Yao et al. BMC Systems Biology 2010, 4:151
http://www.biomedcentral.com/1752-0509/4/151

BMC Systems Biology

RESEARCH ARTICLE Open Access

Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis

Chen Yao[1], Hongdong Li[1], Chenggui Zhou[1], Lin Zhang[1], Jinfeng Zou[1], Zheng Guo[1,2*]

Abstract

Background: It has been suggested that, in the human protein-protein interaction network, changes of co-expression between highly connected proteins ("hub") and their interaction neighbours might have important roles in cancer metastasis and be predictive disease signatures for patient outcome. However, for a cancer, such disease signatures identified from different studies have little overlap.

Results: Here, we propose a systemic approach to evaluate the reproducibility of disease signatures at multiple levels, on the basis of some statistically testable biological models. Using two datasets for breast cancer metastasis, we showed that different signature hubs identified from different studies were highly consistent in terms of significantly sharing interaction neighbours and displaying consistent co-expression changes with their overlapping neighbours, whereas the shared interaction neighbours were significantly over-represented with known cancer genes and enriched in pathways deregulated in breast cancer pathogenesis. Then, we showed that the signature hubs identified from the two datasets were highly reproducible at the protein interaction and pathway levels in three other independent datasets.

Conclusions: Our results provide a possible biological model that different signature hubs altered in different patient cohorts could disturb the same pathways associated with cancer metastasis through their interaction neighbours.

Background

Analysis of gene expression patterns in cancers has greatly enhanced our understanding of the biology of cancer and provided a way to improve the prediction of many cancers. For example, many signature genes have been extracted from microarray data to predict the outcome of breast cancer [1-4]. However, for a particular disease, signature genes identified from different studies are usually highly inconsistent, raising doubts about the biological significance or clinical implication of the signatures identified [5-7]. In attempts to tackle this problem, many approaches have been proposed for the extraction of network-based disease signatures based on protein-protein interaction (PPI) data. Notably, because the PPI data is a union of the interactions activated under various conditions, and currently includes a lot of false positives, it alone can provide limited information for discriminating interactions in different biological pathways such as signal transduction pathways. On the other hand, considering that gene expression is sensitive to disease conditions, it is reasonable to combine gene expression data with PPI data to measure the'activity'of PPI subnetworks in response to the investigated conditions and such subnetworks are often suggestive of functional signaling cascades, metabolic pathways and molecular complexes that are associated with the disease phenotypes [8-11]. For example, Chuang *et al.* identified PPI subnetworks with coherent gene expressions as disease signatures that were suggested to be more accurate than single gene signatures for predicting breast cancer metastasis [11]. However, the subnetworks identified from different datasets were still highly inconsistent [12]. Recently, Taylor *et al.* searched for changes in the global modularity of the human interactome and found that patients who survived breast cancer had an organization of the PPI network different from that in patients who died of the illness [13]. Specifically, they suggested that "hub" proteins with altered co-expression

* Correspondence: guoz@ems.hrbmu.edu.cn
[1]Bioinformatics Centre and Key Laboratory for NeuroInfomation of the Education Ministry of China, School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China
Full list of author information is available at the end of the article

relation with their interaction partners can be used as robust signatures to predict cancer outcome. However, as shown here, such signature hubs selected from different studies for breast cancer metastasis have little overlap.

This irreproducibility problem is usually attributed to deficiency in experimental designs, different platforms and statistical analyses of disease signatures [14,15]. However, it is very likely that the inconsistency of disease signatures discovered from different cancer samples for a particular cancer might reflect the biological variation and heterogeneity of the cancer [5,16]. It is becoming increasingly clear that, for a particular cancer, genetic and epigenetic changes in different patients are extremely heterogeneous. Especially, as demonstrated in recent high-throughput screens of somatic mutation of genes in cancer genomes, the vast majority of gene mutations are different among patients with a particular cancer [17-22]. It is also becoming clear that diverse molecular changes in cancers might actually be consistent in some essential cellular functions (hallmarks) whose alterations might collectively dictate malignant growth for almost all human cancers [23,24]. Therefore, it is reasonable to design scores to evaluate the reproducibility of disease signatures of cancers at multiple levels based on some biological assumptions (or molecular models), taking into account functional relations between the disease signatures such as expression correlation [16] and functional similarity [25]. If a score is significantly higher than expected by chance, it provides statistical evidence that the underlying model could correctly explain a large fraction of diverse but functionally related disease signatures. In this sense, the biological assumptions for designing the scores are testable.

Here, we propose a systemic approach to evaluate the reproducibility of network-based disease signatures derived for a particular cancer, taking into account their functional relations. Specifically, we evaluated the reproducibility of signature hubs for characterizing the changes of global modularity of the human interactome for breast cancer metastasis [13]. First, based on the assumption that proteins with similar interaction neighbours are likely to have similar biological functions [26,27], we proposed a topological overlap score, the percentage of overlap based on topology similarity (POT) score, to measure the reproducibility of signature hubs detected in different datasets. Using the POT score, we found signature hubs detected in two datasets for breast cancer metastasis were highly consistent in terms of frequently sharing neighbourhood proteins in the human PPI network and displaying consistent co-expression changes with the overlapping neighbours. Then, we showed that the interaction neighbour proteins shared by the two lists of signature hubs from the two datasets tended to be cancer susceptibility genes and affect some pathways known to be associated with breast cancer pathogenesis, indicating that these pathways might have important diagnostic and therapeutic implications. Finally, we proved that these results were highly reproducible in three other independent datasets for breast cancer metastasis.

