Computational Biology and Chemistry 35 (2011) 151–158

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research article

Identifying novel prostate cancer associated pathways based on integrative microarray data analysis

Ying Wang[a,b,c], Jiajia Chen[a,d], Qinghui Li[a], Haiyun Wang[c], Ganqiang Liu[a,c,e], Qing Jing[c,f], Bairong Shen[a,*]

[a] Center for Systems Biology, Soochow University, No. 1. Shizi Street, Suzhou 215006, China
[b] Laboratory of Gene and Viral Therapy, Eastern Hepatobiliary Surgical Hospital, the Second Military Medical University, Shanghai, China
[c] School of Life Science and Technology, Tongji University, Shanghai 200092, China
[d] School of Chemistry and Biological Engineering, Suzhou University of Science and Technology, 215009, China
[e] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia
[f] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

ARTICLE INFO

Article history:
Received 31 December 2010
Received in revised form 14 April 2011
Accepted 16 April 2011

Keywords:
Meta-analysis
Pathway enrichment analysis
GeneGo database
KEGG database
Gene set enrichment analysis

ABSTRACT

The development and diverse application of microarray and next generation sequencing technologies has made the meta-analysis widely used in expression data analysis. Although it is commonly accepted that pathway, network and systemic level approaches are more reproducible than reductionism analyses, the meta-analysis of prostate cancer associated molecular signatures at the pathway level remains unexplored. In this article, we performed a meta-analysis of 10 prostate cancer microarray expression datasets to identify the common signatures at both the gene and pathway levels. As the enrichment analysis result of GeneGo's database and KEGG database, 97.8% and 66.7% of the signatures show higher similarity at pathway level than that at gene level, respectively. Analysis by using gene set enrichment analysis (GSEA) method also supported the hypothesis. Further analysis of PubMed citations verified that 207 out of 490 (42%) pathways from GeneGo and 48 out of 74 (65%) pathways from KEGG were related to prostate cancer. An overlap of 15 enriched pathways was observed in at least eight datasets. Eight of these pathways were first described as being associated with prostate cancer. In particular, endothelin-1/EDNRA transactivation of the EGFR pathway was found to be overlapped in nine datasets. The putative novel prostate cancer related pathways identified in this paper were indirectly supported by PubMed citations and would provide essential information for further development of network biomarkers and individualized therapy strategy for prostate cancer.

1. Introduction

Microarray technology has been widely used in biological studies for detecting the simultaneous expression of thousands of genes. Fig. 1 shows the number of articles in PubMed which identify differential gene expression using microarray technology or other methods. This number of papers has been steadily increasing over recent years. Even so, different laboratories frequently describe differentially expressed gene lists that vary significantly. These differences can be attributable to the variation in microarray platforms, experimental samples, normalization and analysis methods, and inherent biological uncertainty. Thus, it is difficult to obtain a reliable analysis result from only one dataset (Rhodes et al., 2004; Cahan et al., 2007; Xu et al., 2007; Tang et al., 2010; Yan et al., 2010). The availability of an increasing number of published microarray expression datasets means that application of the meta-analysis method is both possible and necessary to determine significant patterns from multiple datasets.

It is well known that cancer is a systems biology disease (Khalil and Hill, 2005; Hornberg et al., 2006; Faratian et al., 2009). The meta-analysis of cancer microarray expression data at a systems level, such as the pathway level, network level or even a system network dynamics level will improve the understanding of the complex molecular mechanisms underlying cancer. Previous analyses, however, have not, in general, compared the similarities between gene based analysis and pathway based analysis. To date, most meta-analysis studies performed on prostate cancer still focus on the detection of common signatures at the gene level rather than at the pathway level. Some groups have used a clustering algorithm to discover subtypes of the tumor (Dhanasekaran et al., 2001; Luo et al., 2001; Magee et al., 2001; Welsh et al., 2001; Lapointe et al., 2004), and some of the microarray data meta-analysis studies have emphasized the development of novel meta-analysis models to identify common gene signatures (Rhodes et al., 2002; Ghosh

* Corresponding author. Tel.: +86 512 65110951; fax: +86 512 65110951.
E-mail address: bairong.shen@suda.edu.cn (B. Shen).

1476-9271/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2011.04.003

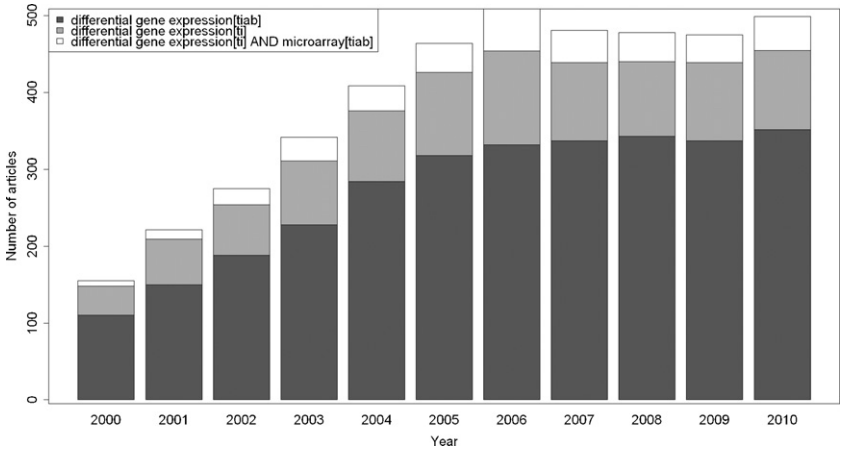

Fig. 1. The differential expression analysis related work found in PubMed database. PubMed queries for "differential gene expression [tiab]", "differential gene expression [ti]" and "differential gene expression [ti] and microarray [tiab]".

