dataset. For the breast cancer datasets, the overall reproducibility is then given by the average Cscore of the inferred pathways over all six inference-validation pairs.

Six methods were compared in this work, including CMI, CORG(Lee *et al*, 2008), Mean[9], Median[9], PCA[8], and Individual Gene. LLR(Log likelihood Ratio, [11]) was not compared here, because it is not discussed in the same gene expression space.

Consensus Feature Elimination (CFE)

In this work, gene expression and activity vectors are generalized as features for classification. Given a set of features {$\boldsymbol{x}_1$, $\boldsymbol{x}_2$,..., $\boldsymbol{x}_n$} with class labels {y_1, y_2,..., y_n}ϵ{ -1, +1}, the task of binary classification is to find a decision function

$$D(x)\begin{cases} >0 \Rightarrow \boldsymbol{x} \in class(+) \\ <0 \Rightarrow \boldsymbol{x} \in class(-) \\ =0 \Rightarrow \boldsymbol{x} \in \boldsymbol{decision\ boundary,} \end{cases} \tag{4}$$

We choose a linear decision function, which can be described as a separating hyperplane:

$$D(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b}, \tag{5}$$

with ***w*** the weight vector and ***b*** the bias value.

Linear classifiers such as Linear Discriminant Analysis (LDA) [26] and linear Support Vector Machines (SVM) [27] use differing optimization criteria to estimate the weight vector. Intuitively, the weights indicate the importance of the associated features. Guyon *et al* proposed Recursive Feature Elimination (RFE), which removes features recursively based on their weights [28]. However, classical RFE exhibits lack of stability in feature selection [29]. In contrast to binary classification tasks that emphasize maximization of classification accuracy, biomarker identification requires features that are both accurate and reproducible across multiple experiments. Thus, we propose a Consensus Feature Elimination (CFE) approach to improve the stability of RFE. As illustrated in Figure 6, we first generate 100 alternative 5-fold random splits of samples, upon which we construct 500 classifiers and record their AUCs (Area Under ROC Curve) and weight vectors. Each feature was then ranked by average square weight $\bar{\mathbf{w}}=\sum_{j=1}^{500}\left(\mathbf{w}^{j}\right)^{2}/500$. The lowest ranking feature was removed recursively until the maximum average AUC was achieved. This process, which has also been called Multiple RFE [30] or ensemble feature selection [31] is known to increase biomarker reproducibility and accuracy by as much as 30% and 15%, respectively. For the breast cancer datasets described in this work, we found the maximum AUC to be very stable, while the corresponding biomarker set was not always unique. Thus we chose to repeat the above procedure 100 times, selecting the most frequently occurring biomarkers as the final marker set.

Seven methods were compared in this work, including CMI, CORG[10], Mean [9], Median[9], PCA [8], LLR(Log likelihood Ratio, [11]), and Individual Gene.

Cancer gene enrichment analysis

The cancer gene enrichment analysis examines over-representation of known 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:

$$P(\#\text{ of cancer genes} > K) = 1-\sum_{i=0}^{K}\frac{\binom{J}{i}\binom{N-J}{M-i}}{\binom{N}{M}}. \qquad (6)$$

Software

COMBINER was implemented in Matlab R2010a with Matlab Bioinformatics toolbox v3.5.

Authors' contributions

RY, BJD, LRP, and FJD conceived and designed the research. RY, and BJD performed the analysis, the statistical computations, and wrote the paper. RY implemented the programs. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge financial support from U.S. Army Research Office (PTSD Grant W911NF-10-2-0111).

