Using Gene Sets to Identify Putative Drugs for Breast Cancer

Tzu-Hung Hsiao[1], Hung-I Harry Chen[1], Yidong Chen[1,2*]
[1]Greehey Children's Cancer Research Institute,
[2]Department of Epidemiology and Biostatistics,
University of Texas Health Science Center at San Antonio, San Antonio, Texas.
[*]Correspondence: cheny8@uthscsa.edu, and chuangey@ntu.edu.tw

Yu-Heng Chen[3], Eric Y. Chuang[3*]
[3]Graduate Institute of Biomedical Electronics and Bioinfomatics,
National Taiwan University, Taiwan.

***Abstract*— The number of current anti-cancer drugs was limited and the response rates were also not high. To "reposition" known drugs as anti-cancer drugs to increase the therapeutic efficiency, we presented a novel analysis framework to identify putative drugs for cancer. Using breast cancer as example, a "cancer – gene sets – drugs" network was constructed through two procedures. First, the "gene sets - drugs" network was built by applying the expression pattern of drugs for gene set enrichment analysis. Secondly, the breast cancer progression associated gene sets were identified by survival analysis of patient cohorts. By integrating the two results, 25 tumor progression associated gene sets and 360 putative anti-cancer drugs were identified. Our method has the ability to identify the "reposition" drugs and the potential affected mechanisms of tumor progression concurrently. It will be useful to speed up the development of anti-cancer drugs from bench to clinical application.**

Keywords: gene set, drug,breast cancer.

I. INTRODUCTION

Lots of drugs were developed to treat cancer based on different pathways. Chemo-drugs affect the cell cycle related mechanism. For example, cisplatin crosslinks DNA to interfere with cell division by mitosis and induces cell apoptosis [1]; Gemcitabine replaces cytidine in the building blocks of nucleic acids during DNA replication and arrests tumor growth [2]; Irinotecan prevents DNA from unwinding by inhibition of topoisomerase I, which is required for DNA replication, and leads to cell death [3]. Target drugs were developed to inhibit tumor driver genes to turn off the dysregulated signal transduction, such as Her2 inhibitor, Trastuzumab [4] ,and EGFR inhibitor, Erlotinib[5]. Drugs that attack other mechanisms, such as angiogenesis [6] and triggers of the immune system [7] , were also developed. Although lots of drugs can be utilized to treat cancers, the response rates were still not high. The survival time of some malignant cancers were not prolonged after treatments. Therefore, to develop drugs to attack other dysregulation pathways of cancers is crucial for the research of anti-cancer drugs.

It takes a lot of time and money to develop anti-cancer drugs [8]. Repositioning drugs to anti-cancer is one of the useful strategies to overcome the bottleneck. Several computational biology approaches have been developed to explore drug repositioning [9-12]. Among those methods, gene expression profile of drug treatment provides higher probability to find the "repositioned" drugs because it profiles the transcriptional response in a whole-genome level. A data resource, i.e. "Connectivity map", which collects the expression profiles of human cancer cell lines treated by more than 1,000 bioactive compounds, provides fundamental information to discover the putative therapeutic compounds of cancer [13]. Several algorithms utilized gene signatures of drug responses to match the expression patterns between drugs and diseases to explore drug candidates for diseases, [13-17]. However, the data in the connectivity map only records the changes in the mRNA level and has no measurements of cell phenotype change. Although associations of expression level between drugs and diseases were identified, the underlying mechanisms and physical effects were still unclear.

In cells, the activities of cellular responses or pathways were regulated by gene expression. Therefore, the underlying phenotype changes or pathway alterations of cancer cells can be delineated through observation of gene expression, such as HIF [18], EMT [19]. Here we proposed an integrated approach which utilizes gene sets to identify putative drugs of breast cancer. The gene sets were applied to Cox proportional regression to identify the relationship between the presented biological function and tumor progression. An enrichment procedure was designed to identify gene set associated drugs. By integrating the information, a "cancer – gene sets- drugs" network was constructed. To demonstrate the ability of our analysis framework, 432 response gene sets [20] were applied to the analysis. Finally 360 drugs which affect the 25 response gene sets in breast tumors were discovered. Those putative drugs provided opportunities to treat breast cancer patients through different mechanisms. The result demonstrated the ability of our system to cooperate with bench work.

