Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Secondary use of existing public microarray data to predict outcome for hepatocellular carcinoma

Curtis J. Wray, MD,[a,*] *Tien C. Ko, MD,*[a] *and Filemon K. Tan, MD, PhD*[b]

[a] *Department of Surgery, University of Texas Medical School at Houston, Houston, Texas*
[b] *Division of Rheumatology, Department of Medicine, University of Texas Medical School, Houston, Texas*

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form
5 December 2013
Accepted 13 December 2013
Available online 6 January 2014

Keywords:
Hepatocellular carcinoma
Gene regulation
Chronic hepatitis
Cancer genetics

ABSTRACT

Background: Since 1990, numerous public repositories of microarray data have been created to store vast genomic data sets. Our hypothesis is that a secondary analysis of an available hepatocellular carcinoma (HCC) public data set could generate new findings and additional hypotheses.

Methods: The Gene Expression Omnibus at the National Center for Biotechnology Information was queried for available data sets specific for 'HCC' and 'clinical data.' Genes that passed filtering and normalization criteria were analyzed using the class comparison and prediction functions in BRB-ArrayTools. Ingenuity pathway analysis software was used to identify potential gene networks up- or down-regulated.

Results: The file GDS274, which measured gene expression in primary HCC lesions with or without hepatic metastases from a cohort of Chinese patients, was identified as an appropriate data set and was imported into BRB-ArrayTools. 9984 genes passed filtering criteria. Clinical data demonstrated alpha fetoprotein (AFP) >100 ng/mL predictive of worse survival (HR 5.87, 95% confidence interval: 1.11–31.0). A class comparison between patients with an AFP >100 and those with AFP <100 demonstrated 92 genes to be differentially expressed. Ingenuity pathway analyses demonstrated the top networks associated with the observed gene expression.

Conclusions: Using available HCC microarray data, we identified genes differentially expressed based on AFP >100. Canonical pathway analysis demonstrated functional gene pathways and associated upstream regulators. This study maximizes the use of publicly available data by generating new findings. Secondary analyses of these data sets should be considered by investigators before embarking on new genomic experiments.

1. Introduction

High throughput genomic technologies are increasingly being used to identify therapeutic targets and risk factors for specific diseases in this era of personalized medicine [1–3]. Gene expression microarrays have been used to differentiate types of leukemia, B-cell lymphoma, breast cancer, and lung cancer [4–6]. The use of high throughput technologies has generated vast amounts of genomic data. Since 1990, numerous public repositories of microarray data have been created. At the present time, a prerequisite to the publication of microarray data is that the results must be publicly available to the research community [7]. The data should be in a form that permits conclusions to be evaluated independently [8].

* *Corresponding author*. Department of Surgery, University of Texas Medical School at Houston, Houston, TX. Tel.: +1 713 566 5095; fax: +1 713 566 4583.
E-mail address: curtis.j.wray@uth.tmc.edu (C.J. Wray).
0022-4804/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jss.2013.12.013

Authors describing a newly sequenced genome, gene, or protein must deposit the primary data in a permanent, public data repository, such as the DNA Data Bank of Japan, European Bioinformatics Institute, and the National Center for Biotechnology Information (NCBI) or ArrayExpress [9].

The established databases allow researchers, at their discretion, to submit some or all of the clinical data associated with a microarray experiment. The standardization of data formatting facilitates further data analyses. This common format makes it easier for researchers to access, query, and share data [10]. Our research aims were [1] to search for a publicly available hepatocellular carcinoma (HCC) gene expression data set that also included clinical patient data and [2] to generate hypotheses using this data set.

2. Methods

2.1. Online search

The online Gene Expression Omnibus, a public functional genomics data repository at the NCBI (http://www.ncbi.nlm.nih.gov/gds) was queried for available data sets. The specific search included ("carcinoma, hepatocellular" [MeSH Terms] OR HCC [All Fields]) AND ("patients" [MeSH Terms] OR patient [All Fields]) AND ("mortality" [Subheading] OR "survival" [MeSH Terms] OR survival [All Fields]). The genomic data (GDS274) file met the search criteria and was imported into BRB-ArrayTools version 4.2 (National Cancer Institute), available at http://linus.nci.nih.gov/BRB-ArrayTools.html [11]. This data set represents primary lesions with or without hepatic metastases in patients with hepatitis B-induced HCC.

2.2. Clinical data

Deidentified available patient data were analyzed for overall survival using Kaplan–Meier survival analysis. Data included age, primary tumor size, type of surgical resection, portal vein involvement, presence of multiple tumors, cirrhosis, serum alpha fetoprotein (AFP), vital status, and survival time. A Cox proportional hazards model was used to determine the effects of multiple independent predictor variables on overall survival. The final multivariate model was created using the backward, stepwise method of covariate elimination to consider a wide range of possible best models [12]. Covariates that were significant below a P value <0.20 were included in the final multivariate model analysis [13]. STATA 12 (StataCorp, College Station, TX) statistical software was used for all analyses.

2.3. Microarray class comparison

Using BRB-ArrayTools, genes that passed filtering and normalization criteria were analyzed using the class comparison, which compares gene expression among predefined classes and presumes the data consists of experiments of different samples representative of the classes. We identified genes that were differentially expressed among classes using a multivariate permutation test [14–17]. The test statistics used were random variance t-statistics for each gene. Although t-statistics were used, the multivariate permutation test is nonparametric and does not require the assumption of Gaussian distributions. In the class comparison analysis, technical replicates of the same sample were averaged.

2.4. Canonical pathway analysis

Interactive pathway analysis (IPA) of complex genomics data software (Ingenuity Systems, www.ingenuity.com, Redwood City, CA) was used to examine differentially expressed genes [18,19]. The analysis settings reference set was the Ingenuity Knowledge Bases (genes + endogenous chemicals). IPA was used to assess for network-associated functions and well-characterized molecular signaling (canonical) pathways. This computational approach investigates the network behavior as a system. The Ingenuity software scans the list of input genes to identify networks (i.e., relationships between genes) using data in the Ingenuity Pathways Knowledge Base, a manually curated database of functional interactions extracted from peer-reviewed publications [20]. A Fisher exact test is performed to determine the likelihood of obtaining at least the equivalent numbers of genes by chance (i.e., from a random input gene set) as actually overlap between the input gene set and the genes present in each identified network. IPA predicts which upstream regulators are activated or inhibited, based on known relationships, to explain the up- and down-regulated genes. The IPA software describes an "upstream regulator" as any molecule that can affect the expression of another molecule.

3. Results

3.1. Online search

The data set GDS274 "HCC metastasis" was identified at Gene Expression Omnibus and imported into BRB-ArrayTools. The data from this microarray experiment were obtained from hepatitis B virus (HBV) positive HCC patients ($n = 40$) in China. GDS274 included primary HCC tumors and matched intrahepatic metastases (i.e., a primary tumor and an intrahepatic metastasis from the same patient). As originally published by Ye *et al.* [11], the mean patient age was 50 y (range: 36–74). The median diameter of the primary HCC was 7.2 cm (range 1.3–17.5). Thirty-two cases (80%) had underlying cirrhosis and 98% of the patients were HBV-positive. Serum AFP was >20 ng/mL in 68% of patients.

3.2. Clinical data

Deidentified, individual patient data were included in the GDS274 data set. A Cox proportional hazards model was created to determine predictors of survival. Age, tumor size, portal vein involvement, stage, and AFP >100 were found to have a P value <0.20 on univariate analysis. In the final multivariate analysis, only AFP >100 (HR 5.87) was predictive of worse survival (Table 1).