References

1. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
2. Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J *et al*: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer**. *Lancet* 2005, **365**(9460):671-679.
3. Barabasi A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nat Rev Genet* 2011, **12**(1):56-68.
4. Beyer A, Bandyopadhyay S, Ideker T: **Integrating physical and genetic maps: from genomes to interaction networks**. *Nat Rev Genet* 2007, **8**(9):699-710.
5. Li J, Lenferink AEG, Deng Y, Collins C, Cui Q, Purisima EO, O'Connor-McCourt MD, Wang E: **Identification of high-quality cancer prognostic markers and metastasis network modules**. *Nat Commun* 2010, **1**:34.
6. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**(2):171-178.
7. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T: **Network-based classification of breast cancer metastasis**. *Mol Syst Biol* 2007, **3**.
8. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi M-B, Harpole D, Lancaster JM, Berchuck A *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**. *Nature* 2006, **439**(7074):353-357.
9. Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, Zhu J, Wang H, Wang C, Topol E *et al*: **Towards precise classification of cancers based on robust gene functional expression profiles**. *BMC Bioinformatics* 2005, **6**(1):58.
10. Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification**. *PLoS Comput Biol* 2008, **4**(11):e1000217.
11. Su J, Yoon B-J, Dougherty ER: **Accurate and reliable cancer classification based on probabilistic inference of pathway activity**. *PLoS ONE* 2009, **4**(12):e8161.

12. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al*: **A gene-expression signature as a predictor of survival in breast cancer**. *N England J Med* 2002, **347**(25):1999-2009.
13. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, d'Assignies MS *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* 2007, **13**(11):3207-3214.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci USA* 2005, **102**(43):15545-15550.
15. Hanahan D, Weinberg R: **The hallmarks of cancer**. *Cell* 2000, **100**:57 - 70.
16. Hanahan D, Weinberg Robert A: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.
17. Vuaroqueaux V, Urban P, Labuhn M, Delorenzi M, Wirapati P, Benz C, Flury R, Dieterich H, Spyratos F, Eppenberger U *et al*: **Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome**. *Breast Cancer Res* 2007, **9**(3):R33.
18. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar G, Venugopal A, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C *et al*: **NetPath: a public resource of curated signal transduction pathways**. *Genome Biol* 2010, **11**(1):R3.
19. Huret J-L, Minor SL, Dorkeld F, Dessen P, Bernheim A: **Atlas of genetics and cytogenetics in oncology and haematology, an interactive database**. *Nucleic Acids Res* 2000, **28**(1):349-351.
20. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nat Rev Cancer* 2004, **4**(3):177-183.
21. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N *et al*: **The consensus coding sequences of human breast and colorectal cancers**. *Science* 2006, **314**(5797):268-274.
22. Mosca E, Alfieri R, Merelli I, Viti F, Calabria A, Milanesi L: **A multilevel data integration resource for breast cancer study**. *BMC Sys Biol* 2010, **4**(1):76.
23. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
25. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P *et al*: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored**. *Nucleic Acids Res* 2011, **39**(suppl 1):D561-D568.
26. Friedman JH: **Regularized discriminant analysis**. *J AM STAT ASSOC* 1989, **84**(405):165-175.
27. Vapnik V: **Statistical Learning Theory**: Wiley-Interscience; 1998.

28. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines**. *Mach Learn* 2002, **46**(1):389-422.

29. Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, Küffner R, Zimmer R: **Reliable gene signatures for microarray classification: assessment of stability and performance**. *Bioinformatics* 2006, **22**(19):2356-2363.

30. Duan K-B, Rajapakse JC, Wang H, Azuaje F: **Multiple SVM-RFE for gene selection in cancer classification with expression data**. *IEEE Trans NanoBiosci* 2005, **4**(3):228-234.

31. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y: **Robust biomarker identification for cancer diagnosis with ensemble feature selection methods**. *Bioinformatics* 2010, **26**(3):392-398.

32. MacDonald TJ, Brown KM, LaFleur B, Peterson K, Lawlor C, Chen Y, Packer RJ, Cogen P, Stephan DA: **Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease**. *Nat Genet* 2001, **29**(2):143-152.

33. Vuaroqueaux V, Urban P, Labuhn M, Delorenzi M, Wirapati P, Benz C, Flury R, Dieterich H, Spyratos F, Eppenberger U *et al*: **Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome**. *Breast Cancer Res* 2007, **9**(3):R33.

34. Giubellino A, Burke TR, Bottaro DP: **Grb2 signaling in cell motility and cancer**. *Expert Opin on Ther Tar* 2008, **12**(8):1021-1033.