II. METHOD

Here we propose an analysis framework to identify putative drugs for breast cancer and also investigate the

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

effected mechanisms. The framework contains two major components: 1) the 'gene sets - drugs' association network and 2) exploration of tumor progression associated gene sets (TPA-GSs). 1) The 'gene sets - drugs' associations were generated through the gene set enrichment analysis of the expression profiles of drug treatments. A novel enrichment method was proposed for multiple times treatments. 2) To identify TPA-GSs, the Cox hazard proportional model was utilized to estimate the hazards of the gene sets. The gene sets of which *p*-value passed a selecting criterion were identified as TPA-GSs. By integrating the information of the two components, the putative drugs will be discovered. The details and mathematical models are described below.

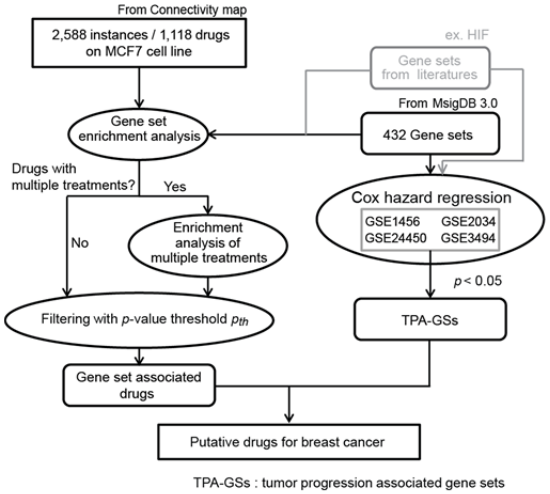

Figure 1. The flow chart to identify the putative drugs for breast cancer. The gene set was applied to gene set enrichment analysis. For the drugs with multiple data, an enrichment analysis of multiple treatments was applied. The gene set associated drugs were identified by filtering with thethreshold, p_{th}. The gene sets were also applied to Cox hazard regression model to identify the TPA-GSs. The putative drugs for breast cancer were identified by integrating the results.

Gene Set Enrichment Analysis

In order to identify the enrichments of gene sets for the expression of drug treatments, a scoring method was developed to quantify each expression instance. Suppose there are *N* genes in a given gene set. Let $\mathbf{x}_{j,l} = \{x_{1,l}, \ldots, x_{N,l}\}$, where $x_{j,l}$ is the $\log_2$-transformed expression level of gene *j* in the instance l, which was the expression profile of the cell line treated with the drug *g*. For the given gene set, the enriched score of the instance *l* is defined as,

$$s_l = \frac{1}{N}\sum_{j=1}^{N} z_{j,l} \quad (1)$$

, where $z_{j,l} = \left(x_{j,l} - \mu_j\right)/\sigma_j^*$,

μ_j is the mean of the untreated cell lines in the same experimental batch and σ_j^* is the standard deviation of gene *j* in the untreated expression profiles. To assess the statistical significance of the enriched score s_l, we provide a permutation based hypothesis tests to estimate the statistical significance based on the concept proposed in [21]. The null distribution *Q1* is generated by randomly selecting gene members of the gene set *D* times. Then the empirical *p*-values of the gene set is calculated as the fractions of the permutation value s_{Q1} exceed (or below) the value s_l :

$$p_{Q1}(s_l) = \#\{|s_{Q1}| > |s_l|\}/D \quad (2)$$

Based on the enriched score s_l, and the *p*-value, we can determine if the instance *l* of drug treatment has enrichment (or depletion) with the gene set.

Enrichment analysis of multiple treatment

For the drugs which have multiple treatments, we propose a method of cluster enriched score based on the concept of π-value [22] to conclude the overall enrichment. Suppose there are *W* clusters of drugs with multiple instances of treatment and the number of instances in the cluster *o* is y_o. The cluster enriched score for cluster *o* is defined as,

$$cs_o = \frac{1}{K}\sum_{i=1}^{y_o} p_i^* s_i \quad (3)$$

, where $p_i^* = -\log_{10}(p_i),\ K = \sum_{j=1}^{y_w}|p_j^*|$,

s_i is the enriched score of the *i*th instance in the cluster o, and p_i is the p-value of s_i. The statistical significance of the cluster signature score cs_o can be also tested by the permutation based hypothesis tests described above. The null distributions *Q2* are generated by randomly assigning y_o instances in the cluster *D* times. Then the empirical *p* values of the score cs_o is calculated as the fractions of the permutation value cs_{Q2} exceed (or below) the value s_l :