Results

Network topology consistency of the hub protein lists

We first searched for signature hubs whose co-expressions with their interacting partners were significantly different between patients labelled non-metastatic and metastatic. We used the method proposed by Taylor *et al*. [13], as described briefly in *Methods*, in the dataset (the Wang dataset) compiled by Wang *et al*. [28] and in the dataset (the Desmedt dataset) compiled by Desmedt *et al*. [29]. Here, we did not apply the FDR control at the step of finding signature hubs because the statistical powers of most multiple test adjustment methods are decreased in the presence of wide and correlated expression changes of genes in cancers [30,31]. Instead, we used a P value of 0.01 to find candidate signature hubs, as in the work by Taylor *et al*. [13]. With $P < 0.01$, we identified a total of 65 and 72 signature hubs in the Wang dataset and Desmedt dataset, respectively (See Additional file 1-Table S1 for the signature hubs.). Only 4 signature hubs appeared in both datasets and the percentage of overlaps (PO) score of the hub lists was only 5.9%. Thus, at the level of individual proteins, the signature hubs detected in different studies were extremely inconsistent, although the PO score was significantly larger than expected by chance alone (hypergeometric test $P = 0.027$).

Then, we evaluated the reproducibility of two lists of signature hubs by the POT score which measures the percentage of overlapped interaction neighbours of signature hubs extracted from different studies (see *Methods*). First, by the hypergeometric distribution model, with FDR < 0.05, we tested whether the interaction neighbours of a hub in a list overlapped significantly with the neighbours of at least one of the hubs in another list. Then, considering that signature hubs with significant neighbourhood overlaps might have similar functional roles, we calculated the POT score for two lists of signature hubs. The POT score between the lists of signature hubs extracted from the Wang dataset and the Desmedt dataset was as high as 73%.

Next, we did three random experiments to test whether the increased overlap might be introduced by some factors irrelevant to the disease status. First, for each dataset, we assigned phenotype labels randomly to patients to generate expression data with the same correlation structure as the original dataset, and then searched for signature hubs in the PPI network by the

approach used with the real data. Because the phenotype information was randomised, the detected signature hubs should be irrelevant to disease status. Repeating this process 1000 times, we found the average of the POT scores for the random pairs of protein lists was 41%, which was significantly smaller than the score (73%) observed with the real data ($P < 0.005$). Second, we tested whether the increased reproducibility might be due to the network topology. From the same PPI network, we randomly selected 1000 pairs of protein lists with the same lengths as the signature hub lists and then computed their POT scores. The average of the POT scores for these random pairs of protein lists was 44%, which was significantly smaller than that observed ($P < 0.005$). Third, we tested whether the high level of reproducibility might be due to the high degrees (numbers of interaction partners) of signature hubs. Using a local rewiring algorithm [32], we produced 1000 random PPI networks in each of which all proteins had exactly the same connectivity as in the original PPI network and the choice of their interaction partners was random. Then, from each random network we selected the pairs of hub lists that had exactly the same lengths and degree distributions as the two lists of signature hubs extracted from the actual PPI network. Then, we recalculated the POT score for this random pair of hub lists. This process was repeated 1000 times. The average POT score for 1000 pairs of random hub lists was 42%, significantly smaller than that observed ($P < 0.005$).

Both false negatives and false positives are concerned for the PPI data quality [33,34]. To tackle the low coverage problem introduced by false negatives, we integrated 8 databases to generate a large PPI network for our study. To reduce the effect of false positives, we also used a small PPI network which contained only the hand-curated PPI interaction data from OPHID [35] and MINT [36]. The POT score was decreased a little to 62% due to the smaller network size based on this PPI dataset. However, the POT score was significantly higher than those (20%, 29% and 17%) based on each of the three random experiments described above ($P < 0.005$), respectively. Two PPI networks generated similar POT scores, suggesting that our results were rather robust against false negatives and false positives in the PPI data.

Pathway consistency of the hub protein lists

If two signature hubs share many interaction neighbour proteins, then they might participate in the same or similar functions [26,27]. To reveal the consistency of signature hub lists at the pathway level, for each signature hub identified from each dataset, we analysed the enrichment of its interaction neighbours in pathways collected in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [37] (see *Methods*). With FDR < 0.01, we found that 34 pathways were enriched significantly with the neighbours of at least one of the signature hubs detected in the Desmedt dataset, among which 26 pathways were included in the 38 significant pathways detected in the Wang dataset (See Additional file 1-Table S2 for the list of 26 pathways.). Notably, among the other 12 pathways detected in the Wang dataset but not in the Desmedt dataset, 11 were marginally significant in the Desmedt dataset with $P < 0.05$. Similarly, among the 8 pathways detected in the Desmedt dataset but not in the Wang dataset, 6 were marginally significant in the Wang dataset with $P < 0.05$. Thus, some inconsistency between the two datasets might come from a reduction of the statistical power by using the stringent FDR control for adjusting multiple tests when the multiple tests are not independent of each other [30,31].