35. Van Laere SJ, Van der Auwera I, Van den Eynden GG, Elst HJ, Weyler J, Harris AL, van Dam P, Van Marck EA, Vermeulen PB, Dirix LY: **Nuclear Factor-κB Signature of Inflammatory Breast Cancer by cDNA Microarray Validated by Quantitative Real-time Reverse Transcription-PCR, Immunohistochemistry, and Nuclear Factor-κB DNA-Binding**. *Clin Cancer Res* 2006, **12**(11):3249-3256.

36. Hamann U, Herbold C, Costa S, Solomayer E-F, Kaufmann M, Bastert G, Ulmer HU, Frenzel H, Komitowski D: **Allelic Imbalance on Chromosome 13q: Evidence for the Involvement of BRCA2 and RB1 in Sporadic Breast Cancer**. *Cancer Res* 1996, **56**(9):1988-1990.

37. Rakha EA, Reis-Filho JS, Ellis IO: **Basal-Like Breast Cancer: A Critical Review**. *J Clin Oncol* 2008, **26**(15):2568-2581.

38. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B *et al*: **Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis**. *J Natl Cancer Inst* 2006, **98**(4):262-272.

39. Smid M, Wang Y, Klijn JGM, Sieuwerts AM, Zhang Y, Atkins D, Martens JWM, Foekens JA: **Genes Associated With Breast Cancer Metastatic to Bone**. *J Clin Oncol* 2006, **24**(15):2261-2267.

40. Campbell IG, Russell SE, Choong DYH, Montgomery KG, Ciavarella ML, Hooi CSF, Cristiano BE, Pearson RB, Phillips WA: **Mutation of the PIK3CA Gene in Ovarian and Breast Cancer**. *Cancer Res* 2004, **64**(21):7678-7681.

41. Woelfle U, Cloos J, Sauter G, Riethdorf L, Jänicke F, van Diest P, Brakenhoff R, Pantel K: **Molecular Signature Associated with Bone Marrow Micrometastasis in Human Breast Cancer**. *Cancer Res* 2003, **63**(18):5679-5684.

42. Ursini-Siegel J, Hardy WR, Zuo D, Lam SHL, Sanguin-Gendreau V, Cardiff RD, Pawson T, Muller WJ: **ShcA signalling is essential for tumour progression in mouse models of human breast cancer**. *EMBO J* 2008, **27**(6):910-920.

43. Wolfer A, Wittner BS, Irimia D, Flavin RJ, Lupien M, Gunawardane RN, Meyer CA, Lightcap ES, Tamayo P, Mesirov JP *et al*: **MYC regulation of a “poor-prognosis” metastatic cancer cell state**. *Proc Natl Acad Sci USA* 2010, **107**(8):3698-3703.

Figures

Figure 1 - Schematic overview of COMBINER

COMBINER uses Core Module Inference (CMI) to infer candidate pathway activities from each pathway in an inference dataset, Consensus Feature Elimination (CFE) to filter out irreproducible activities in validation datasets, and a multi-level reproducibility validation framework to conduct pair-wise validations to find common reproducible activities which make up the "core module". To identify the "driver" genes, we reassemble the resulting core module markers in both intracellular signalling pathways and a large overall regulatory network reflecting interactions between pathways.

Figure 2 - Reproducible power of pathway inference methods. The reproducibility of a pathway is measured by $C_{score}(N) = \frac{1}{N}\sum_{i=1}^{N} t_{score}(P_I^i) \cdot t_{score}(P_V^i)$, where P_I^i is the ith PA in descending order in the inference dataset, and P_V^i is its corresponding PA in the validation dataset. The overall reproducibility is then defined as the average Cscore of selected top inferred pathway activities over all six inference-validation pairs. We did not compare LLR method, which transfers gene expression to a log-likelihood space. We compared CMI with five inference methods, including the CORG, mean, median, first component score of PCA, as well as no-inferring gene method. Comparing by different ranges of top inferred activities, the CMI showed significant better overall reproducibility over other methods.