$$p_{Q2}(cs_o) = \#\{|cs_{Q2}| > |cs_o|\}/D \quad (4)$$

Construction of gene sets drug network

Propose there are G gene sets and H drugs. The enriched score and *p*-value of the gene set *p* for the drug *q* are defined as ES_{pq} and pES_{pq}. For the drugs that only have one instance of expression profile, the enriched score and *p*-value of the drug is the enriched score *s* and p_{Q1} of the instance. For the drugs with multiple treatments, the enriched score and *p*-value of the drug is set as the cluster enriched score *cs* and p_{Q2} of the cluster. By filtering with *p*-value threshold p_{th}, the 'gene sets – drugs' network is constructed and represented by a matrix **Y**.

$$Y_{pq} = \begin{cases} ES_{pq} & pES_{pq} < p_{th} \\ 0 & ,else \end{cases} \quad (6)$$

Survival Analysis of the Enriched score

To explore the tumor progression associated gene sets, the Cox hazard proportional model [23] was applied to

analyze the enriched scores. The enriched score of each patient was calculated as Equation (1) described above, and then the values were applied to Cox hazard regression model with survival information. The gene sets which had *p*-values of Cox model < 0.05 were identified as tumor progression associated gene sets. If the hazard rate > 1, the gene sets were defined as promoted TPA-GSs, otherwise they were defined as inhibited TPA-GSs. Kaplan–Meier plots were utilized to visualize the survival information. Two groups of patients was divided through 50% percentile of enriched scores, and then applied to Kaplan–Meier plots.

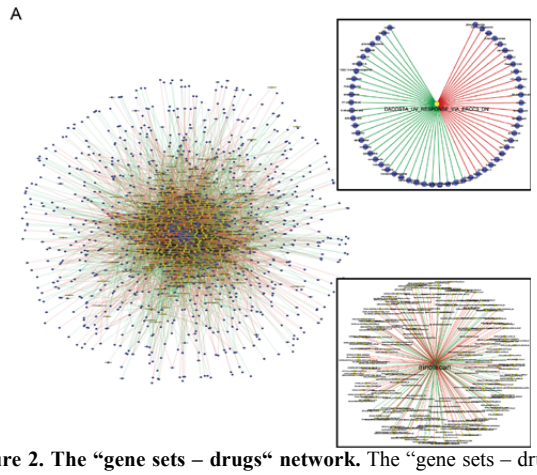

Figure 2. The "gene sets – drugs" network. The "gene sets – drugs" network was constructed. The red and green lines indicate the relationships of enrichment and depletion. The gene set, 'DACOSTA UV RESPONSE VIA ERCC3 DN' has the highest number, 57, of associated drugs. The drug irinotecan had then maximum number of associated gene sets.

III. RESULT

"Gene sets – drugs" assocaited netwrok

A total of 2,588 expression profiles treated by 1,118 kinds of drugs were downloaded from Cmap [12]. Only the expression profiles of MCF7 cell lines were used for the analysis. 432 gene sets which responded to different stimulation downloaded from CGP 3.0 of MSigDB were applied to our system. By applying those data to our analysis framework as described in the Method section, 10,249 "gene set – drug" associations were identified (Figure 2). 5,217 of them were enrichment and 5,032 were depletion. An average of 23.7 drugs (12.1 for positive and 11.6 for negative associations) were associated with one gene set. An average of 9.16 gene sets (4.7 for enrichment and 4.5 for depletion) were enriched in one drug. All of the gene sets have at least one associated drug and 116 drugs (10%) had no enriched gene sets. The gene set, 'DACOSTA UV RESPONSE VIA ERCC3 DN' has the highest number, 57, of associated drugs. The drug irinotecan had 217 associated gene sets.

Tumor progression associated gene sets

To identify the tumor progression associated gene sets (TPA-GSs), four expression datasets of breast cancers, GSE1456, GSE2034, GSE3494, and GSE24450 were used to investigate the prognosis of the gene sets through the Cox hazard proportional model. Using the *p*-values off the 4 data sets all < 0.05 as criteria, 25 gene sets were identified as TPA-GSs, which are listed in the right side of heatmap in the Figure 3. 24 of 25 TPA-GSs were defined as promoted TPA-GSs with a hazard ratio >1. Only the gene set 'YING SLIENCED BY DICER' was the inhibited TPA-GSs (HR<1).