We did a random experiment to test the significance of the high concordance of pathway enrichment (see *Methods*). First, we took the 38 pathways identified from the Wang dataset as the gold standard. From each of the random networks produced by a local rewiring algorithm [32], we extracted a random hub list of the same length and degree distribution with the list of signature hubs identified from the Desmedt dataset. Then, we detected the pathways enriched with the neighbours of random hubs and compared them with the gold standard. Repeating this process 1000 times, we found the average number of overlapping pathways was 1, significantly fewer than the 26 overlaps observed in the real data ($P < 0.001$). The result was the same when taking the pathways detected from the Desmedt dataset as the gold standard.

The 26 pathways detected in both datasets included many pathways known to be deregulated in breast cancer pathogenesis, such as cell cycle, apoptosis, Jak-STAT, MAPK, ErbB, Wnt and P53 signalling pathways [38]. Among these 26 pathways, there were 191 and 238 interaction neighbours of the signature hubs identified from the Wang and Desmedt datasets, respectively, and they shared 114 proteins, which was significantly more than expected by chance alone (hypergeometric test $P < 2.2 \times 10^{-16}$). These common interaction neighbour proteins might have important roles in cancer. To test this, we assembled a list of 427 cancer susceptibility genes from the Cancer Gene Census database [39] and found 50 out of 114 neighbour proteins were known cancer proteins (hypergeometric test $P = 6 \times 10^{-4}$). When using the 685 genes collected in our F-census database [25], 100 out of 114 neighbour proteins were included (hypergeometric test with $P < 2.2 \times 10^{-16}$).

The above results suggested that the two lists of signature hubs might affect the same pathways. In one situation,

in different cohort patients, a cancer-associated pathway could be affected by the co-expression changes of different signature hubs with the same set of neighbours enriched in this pathway. For example (Figure 1a), the interleukins IL2 and IL6 were identified as signature hubs from the Wang and Desmedt datasets separately and their overlapped neighbours were enriched in the Jak-STAT signalling pathway. Thus, changes of co-expression of these shared neighbours with either IL2 or IL6 might disrupt the Jak-STAT signalling pathway and contribute to the progression of cancer [40]. For another example (Figure 1b), 6 signature hubs identified from the Wang dataset and another 3 signature hubs identified from the Dsemedt dataset are all subunits of a ribosome complex for protein biosynthesis. They share other subunits as interaction neighbours and their deregulation might be associated with cell growth and proliferation [41]. In another situation, a cancer-associated pathway could be affected by changes of different signature hubs interacting with different sets of neighbours that were separately enriched in this pathway. For example (Figure 1c), proteins DUSP3 with degree 18 and CAD with degree 39 were identified as signature hubs in the Wang and Desmedt datasets separately. The neighbours of each of these two proteins were enriched in the MAPK signalling pathway associated with cancer metastasis [42], but their neighbours shared only 1 protein. It has been suggested that DUSP3 can negatively regulate members of the MAP kinase superfamily (MAPK) [43], while the deregulation of CAD proteins might be associated with activation of the MAPK cascade [44]. Notably, this functional relation between two signature hubs was not reflected by the POT score, which considers only overlapping neighbours between the signature hubs (see *Discussion*).

Co-expression consistency of the hub proteins lists

Considering that a signature hub disturbs functions through differential co-expression with their interaction neighbours [13], we further assumed that two functionally similar hubs should display consistent co-expression changes with their overlapping neighbours across different datasets [45,46]. Therefore, for two hubs detected from two datasets separately, we additionally tested the consistency of the directions of their correlations with the shared neighbours across the datasets by the Bernoulli distribution model (see *Methods*).

With the co-expression restriction, for Wang and Desmedt dataset, the POT score (denoted as POG-e score) decreased a little from 73% to 67%, largely explainable when considering that any extra restriction may miss some true relations. On the other hand, the random POT-e score decreased greatly from 44% to 26%. The results suggested that signature hubs sharing neighbours were significantly consistent in the change directions of correlations with their shared neighbors. For example, from the Wang and Desmedt datasets separately, the interleukins IL2 and IL6 were identified as signature hubs and their 6 overlapped neighbours were enriched in the Jak-STAT signaling pathway. In both Wang and Desmedt datasets, the expressions of IL2 and IL6 were both positively correlated with the expressions of these shared neighbours in non-metastatic patients, but negatively correlated with the expressions of the shared neighbours in metastatic patients. These results suggest that Jak-STAT signaling pathway could be perturbed by the disruption of co-expressions of either IL2 or IL6 with the shared neighbours during the breast cancer metastasis.

Validation in three independent breast cancer datasets

We validated our results by analyzing three other independent datasets for breast cancer metastasis [2,47,48]. For lists of signature hubs extracted from every two breast cancer datasets, the PO score was less than 4%. However, the corresponding POT scores took values ranging from 61% to 75% which were all significantly larger than expected by chance according to the three random experiments as described in *Methods*. Similar results were observed based on the POT-e score ($P < 0.005$, see Additional file 1- Table S3 for details).

For example, 80 signature hubs were identified from the Vijver dataset, among which only 4 and 1 overlapped with the signature hubs found in the Wang and Desmedt datasets, respectively. However, the corresponding POT scores were 64% and 75%, respectively, and they were both significantly larger than expected by chance ($P < 0.005$), according to each of the three random experiments as described in *Methods*. Notably, although the average POT score between the Wang and Vijver datasets was only 64%, the POT score for the signature hub list extracted from the Vijver dataset to the signature hub list extracted from the Wang dataset was 71%, suggesting that many of the signature hubs detected from the Vijver dataset could be represented by the signature hubs from the Wang dataset in terms of neighbourhood similarity. The score in the opposite direction was only 57%, indicating that the samples used in the Vijver dataset might be insufficient for capturing enough signature hubs to cover the signature hubs extracted from the Wang dataset.

