

Network-based classification of breast cancer metastasis

Han-Yu Chuang^{1,4}, Eunjung Lee^{2,3,4}, Doheon Lee³ and Trey Ideker^{1,2}

¹ Bioinformatics Program, University of California San Diego, La Jolla, California 92093, USA. ² Department of Bioengineering, University of California San Diego, La Jolla, California 92093, USA. ³ Department of Biosystems, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea. ⁴ These authors contributed equally to this work

Email contact: Han-Yu Chuang (hchuang@ucsd.edu) and Trey Ideker (trey@bioeng.ucsd.edu)

Mapping the pathways that give rise to metastasis is one of the key challenges of breast cancer research. Recently, several large-scale studies have shed light on this problem through analysis of gene expression profiles to identify markers correlated with metastasis. However, each study identifies a different set of marker genes, and it remains unclear how these genes interrelate within a larger functional network.

To address this limitation, we applied a protein-network-based approach that identifies markers not as individual genes but as subnetworks extracted from protein interaction databases. Gene expression profiles from two cohorts of patients (295¹ and 286² breast cancer patients, respectively) were mapped to a human protein interaction network consisting of 57,235 interactions among 11,203 proteins. A search over this network was performed to identify *discriminative subnetworks* which could be used to classify the disease status of a patient. Specifically, each subnetwork was scored by a vector of activities across the patients, where the activity for a given patient is a function of the expression levels of its member genes. A subnetwork's discriminative potential was computed as the mutual information between the activity vector and the disease status (metastatic vs. non-metastatic).

The resulting ~250 discriminative subnetworks identify new putative cancer markers and suggest novel regulators involved in tumor progression. This network-based analysis has four major advantages over previous approaches. First, the discriminative networks provide an array of “small-scale” models charting the molecular mechanisms correlated with metastasis. Each of these may be further investigated through directed experimental and computational methods. Example subnetworks are shown in **Fig. 1a**, detailing interactions between P53, Ras, SMAD4 and ECM-related proteins. Second, although genes with known breast cancer mutations are typically not detected through analysis of differential expression (e.g. P53, KRAS, HRAS and PIK3KA in the figure), they play a central role in the protein network by interconnecting many expression-responsive genes. Consequently, the enrichment of known cancer-susceptibility genes is substantially higher in discriminative subnetworks than in gene signatures identified by expression profiling alone. Third, the identified subnetworks are significantly more reproducible between the two breast cancer cohorts than individual marker genes selected without network information (**Fig. 1b**). Finally, network-based classification achieves higher accuracy in prediction, as ascertained by selecting markers from one dataset and applying them to a second independent validation dataset (**Fig. 1c**). Thus, a network-based approach integrating protein interactions with breast cancer expression profiles leads to increased classification accuracy and, simultaneously, provides a view of the biological processes underlying metastasis.

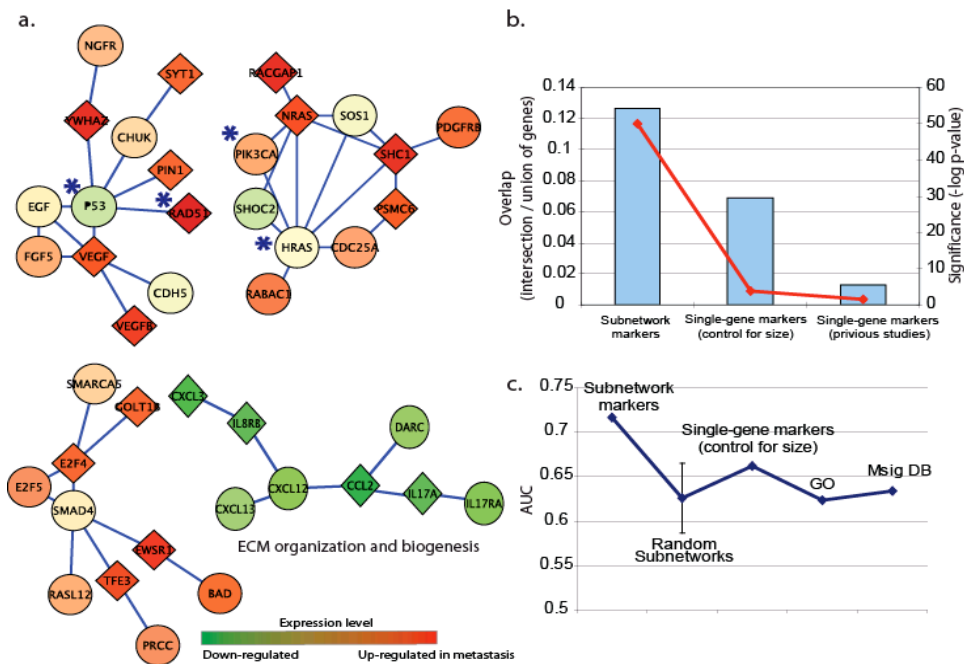


Figure 1. [a] Four example discriminative subnetworks. Nodes and links represent human proteins and previously-characterized protein interactions, respectively. The color of each node scales with the change in expression of the corresponding gene for metastatic versus non-metastatic cancer. Diamonds represent significant differential-expression ($p < 0.05$). Known breast cancer susceptibility genes are marked by a blue asterisk. **[b] Agreement in markers selected from two datasets (DS1, DS2).** Blue bars chart the overlap on the left axis; the red line charts the hypergeometric p-values of overlap on the right axis. The first “single-gene” analysis was done by using the same number of genes covered by subnetwork markers. **[c] Five-year metastasis-free prediction in DS1¹ using the discriminative subnetworks selected from DS2².** AUC (Area Under ROC Curve) was used as a metric to evaluate performance.

Reference

1. van de Vijver, M.J., *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).
2. Wang, Y., *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-679 (2005).