Putative drugs for breast cancer

To find out putative anti-cancer drugs for breast cancer, the drugs with negative association with the 24 promoted TPA-GSs and the positive associated drugs for inhibited TPA-GSs were extracted (Figure 3). A total of 360 drugs were identified with the anti-cancer ability for breast cancer. By ranking the number of the associated TPA-GSs, 5 drugs, exverapamil, H-7, irinotecan, puromycin, and pyrvinium, showed the association with more the half of the TPA-GSs and are indicated as red lines in Figure 3. The numbers of associated TPA-GSs were 20, 15, 15, 13, and 13, respectively. The drug dexverapamil is the ABCB1 modulator, which has been used as a sensitizer of chemo therapy [24]. H-7 is a protein kinase inhibitor [25]. The chemo-drug irinotecan is an inhibitor of topoisomerase 1, which is an essential protein for cell mitosis. Puromycin is a kind of antibiotic which can inhibit translation. Pyrvinium is an old anthelminthic medicine and has been found as a potential Wnt inhibitor [26]. Although some of the drugs have not been used in cancer, our results showed those dugs affected TPA-GSs and have potential to be anti-cancer drugs for breast cancer.

The drug fulvestrant (indicated as a blue line) was used to treat breast cancer through the mechanism of inhibition of the estrogen receptor. However, our data showed it also affects the promoted TPA-GS 'LIANG SILENCED BY METHYLATION DN', which was derived from an experiment of inhibition of methylation. The result speculated that the anti-cancer effect of fulvestrant may also go through the inhibition of methylation. The mTOR inhibitor, sirolimus, and PI3K inhibitor, wortmannin, (green lines in Figure 3) were shown as having enrichments of the gene sets 'PENG RAPAMYCIN RESPSONE DN' and 'PENG LEUCINE DEPRIVATION DN', where they were derived from mTOR inhibition and nutrition deprivation [27]. The result demonstrated the accuracy of our result.

IV. CONCLUSION

In this study, we proposed an analysis framework to construct the "cancer - gene sets - drugs" network. Different from other methods which directly map the

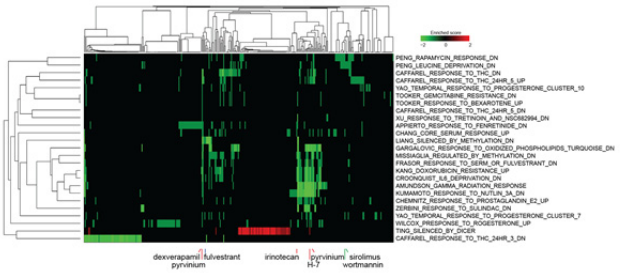

Figure 3. **The heatmap of enriched scores of TPA-GS associated drugs.** A total of 360 drugs were shown enrichment or depletion with the 25 TPA-GSs. For the inhibited TPA-GS, "TING SILENCED BY DICER", the drugs which have positive scores (red) were selected as putative drugs. For the other 24 promoted TPA-GSs, the drugs of negative scores (green) were chosen.

expression profiles of drug treatment to disease, the gene sets played an interpretation level for drug treatments and diseases. For drug treatment, the gene sets can be utilized to explore the cellular responses or altered pathways after drug stimulations. They were also used to identify the cellular responses or pathways correlated with tumor progression through patients' survival information. By integrating the two results, the putative drugs for breast cancer and the potential affected mechanism in tumor cells will be systemically discovered. Most available chemo-drugs for breast cancer only worked through the cell-cycle or apoptosis mechanisms. As demonstrated in the result section, we identified 25 TPA-GSs and also 360 associated drugs. Our result provided valuable information to revisit the drugs to "reposition" for cancer therapy through different mechanisms. Not limited to the response gene sets we used in the Result section, our method can extend to other gene sets, such as 'KEGG' pathway or microRNA target genes, to explore new putative drugs through new mechanisms to increase the response rate of cancer therapy. Our method also can be utilized for other diseases through the same principle to speed up the progression of drug development to clinical particle for those identified tumor associated mechanisms.