Figure 3 - Comparison of CMI and other inference methods-based COMBINER using LDA-CFE classifiers focused on the top 100 inferred pathways. Seven methods were compared here, including CMI, CORG, Mean, Median, PCA, LLR and Individual Gene. (a) Classification accuracy for best feature set: pair-wise comparisons. Starting from all 100 inferred pathway activities, we recursively removed the activity with the lowest average weight from 500 LDA classifiers, until the maximum average AUC was reached. The process was repeated 100 times and the most frequently occurring marker set was regarded as the ultimate marker. We measured classification accuracy of each method by computing AUC mean ± standard error for the final feature set. (b) Classification accuracy overall. The overall classification accuracy was measured by computing the average maximum mean AUC of all six inference-validation pairs. On average, CMI was superior to the other

methods, even though its activity vector consisted of expression values from only a few genes in each pathway.

Figure 4 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. The common core modules of CMI- and CORG-COMBINER describe growth factors, survival factors, the cell cycle, and the extracellular matrix. Unique pathways to CMI-COMBINER include the anti-apoptosis and JAK-STAT cascade, while anti-growth factor and death factor pathways were discovered uniquely by CORG-COMBINER.

Figure 5 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. The three well-known oncogenes for breast cancer metastasis–TP53, BRCA1, and ERBB2–were enlarged further. The core module markers were reassembled into an overall interaction network. Known functional modules neatly overlay well-connected clusters. Many of the highly connected genes are known "driver" genes playing an important role in breast cancer metastasis.

Figure 6 Diagram of Consensus Feature Elimination. We first generated 100 alternative 5-fold random splits of samples, upon which it constructs 500 classifiers with their AUCs as well as weight vectors. Each feature is then ranked by its average square weight. The lowest ranking feature was removed backward until the maximum average AUC was achieved. The procedure is repeated for 100 times, and the most frequently occurring marker set was regarded to be the ultimate marker.

Tables

Table 1 Cancer Gene Enrichment rate of various breast cancer gene signatures

	CMI-COMBINER	CORG-COMBINER	Subnetwork	G70	G76
NetPath	54.17%*	50.41%*	26.33%*	10.00%	10.53%
Atlas	60.42%*	46.34%	32.87%	15.71%	18.42%
Census	11.46%*	13.82%*	5.42%*	2.86%	0.00%
CANgene	1.04%	1.63%	0.52%	0.00%	0.00%
G2SBC	43.75%*	46.34%*	19.02%	21.43%	10.53%
COSMIC	16.67%	17.89%*	7.06%	4.29%	1.32%
KEGG	35.42%*	29.27%*	9.90%*	8.57%	1.32%

* p-value < 0.05 for hypergeometric tests

Table 2 Confident "driver" genes for breast cancer metastasis

Symbol	Entrez	Description
MAP2K1 [32]	5604	mitogen-activated protein kinase kinase 1
E2F1 [33]	1869	E2F transcription factor 1
GRB2[34]	2885	growth factor receptor-bound protein 2
NFKB1 [35]	4790	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
RB1[36]	5925	retinoblastoma 1
BRCA1 [37]	672	breast cancer 1, early onset
FOS [38]	2353	v-fos FBJ murine osteosarcoma viral oncogene homolog
SOS1[39]	6654	son of sevenless homolog 1 (Drosophila)
PIK3CA[40]	5290	phosphoinositide-3-kinase, catalytic, alpha polypeptide
JAK1 [41]	3716	Janus kinase 1
SHC1[42]	6464	SHC (Src homology 2 domain containing) transforming protein 1
MYC[43]	4609	v-myc myelocytomatosis viral oncogene homolog (avian)
CCNA2 [38]	890	cyclin A2

Additional files

Additional file 1 – Supplemental materials

Figure S1 Comparison of CMI and other pathway inference methods using SVM-MRFE classifiers subject to top 100 inferred pathways.

Figure S2 Unique core modules of cancer pathway identified by CORG-COMBINER method.

Additional file 2 – Table S1: List of core genes identified by CMI and CORG

Additional file 3 – Table S2 List of core module genes overlaid in KEGG pathway of cancers

Additional file 4 – Table S3 Pathway markers identified by all methods