and Chinnaiyan, 2009). Several studies (Varambally et al., 2005; Nanni et al., 2006; Ghosh and Chinnaiyan, 2009) have mapped sets of significant genes present in at least two datasets to reveal related biological pathways. In addition, Gorlov et al. (2009) recently compared the consistency of the gene functional annotation analysis data between genome-wide association studies (GWASs) and microarrays, and proposed that the gene function based analysis might be more reproducible than the gene based analysis. Based on these reports, we integrated and analyzed the microarray data from normal prostate and tumor prostate samples at the pathway level. Firstly we verified that the expression signatures of different prostate cancer microarray datasets are more similar at the pathway level than at gene level. We then identified novel prostate cancer associated pathways.

In this study, primary differential gene expression analysis was performed using the Cancer Outlier Profile Analysis (COPA) package (MacDonald and Ghosh, 2006) in the R programming environment. We then used GeneGo's MetaCore (GeneGo, Inc.), a commercial integrated knowledge database, and KEGG, an open access pathway database, for pathway enrichment analysis (Liu et al., 2010). We also applied the GSEA method for gene set enrichment analysis to further prove our hypothesis. Text-mining searches in the Entrez PubMed database were performed for the candidate pathways to identify novel prostate-associated pathways.

2. Materials and methods

2.1. Dataset

We used 10 publicly available prostate cancer microarray expression datasets, which had been generated by nine independent laboratories. Five were measured with cDNA spotted technologies (Dhanasekaran et al., 2001; Luo et al., 2001; Lapointe et al., 2004; Tomlins et al., 2007) and five with Affymetrix arrays (Magee et al., 2001; Welsh et al., 2001; Singh et al., 2002; Varambally et al., 2005; Nanni et al., 2006). Meta-analysis requires a certain level of homogeneity in order to compare identified genes or pathways. According to Rhodes et al. (2004), comparative cancer analyses included cancer versus respective normal tissue, high grade versus low grade cancer, poor outcome versus good outcome, metastatic versus primary cancer, and subtype 1 versus subtype 2. Thus, our analysis across multiple datasets, based on normal prostate versus tumor prostate samples, was comparable. The individual analysis of each dataset consisted of three major steps: preprocessing, differential expression analysis, and pathway enrichment analysis.

2.2. Data preprocessing

During the preprocessing procedure, the datasets of two platforms were normalized separately using the Locally Weighted Scatter Plot Smoothing (LOWESS) method for within-slide normalization of the cDNA array datasets and the Median Absolute Deviation (MAD) method for between-slide normalization of all datasets. All expression values were transformed to base-two log. Low-qualified genes were filtered and missing data were imputed by the k-nearest neighbors ($k = 5$) imputation approach. We wrote the R scripts to run the preprocessing procedures of all datasets.

2.3. Differential expression analysis

In the differential expression analysis, we applied Cancer Outlier Profile Analysis (COPA) as originally proposed by Tomlins et al. (2005) and implemented in an R package by MacDonald and Ghosh (MacDonald and Ghosh, 2006) to identify significant genes between two sample classes. According to the COPA package procedure, we centered and scaled the data on a row-wise basis using the median and median average difference. The columns of microarray expression data matrix were samples and the rows were genes. Percentile was used to pre-filter the data. All genes with outlier samples that numbered less than that of the 95th percentile gene were removed from further consideration. A threshold cut-off for 'outlier' status was set and applied to all genes. The number of normal samples that could be considered 'outliers' was 0, which meant that no normal samples could be outliers.

2.4. Pathway and gene set enrichment analysis

After COPA analysis, we identified sets of significantly differentially expressed genes (outliers), which were mapped to the GeneGo database by MetaCore™ and the KEGG database using the Pathway-Express tool developed by the Intelligent Systems and Bioinformatics Laboratory (ISBL) for pathway enrichment analysis. In the Pathway-Express tool, we chose hypergeometric distribution to calculate the significance values (p values) and the FDR method to correct the p value. In MetaCore™, according to the MetaCore Manual (GeneGo, Inc.), the statistical significance (p-value) was also calculated using hypergeometric distribution. We defined the number of intersecting objects in the experiment as r, the number of network objects in the experiment as n, the total number of intersecting network objects in the database as R, and the total number of network objects in the database as N. A p-value was calculated for each object in the experiment based on its number of intersections.