V. REFERENCES

[1] B. Rosenberg, *et al.*, "Platinum compounds: a new class of potent antitumour agents," *Nature,* vol. 222, pp. 385-6, Apr 26 1969.

[2] N. M. Cerqueira, *et al.*, "Understanding ribonucleotide reductase inactivation by gemcitabine," *Chemistry,* vol. 13, pp. 8507-15, 2007.

[3] B. L. Staker, *et al.*, "The mechanism of topoisomerase I poisoning by a camptothecin analog," *Proc Natl Acad Sci U S A,* vol. 99, pp. 15387-92, Nov 26 2002.

[4] C. A. Hudis, "Trastuzumab--mechanism of action and use in clinical practice," *N Engl J Med,* vol. 357, pp. 39-51, Jul 5 2007.

[5] R. Sordella, *et al.*, "Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways," *Science,* vol. 305, pp. 1163-7, Aug 20 2004.

[6] G. Kesisis, *et al.*, "Angiogenesis inhibitors. Drug selectivity and target specificity," *Curr Pharm Des,* vol. 13, pp. 2795-809, 2007.

[7] M. A. Postow, *et al.*, "The antitumor immunity of ipilimumab: (T-cell) memories to last a lifetime?," *Clin Cancer Res,* vol. 18, pp. 1821-3, Apr 1 2012.

[8] A. Kamb, *et al.*, "Why is cancer drug discovery so difficult?," *Nat Rev Drug Discov,* vol. 6, pp. 115-20, Feb 2007.

[9] J. Li, *et al.*, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts," *PLoS Comput Biol,* vol. 5, p. e1000450, Jul 2009.

[10] A. Gottlieb, *et al.*, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Mol Syst Biol,* vol. 7, p. 496, 2011.

[11] H. Luo, *et al.*, "DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome," *Nucleic Acids Res,* vol. 39, pp. W492-8, Jul 2011.

[12] J. Lamb, "The Connectivity Map: a new tool for biomedical research," *Nat Rev Cancer,* vol. 7, pp. 54-60, Jan 2007.

[13] J. Lamb, *et al.*, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science,* vol. 313, pp. 1929-35, Sep 29 2006.

[14] D. Shigemizu, *et al.*, "Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer," *PLoS Comput Biol,* vol. 8, p. e1002347, Feb 2012.

[15] F. Iorio, *et al.*, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proc Natl Acad Sci U S A,* vol. 107, pp. 14621-6, Aug 17 2010.

[16] M. Sirota, *et al.*, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Sci Transl Med,* vol. 3, p. 96ra77, Aug 17 2011.

[17] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PLoS One,* vol. 4, p. e6536, 2009.

[18] Y. Benita, *et al.*, "An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia," *Nucleic Acids Res,* vol. 37, pp. 4587-602, Aug 2009.

[19] A. Loboda, *et al.*, "EMT is the dominant program in human colon cancer," *BMC Med Genomics,* vol. 4, p. 9, 2011.

[20] A. Subramanian, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A,* vol. 102, pp. 15545-50, Oct 25 2005.

[21] L. Tian, *et al.*, "Discovering statistically significant pathways in expression profiling studies," *Proc Natl Acad Sci U S A,* vol. 102, pp. 13544-9, Sep 20 2005.

[22] Y. Xiao, *et al.*, "A Novel Significance Score for Gene Selection and Ranking," *Bioinformatics,* Feb 9 2012.

[23] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society Series B-Statistical Methodology,* vol. 34, pp. 187-&, 1972.

[24] M. Lehnert, *et al.*, "Phase II trial of dexverapamil and epirubicin in patients with non-responsive metastatic breast cancer," *Br J Cancer,* vol. 77, pp. 1155-63, Apr 1998.

[25] T. Volberg, *et al.*, "Effect of protein kinase inhibitor H-7 on the contractility, integrity, and membrane anchorage of the microfilament system," *Cell Motil Cytoskeleton,* vol. 29, pp. 321-38, 1994.

[26] C. A. Thorne, *et al.*, "Small-molecule inhibition of Wnt signaling through activation of casein kinase 1alpha," *Nat Chem Biol,* vol. 6, pp. 829-36, Nov 2010.

[27] T. Peng, *et al.*, "The immunosuppressant rapamycin mimics a starvation-like signal distinct from amino acid and glucose deprivation," *Mol Cell Biol,* vol. 22, pp. 5575-84, Aug 2002.