Gene xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Extracting a few functionally reproducible biomarkers to build robust subnetwork-based classifiers for the diagnosis of cancer

Lin Zhang [a], Shan Li [a], Chunxiang Hao [a], Guini Hong [a], Jinfeng Zou [a], Yuannv Zhang [b], Pengfei Li [b], Zheng Guo [a,b,*]

[a] Bioinformatics Centre, Key Laboratory for NeuroInformation of Ministry of Education and School of Life Science and Technology, School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

[b] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

ARTICLE INFO

Article history:
Accepted 10 May 2013
Available online xxxx

Keywords:
Cancer
Gene expression profiling
Reproducibility of findings
Protein interaction networks
Diagnosis

ABSTRACT

In microarray-based case–control studies of a disease, people often attempt to identify a few diagnostic or prognostic markers amongst the most significant differentially expressed (DE) genes. However, the reproducibility of DE genes identified in different studies for a disease is typically very low. To tackle the problem, we could evaluate the reproducibility of DE genes across studies and define robust markers for disease diagnosis using disease-associated protein–protein interaction (PPI) subnetwork. Using datasets for four cancer types, we found that the most significant DE genes in cancer exhibit consistent up- or down-regulation in different datasets. For each cancer type, the 5 (or 10) most significant DE genes separately extracted from different datasets tend to be significantly coexpressed and closely connected in the PPI subnetwork, thereby indicating that they are highly reproducible at the PPI level. Consequently, we were able to build robust subnetwork-based classifiers for cancer diagnosis.

1. Introduction

Numerous microarray studies have been performed to identify genes that are differentially expressed (DE) between cancer samples and normal controls with the objective of discovering diagnostic or prognostic biomarkers (Berchuck et al., 2009; Finak et al., 2008). However, DE genes extracted from different datasets for a particular cancer are often very inconsistent (Ein-Dor et al., 2005) mainly due to insufficient statistical power of detecting DE genes in small datasets (Zhang et al., 2008). It is known that thousands of samples are required in microarray studies to find reproducible biomarkers for cancer (Ein-Dor et al., 2006). Thus, new approaches to evaluating the reproducibility of biomarkers extracted from high-throughput biological data are needed (Qiu et al., 2006; Ransohoff, 2005). Considering that diverse molecular changes in cancers are functionally correlated (Klebanov et al., 2006; Subramanian et al., 2005), we have proposed the use of functional relationships between disease biomarkers for evaluating reproducibility (Gong et al., 2010; Gong et al., 2011; Yao et al., 2010; Zhang et al., 2009). For example, non-overlapping DE genes identified in different datasets for a specific type of cancer tend to be highly consistent when considering their coexpression relationship (Zhang et al., 2009).

Using scores based on certain reasonable biological assumptions (or molecular models), we can specify the reproducibility of DE gene discovery at different functional levels. Importantly, the biological assumption underlying a functional consistency score is statistically testable: if the score is significantly higher than expected by chance, then the assumption can explain a large fraction of diverse disease biomarkers. Based on this general framework, we determined the specific functional relationships between disease biomarkers on the protein–protein interaction (PPI) network level. Yao et al. (Yao et al., 2010) showed that non-overlapping PPI network signatures for breast metastasis identified from different studies may actually regulate the same sets of interacting protein neighbours. Moreover, Gong et al. found that cancer genes extracted from different databases tend to share significantly more PPI links (Gong et al., 2010). Given that genes encoding interacting proteins tend to share similar functions (Sharan et al., 2007) and DE genes for a disease are often connected in an active PPI subnetwork in response to a disease condition (Guo et al., 2007;

Abbreviations: DE, differentially expressed; PPI, protein–protein interaction; SAM, significance analysis of microarray; FDR, false discovery rate; POD, percentage of overlapping deregulations; PO, percentage of overlap; PON, percentage of overlap in the PPI network; SVM, support vector machine; RFE, recursive feature elimination.

* Corresponding author at: College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China. Tel.: +86 451 8661 5922, +86 28 8320 7187.

E-mail addresses: linzhang.bioinformatics@gmail.com (L. Zhang), shineyong27@yahoo.cn (S. Li), giselle118@yahoo.cn (C. Hao), hongguini0356052@yahoo.com.cn (G. Hong), zou.jinfeng@yahoo.com (J. Zou), zhangyuannv@gmail.com (Y. Zhang), lpfangle@126.com (P. Li), guoz@ems.hrbmu.edu.cn (Z. Guo).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.05.011

Ideker et al., 2002), it is reasonable to combine gene expression data with PPI data to evaluate the 'activated' functional relevance of DE gene lists extracted from different studies (Dittrich et al., 2008; Ulitsky and Shamir, 2009).

It is common practice to select a few of the most significant DE genes from thousands of genes as diagnostic or prognostic markers. However, most markers that are selected from a cohort of samples using this simple method, as well as using other complicated feature selection algorithms, often fail to work in other independent studies. In an attempt to address this problem, researchers have proposed building classifiers at a "meta-gene" level (Huang et al., 2003; Tamayo et al., 2007) using module-based (Mi et al., 2010), pathway-based (Lee et al., 2008) or PPI subnetwork-based approaches (Auffray, 2007; Chuang et al., 2007; Dao et al., 2011) rather than at the level of the individual gene. More specifically, several studies have suggested using a combination of gene expression and PPI data to identify "active PPI subnetworks" as relatively reproducible diagnostic or prognostic biomarkers for a disease (Chuang et al., 2007; Dao et al., 2011; Su et al., 2010). However, the performance of a classifier based on "meta-genes", such as subnetworks extracted from a dataset for a particular disease, still tends to decline in other independent datasets for the same disease (Chuang et al., 2007; Su et al., 2010). This trend could be because certain subnetworks extracted from one cohort of samples using a heuristic optimisation method may consist of genes with less prominent changes in gene expression in other cancer samples.