According to pathway enrichment analysis, the signature hubs extracted from the Vijver dataset and those from both the Wang dataset and the Desmedt dataset were highly consistent. Among the 26 pathways shared by the Wang and Desmedt datasets, 19 were included in the 34 pathways identified from the Vijver dataset, significantly more than expected by chance alone (hypergeometric test $P = 5.2 \times 10^{-5}$). All the other 7 pathways detected in both the Wang and Desmedt datasets were marginally

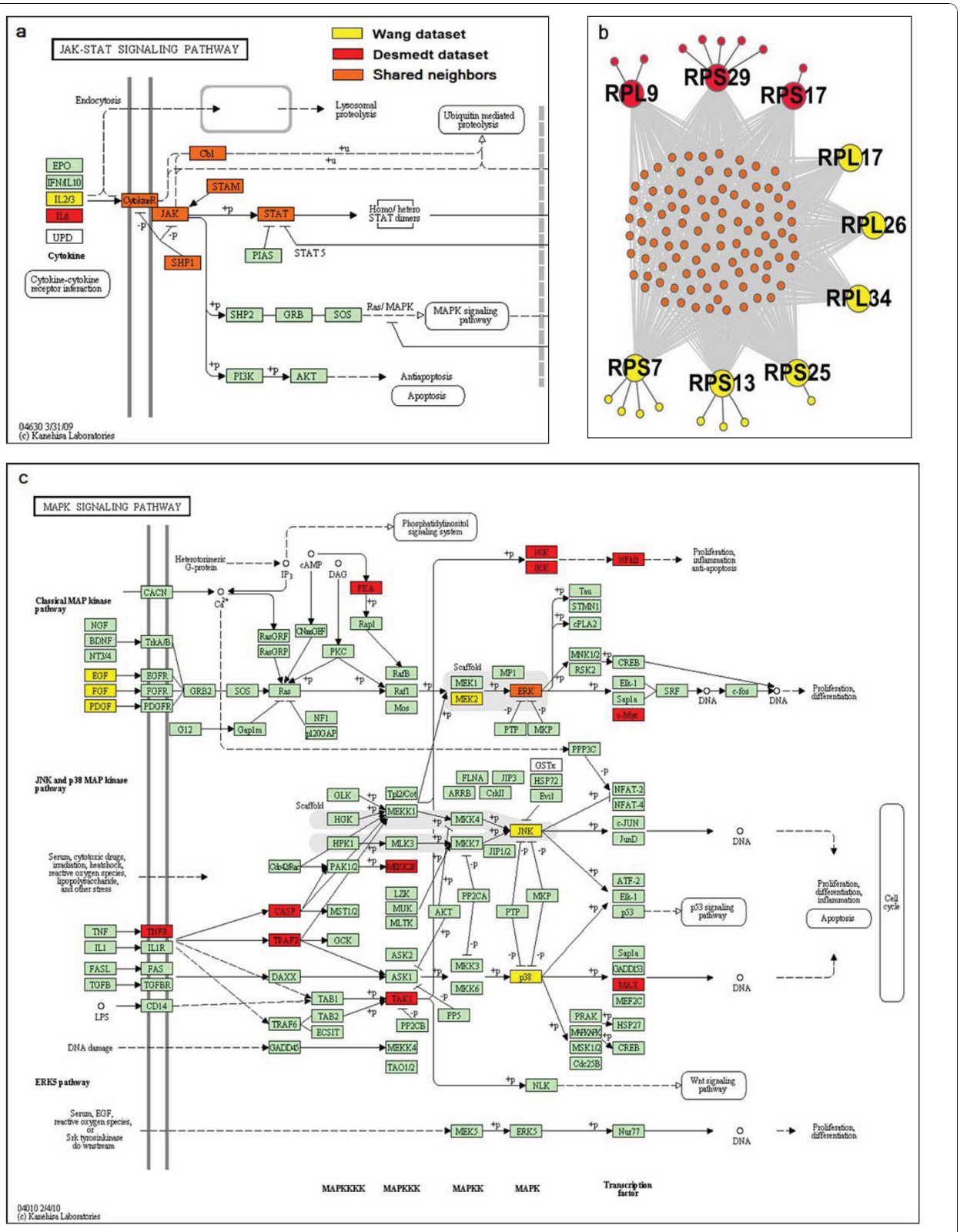

Figure 1 Examples of pathways shared by a signature hub from the Wang dataset and a signature hub from the Desmedt dataset. (a) JAK-STAT signaling pathway; (b) Ribosome complex; (c) MAPK signaling pathway. The yellow and red colors represent proteins (both hubs and their neighbors) identified from the Wang and Desmedt datasets, respectively. The orange colors represent the overlapped neighbors of these two hub proteins. Please see the main text for detailed explanation.

significant in the Vijver dataset with $P < 0.05$. These results indicated that these pathways, such as MAPK signaling and apoptosis pathways which were also founded in other studies [11,49], might be disturbed in the breast metastatic progression.

The above results confirmed that signature hubs detected from different datasets for breast cancer metastasis were reproducible in terms of neighbourhood protein overlap and, more generally, pathway overlap. Notably, approximately half of the patients in the Vijver dataset were lymph node-positive and underwent adjuvant therapy before expression profiling, whereas all patients in the Wang dataset had lymph node-negative breast cancer [11]. However, our results indicated that the two types of samples might have similar molecular changes at the pathway level.

Discussion

Changes in the global modularity of the human interactome might provide important insights into the mechanism underlying cancer metastasis [13]. As shown in this study, although signature hubs detected from different studies for breast cancer metastasis have little overlap, they are highly consistent in terms of frequently sharing interaction proteins and displaying consistent co-expression changes with their overlapping neighbours, indicating that they might alter the same pathways through differential co-expression with their interaction neighbours. To some extent, this finding is similar to the observation made in microRNA studies that a cancer pathway could be changed in cancer cases by either aberration of some cancer genes or modification of microRNAs regulating these genes [50]. Recently, using several microarray datasets, Li *et al*. identified gene signature modules with high predictive accuracy for breast cancer metastasis [49]. These modules contained two parts: a set of signature genes that are dynamically modulated between 'high-risk' and 'low-risk' patients and a unique set of cancer driver-mutating genes that are the direct protein interacting partners of the signature genes. At the conceptual level, their results also suggested that, despite low overlap, disease signatures detected from different datasets may reflect consistent function disruptions. Especially, many modules identified by Li *et al*., such as cell cycle, apoptosis and immune response, were functionally consistent with our KEGG pathways enriched with proteins targeted by different signature hubs.

The POT score proposed in this paper considers the functional concordance between signature hubs only according to their overlapping neighbours. The significantly high POT scores between signature hubs derived from different studies for breast cancer metastasis indicates that the biological assumption included in this score could explain a large fraction of diverse signature hubs. However, the POT scores were only about 70% for the five datasets in this study and some inconsistent signature hubs could not be explained by this model. One explanation is that the incomplete PPI data might be insufficient for capturing all functional links among signature hubs. Another possibility is that there might be other molecular models that can explain the remaining inconsistent discoveries. For example, as illustrated by a case presented in *Results*, two signature hubs with non-overlapping neighbours might be functionally consistent if their neighbours are enriched separately in the same pathway, but such a functional relation is not measured by the POT score. Principally, we could further consider this and other possible relations of signature hubs to reveal the consistency of signature hub lists. For example, we could evaluate the consistency of signature hub lists at the pathway level by counting overlaps of the enrichment pathways associated with different signature hub lists. However, such a pathway level analysis might have only a limited application scope because many proteins have not been annotated to current pathway databases such as KEGG used in this study. The limited annotation can reduce the power of finding true enrichment pathways and introduce some inconsistency [11]. More problematically, pathways defined in current databases are often inconsistent and their boundaries are unclear [51]. For example, it is possible that a pathway documented in a pathway database consists of several sub-pathways, and only alterations of genes within a sub-pathway have the same or a similar role in cancer development, and the genes within the other sub-pathways might be irrelevant to, or have other roles, in the disease mechanism. In such a situation, it would be ambiguous if we consider two signature hubs as functionally equivalent (reproducible) when they are associated with different parts of the pathway through their interaction proteins. Thus, to interpret the consistency of signature hubs at the pathway level, we need to determine pathways or their sub-pathways that are most relevant to a disease. Compared with KEGG and other pathway databases, Gene Ontology (GO) [52] could help us tackle this problem to some degree because it describes biological functions from general to specific in a hierarchy. However, currently, it is still a difficult task to treat the redundant annotations in GO properly [53,54] and this problem deserves future research efforts [51,55]. Thus, currently, the pathway analysis can only partially support the POT score analysis. When the pathway definition and gene annotation are improved, the pathway analysis will become an efficient way of explaining inconsistent signatures generated from different studies.

The irreproducibility of molecular signatures detected for a complex disease is also a common problem in

many other research areas based on high-throughput biotechnology such as proteomics [56] and metabolomics [57]. Also, it is very likely that the small samples typically used in current studies of these areas might reflect the wide and diverse molecular changes in a complex disease only partially. In general, taking into account the diverse but correlated molecular changes in a complex disease such as a cancer, our approach provides a framework for explaining the reproducibility of biological findings at the systems biology level. However, even when we could find functionally consistent disease signatures from currently available samples, it might still need thousands of samples to find a few reproducible individual signatures. Thus, it would be a difficult task to build a consensus prognostic classifier on the basis of a few signatures for a complex disease [6]. To circumvent the difficulty of finding consistent signatures themselves, we could use some biological pathways commonly affected by diverse molecular changes as modular signatures to build robust diagnostic classifiers [58]. The identification of such clearly defined key pathways of cancer metastasis might provide crucial guidance for designing diagnostic classifiers and, perhaps, appropriate drug combinations [59].

Conclusions

Distant metastases are the major cause of death in cancer patients. The heterogeneous nature of tumours leads to different responses from different patients with the same type of cancer. Therefore, as a sign that two studies have detected the same result for a disease, it is not necessary that the signature lists themselves are consistent. They could be probably tracking a common set of biologic phenotype, as we shown here, in protein network, signature hubs with low reproducibility may actually have similar functions by interacting with the same sets of neighbour proteins.