In this paper, we evaluated the reproducibility of the 5 (or 10) most significant DE genes extracted from one study in other independent studies for a particular cancer type. First, for each of four cancer types, we evaluated the consistency of the deregulation directions of DE genes (i.e., up- or down-regulated in cancer samples relative to normal controls) extracted from different datasets. Then, we proposed a scoring system to evaluate the reproducibility of two lists of the n most significant DE genes in terms of their significant coexpression and close connection in the human PPI network. Our results supported the assumption that the n most significant DE genes that are separately identified in different datasets for a particular type of cancer tend to be significantly coexpressed and closely connected in an active PPI subnetwork associated with the cancer. Finally, for each of the four cancer types, we built a classifier using the active PPI subnetworks based on the n most significant DE genes extracted from one dataset and evaluated its robustness in another independent dataset.

2. Results

2.1. Reliability of DE gene detection

First, we evaluated the consistency of the deregulation directions, representing increased or decreased average expressions of cancer samples compared to normal samples, between the DE genes that were separately identified in two datasets for each cancer type (see Table 1). For each dataset, we selected DE genes using SAM with a 1% FDR level. For colon cancer, 1149 DE genes were found in the Colon23 dataset, and 1127 (98%) of these genes were included in the set of 5478 DE genes identified in the Colon64 dataset, which were significantly more than expected by chance ($P = 2.46 \times 10^{-12}$, hypergeometric test). All of the DE genes shared between these two datasets were found to be deregulated in the same directions in the two datasets, which was unlikely to occur by chance ($P < 1 \times 10^{-12}$, binomial test). Similarly, for each of the other three cancer types, almost all DE genes shared between the two datasets were found to be deregulated in the same directions in the two datasets, as indicated by the POD_1 score shown in Table 2.

Many of the genes that were selected as DE genes in a dataset but not in another dataset may actually be differentially expressed in the latter cases. For example, 97% of the 4351 DE genes that were solely identified in the Colon64 dataset showed consistent deregulation directions in the Colon23 dataset, which was unlikely to occur by chance ($P < 1 \times 10^{-12}$, binomial test), indicating that the differential expression signals of most of these DE genes were actually represented in the Colon23 dataset. Similarly, for each of the other three cancer types, we also observed that nearly 90% of the DE genes solely identified in the dataset with greater statistical power showed the same deregulation directions in the dataset with the smaller power, as indicated by the POD_2 score shown in Table 2.

Taken together, the above results suggested that effective differential expression signals also widely exist in the smaller dataset for each of these cancer types. The high consistency of deregulation directions between the lists of DE genes determined from independent datasets for a particular cancer also validated the reliability of the majority of the DE genes that were identified in different studies for each type of cancer.

Table 1
Eleven datasets analysed in this study.

Cancer	Datasets[a]	T[b]	N[c]	GEO ACC No.	Platforms
Colon	Colon23	15	8	GSE4183	HG-U133_Plus_2
	Colon64	32	32	GSE8671	HG-U133_Plus_2
Gastric	Gastric24	12	12	GSE19826	HG-U133_Plus_2
	Gastric62	31	31	GSE13911	HG-U133_Plus_2
Breast	Breast58	31	27	GSE10810	HG-U133_Plus_2
	Breast185	42	143	GSE10780	HG-U133_Plus_2
Lung	Lung52	26	26	GSE7670	HG-U133A
	Lung107	58	49	GSE10072	HG-U133A
	Lung88[d]	44	44	GSE18842	HG-U133_Plus_2
	Lung120[d]	60	60	GSE19804	HG-U133_Plus_2
	Lung156[d]	91	65	GSE19188	HG-U133_Plus_2

[a] Each dataset is denoted by the following nomenclature: cancer type followed by the total number of samples.
[b] T denotes the number of tumour samples.
[c] N denotes the number of normal samples.
[d] These three datasets were used to further evaluate the stability of the classifiers trained for this cancer.

2.2. Reproducibility of top-ranked most significant DE genes at the PPI level

For each cancer type, most of the top n1 (n1 = 5, 10) DE genes extracted from one study were not among the top n2 (n2 = 5, 10) DE genes extracted from another study, as indicated by the low PO_{n1-n2} scores shown in Table 3. However, all of the top 10 DE genes identified in one dataset showed the same deregulation directions in another dataset, thereby indicating that these most significant DE genes are likely to show differential expressions in other independent cohorts of samples for the same cancer type.

We assumed that two DE genes were functionally related if they were significantly coexpressed and connected within two steps of PPI links in the PPI network (see details in Materials and methods,

Table 2
POD scores for two lists of DE genes for each cancer type.

Dataset	DEG_1[a]	DEG_2[b]	POD_1[c]	POD_2[d]
Colon23 vs. Colon64	1149	5478	100%	97%
Gastric24 vs. Gastric62	100	5335	100%	95%
Breast58 vs. Breast185	4919	6742	99%	89%
Lung52 vs. Lung107	2967	4928	99%	94%

[a] DEG_1 (or [b] DEG_2) denotes DE genes selected from the former (or the latter) dataset.
[c] POD_1 (or [d] POD_2) denotes the proportions of genes that showed the same deregulation directions in both datasets among the DE genes shared between the two lists (or among the DE genes that solely appeared in DEG_2). All P-values for the POD_1 and POD_2 scores in Table 2 are less than 1×10^{-12}.