Methods

Datasets

Five datasets of gene expression profiles for breast cancer metastasis are described in Table 1. Patients who had been detected metastasis within 5 years during follow-up visits were assigned to metastatic and the remaining patients were assigned to non-metastatic. We mainly analyzed the Wang dataset [28] and Desmedt dataset [29], and the other three datasets [2,47,48] were used for validation.

The human PPI data were downloaded from MINT [36], BIND [34], IntAct [60], HPRD [61], MIPS [62], DIP [63], KEGG (PPrel for protein-protein interactions, ECrel for enzymes involved in neighboring steps) [64] and Reactome [65]. To increase the coverage of the PPI network, we pooled together these 8 PPI datasets to construct an integrated PPI network that consists of 101,729 distinct interactions involving 12,372 human proteins [66]. We restricted our analysis to the 5470 genes encoding proteins in this PPI network and presenting in all five breast datasets. We also did the analysis using the OPHID data [35] combined with MINT data [36], which was the same as the PPI data used by Taylor *et al.* [13].

The 60 pathways analysed here were collected from the categories "Environmental Information Processing" and "Cellular Processes" in the KEGG database [37] at March 2010.

Selection of signature hubs

As Taylor *et al.* did, proteins with at least 3 interaction neighbours in the PPI network were defined as hubs and used for further study [13]. To determine the difference of co-expression of a hub with its interaction partners between metastatic and non-metastatic patients in a dataset, we calculated the Pearson correlation coefficients (PCCs) between this hub and its interaction partners in each patient group and then the absolute difference of the PCCs between two groups. We randomly permuted patients in the two groups 1000 times to calculate the random distribution of the absolute difference of PCCs between the groups. Then the real absolute difference of the PCCs for this hub between patient groups was compared to the random distribution to generate its P value. Hub proteins with P values < 0.01 were selected as candidate signature hubs, also referred to as signature hubs for short in the text.

Multi-level evaluation of reproducibility

In the following, we describe some scores for measuring the consistency between signature hubs derived from different studies for breast cancer metastasis.

At the individual protein level, the consistency of two lists of signature hubs was measured by the percentage of overlaps (PO) score [30]. Suppose hub list 1 with length L_1 and list 2 with length L_2 share k proteins, then the POG score from list 1 (or 2) to list 2 (or 1) is:

$$\mathrm{PO}_{12} = k / L_1$$

$$\mathrm{PO}_{21} = k / L_2$$

The average of the scores in the two directions is:

$$\mathrm{PO} = \left(\mathrm{PO}_{12} + \mathrm{PO}_{21} \right) / 2$$

Based on the assumption that proteins with significantly overlapped interaction neighbours are likely to share the same function [26,27], we designed a score, named the percentage of overlap based on topology (POT) similarity score, to measure the consistency of two lists of signature hubs at the PPI topology similarity

Table 1 The five datasets analyzed in this study

Datasets [a]	No. of Patients	metastatic	non-metastatic	Platforms
Wang dataset [28]	286	106	180	Affymetrix HG-U133a
Desmedt dataset [29]	198	35	163	Affymetrix HG-U133a
GSE1456 [47]	159	40	119	Affymetrix HG-U133a
GSE3494 [48]	218	37	181	Affymetrix HG-U133a
Vijve dataset [2]	295	78	217	Agilent Hu25K

[a] Datasets were named according to the authors or GEO numbers.

level. Let n and m be the number of neighbours interacting with proteins i and j, respectively, and g is the number of neighbours shared by these two proteins, the probability P of observing no fewer than g neighbours shared by proteins i and j by chance was calculated by the hypergeometric distribution model as [67]:

$$P = 1 - \sum_{i=0}^{g-1} \frac{C_n^i C_{N-n}^{m-i}}{C_N^m}$$

where N is the number of proteins with both PPI and gene expression data. With FDR < 0.05, the P value was adjusted by the Bonferroni-Hochberg procedure to account for multiple tests [68].

Then, let T_{12} (or T_{21}) be the number of proteins in list 1 (or list 2) whose neighbours are overlapped significantly with the neighbours of at least one of the proteins in list 2 (or list 1), then the POT score is defined as the average of the scores in the two directions:

$$\mathrm{POT} = \left(\mathrm{POT}_{12} + \mathrm{POT}_{21}\right)/2$$

The score in one direction is:

$$\mathrm{POT}_{12} = \left(k + T_{12}\right)/L_1$$

and the score in the other direction is:

$$\mathrm{POT}_{21} = \left(k + T_{21}\right)/L_2$$

The significance of a PO score was calculated as the probability of observing at least k overlapped proteins by chance using the hypergeometric probability model [30]. The significance of an observed POT score was assessed by three random experiments, testing whether the POT score might be due to (1) the correlation structures of expression profiles, (2) the PPI network topology or (3) the degree distribution of the signature hubs. The details of the experimental procedures are described in *Results*.

Both the PO and POT scores are dependent on the list lengths. In this study, our major objective was to find consistent results from the hub lists detected from different studies for a disease and we did not intend to compare the consistency level of hub lists with different lengths. Thus, we did not normalize PO or POT scores as we did in our earlier work [16].

Considering the basic assumption that signature hubs may disturb functions through differential co-expression with their interaction neighbours [13], we further assumed that two functionally similar hubs should display consistent co-expression changes with their overlapping neighbours across different datasets [45,46]. Let n be the number of neighbours shared by two signature hubs detected separately from two datasets, and k is the number of these shared neighbours whose correlation changes with these two hubs are in the same direction across the two datasets. Then, the probability P of observing no fewer than k neighbours with the same directions of correlation changes by chance was calculated by the Bernoulli distribution model as:

$$P = 1 - \sum_{i=0}^{k-1} \binom{n}{k-1} p^{k-1}(1-p)^{n-k+1}$$

With $P < 0.05$, we calculated the POT score with co-expression restriction, denoted as POT-e. Then, we did a random experiment to determine if an observed POT-e score is significantly larger than expected by chance when the change directions of correlations between hubs and their shared neighbours are irrelevant to the disease conditions. We randomly reassigned phenotype labels of samples to generate expression data with the same correlation structure as the original data, and then recalculated the POT-e score. This process was repeated 1000 times and the P value was calculated as, among the scores of the 1000 datasets with random phenotypes, the proportion of the scores exceeding the observed one.

Pathway enrichment analysis

For each signature hub, we detected pathways enriched with its interaction neighbours by the hypergeometric probability model [67]. The P value was adjusted by the Bonferroni-Hochberg procedure with FDR < 0.01 [68].

For a disease, we took the pathways enriched with the neighbours of signature hubs detected from one dataset as the gold standard, and then calculated the overlap with the pathways enriched with the neighbours of signature hubs detected from another dataset. To test the significance of the observed overlapping, we produced a random network by using a local rewiring algorithm [32]. In the random PPI network, all proteins had exactly the same connectivity as in the original PPI network and the choice of their interaction partners was random. Then, from the random network we selected a random protein list with the same length and degree distribution as the list of signature hubs identified from another dataset. Then, with FDR < 0.01, we detected the pathways enriched with random hubs and compared them with the gold standard. Repeating this process 1000 times, we calculated the *P* values for the observed overlaps.

Additional material

Additional file 1: Supplemental Tables. This file contains Tables S1-S3. Table S1 Two hub protein lists separately identified from Wang and Desmedt datasets. Table S2 List of 26 common KEGG pathways. Table S3 POT and POT-e scores for five breast cancer datasets

Authors' contributions

CY and ZG made contributions to the concepts, CY, HDL and CGZ carried out the analyses of the data. LZ and JFZ helped to interpret the results. CY and ZG drafted the paper, and all authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant NO: 30770558, 30970668, 81071646), Excellent Youth Foundation of Heilongjiang Province (grant No. JC200808) and Scientific Research Fund of Heilongjiang Provincial Education Department (NO: 11541156).

Author details

[1]Bioinformatics Centre and Key Laboratory for NeuroInfomation of the Education Ministry of China, School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China. [2]Colleges of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China.

Received: 26 June 2010 Accepted: 9 November 2010
Published: 9 November 2010

References

1. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, *et al*: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci USA* 2003, **100**:8418-8423.
2. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, 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 Engl J Med* 2002, **347**:1999-2009.
3. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, *et al*: **Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.** *Proc Natl Acad Sci USA* 2005, **102**:3738-3743.
4. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, *et al*: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351**:2817-2826.
5. 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**:171-178.
6. Ein-Dor L, Zuk O, Domany E: **Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.** *Proc Natl Acad Sci USA* 2006, **103**:5923-5928.
7. Dupuy A, Simon RM: **Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.** *J Natl Cancer Inst* 2007, **99**:147-157.
8. Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH: **Modular organization of protein interaction networks.** *Bioinformatics* 2007, **23**:207-214.
9. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach.** *Bioinformatics* 2008, **24**:i223-231.
10. Supper J, Spangenberg L, Planatscher H, Drager A, Schroder A, Zell A: **BowTieBuilder: modeling signal transduction pathways.** *BMC Syst Biol* 2009, **3**:67.
11. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Syst Biol* 2007, **3**:140.
12. Auffray C: **Protein subnetwork markers improve prediction of cancer outcome.** *Mol Syst Biol* 2007, **3**:141.
13. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL: **Dynamic modularity in protein interaction networks predicts breast cancer outcome.** *Nat Biotechnol* 2009, **27**:199-204.
14. Ntzani EE, Ioannidis JP: **Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment.** *Lancet* 2003, **362**:1439-1444.
15. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365**:488-492.
16. Zhang M, Zhang L, Zou J, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang C, Guo Z: **Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes.** *Bioinformatics* 2009, **25**:1662-1668.
17. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, *et al*: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**:1108-1113.
18. Sjoblom 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**:268-274.
19. Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, *et al*: **A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer.** *Nat Genet* 2005, **37**:590-592.
20. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, *et al*: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**:153-158.
21. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, *et al*: **Somatic mutations affect key pathways in lung adenocarcinoma.** *Nature* 2008, **455**:1069-1075.
22. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458**:719-724.
23. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57-70.
24. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439**:353-357.
25. Gong X, Wu R, Zhang Y, Zhao W, Cheng L, Gu Y, Zhang L, Wang J, Zhu J, Guo Z: **Extracting consistent knowledge from highly inconsistent cancer gene data sources.** *BMC Bioinformatics* 2010, **11**:76.
26. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
27. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.

28. Wang Y, Klijn JG, 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**:671-679.
29. 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**:3207-3214.
30. Zhang M, Yao C, Guo Z, Zou J, Zhang L, Xiao H, Wang D, Yang D, Gong X, Zhu J, *et al*: **Apparently low reproducibility of true differential expression discoveries in microarray studies.** *Bioinformatics* 2008, **24**:2057-2063.
31. Carvajal-Rodriguez A, de Una-Alvarez J, Rolan-Alvarez E: **A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests.** *BMC Bioinformatics* 2009, **10**:209.
32. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296**:910-913.
33. Liu Y, Liu N, Zhao H: **Inferring protein-protein interactions through high-throughput interaction data from diverse organisms.** *Bioinformatics* 2005, **21**:3279-3285.
34. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31**:248-250.
35. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21**:2076-2082.
36. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res* 2007, **35**:D572-574.
37. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34**:D354-357.
38. Yu JX, Sieuwerts AM, Zhang Y, Martens JW, Smid M, Klijn JG, Wang Y, Foekens JA: **Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer.** *BMC Cancer* 2007, **7**:182.
39. 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**:177-183.
40. Gadina M, Hilton D, Johnston JA, Morinobu A, Lighvani A, Zhou YJ, Visconti R, O'Shea JJ: **Signaling by type I and II cytokine receptors: ten years after.** *Curr Opin Immunol* 2001, **13**:363-373.
41. Ruggero D, Pandolfi PP: **Does the ribosome translate cancer?** *Nat Rev Cancer* 2003, **3**:179-192.
42. Rosen LS, Ashurst HL, Chap L: **Targeting signal transduction pathways in metastatic breast cancer: a comprehensive review.** *Oncologist* 2010, **15**:216-235.
43. Rahmouni S, Cerignoli F, Alonso A, Tsutji T, Henkens R, Zhu C, Louis-dit-Sully C, Moutschen M, Jiang W, Mustelin T: **Loss of the VHR dual-specific phosphatase causes cell-cycle arrest and senescence.** *Nat Cell Biol* 2006, **8**:524-531.
44. Sigoillot FD, Kotsis DH, Serre V, Sigoillot SM, Evans DR, Guy HI: **Nuclear localization and mitogen-activated protein kinase phosphorylation of the multifunctional protein CAD.** *J Biol Chem* 2005, **280**:25611-25620.
45. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14**:1085-1094.
46. Yan X, Mehan MR, Huang Y, Waterman MS, Yu PS, Zhou XJ: **A graph-based approach to systematically reconstruct human transcriptional regulatory modules.** *Bioinformatics* 2007, **23**:i577-586.
47. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, *et al*: **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.** *Breast Cancer Res* 2005, **7**:R953-964.
48. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.** *Proc Natl Acad Sci USA* 2005, **102**:13550-13555.
49. Li J, Lenferink A, Deng Y, Collins C, Cui Q, Purisima E, O ' Connor-McCourt M, Wang E: **Identification of high-quality cancer prognostic markers and metastasis network modules.** *Nat Commun* 2010, **1**:34.
50. Gaire RK, Bailey J, Bearfoot J, Campbell IG, Stuckey PJ, Haviv I: **MIRAGAA–a methodology for finding coordinated effects of microRNA expression changes and genome aberrations in cancer.** *Bioinformatics* 2010, **26**:161-167.
51. Lu LJ, Sboner A, Huang YJ, Lu HX, Gianoulis TA, Yip KY, Kim PM, Montelione GT, Gerstein MB: **Comparing classical pathways and modern networks: towards the development of an edge ontology.** *Trends Biochem Sci* 2007, **32**:320-331.
52. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
53. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22**:1600-1607.
54. Grossmann S, Bauer S, Robinson PN, Vingron M: **Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.** *Bioinformatics* 2007, **23**:3024-3031.
55. Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, *et al*: **Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network.** *Bioinformatics* 2007, **23**:2121-2128.
56. Ransohoff DF: **Lessons from controversy: ovarian cancer screening and serum proteomics.** *J Natl Cancer Inst* 2005, **97**:315-319.
57. David IB, Douglas BK: **statistical strategies for avoiding false discoveries in metabolomics and related experiments.** *Metabolomics* 2006, **2**: s11306-11006-10037.
58. Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, Zhu J, Wang H, Wang C, Topol EJ, *et al*: **Towards precise classification of cancers based on robust gene functional expression profiles.** *BMC Bioinformatics* 2005, **6**:58.
59. Massague J: **Sorting out breast-cancer gene signatures.** *N Engl J Med* 2007, **356**:294-297.
60. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, *et al*: **IntAct–open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35**: D561-565.
61. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, *et al*: **Human protein reference database as a discovery resource for proteomics.** *Nucleic Acids Res* 2004, **32**:D497-501.
62. Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KF, Munsterkotter M, Ruepp A, Spannagl M, Stumpflen V, Rattei T: **MIPS: analysis and annotation of genome information in 2007.** *Nucleic Acids Res* 2008, **36**:D196-201.
63. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**: D449-451.
64. Limviphuvadh V, Tanaka S, Goto S, Ueda K, Kanehisa M: **The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs).** *Bioinformatics* 2007, **23**:2129-2138.
65. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, *et al*: **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Res* 2005, **33**: D428-432.
66. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, *et al*: **A human phenome-interactome network of protein complexes implicated in genetic disorders.** *Nat Biotechnol* 2007, **25**:309-316.
67. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci USA* 2003, **100**:4372-4376.
68. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, **Series B (Methodological)57(1)**:289-300.

doi:10.1186/1752-0509-4-151
Cite this article as: Yao *et al.*: Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis. *BMC Systems Biology* 2010 **4**:151.