Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Original Article

Differentially expressed genes in tumors of prostate cancer in American patients with European and African origin

K. Venkatesh [a,*], D. Lajwanti [b], Sandhya.P. Kiran [c], D.V. Raje [d], Anuraj Nayarisseri [e]

[a] Statistician, Ocimum Biosolutions Ltd., Road #1, Banjara Hills, Hyderabad 560034, India
[b] Bioinformatician, Ocimum Biosolutions Ltd., Road #1, Banjara Hills, Hyderabad 560034, India
[c] Statistician, Ocimum Biosolutions Ltd., Road #1, Banjara Hills, Hyderabad 560034, India
[d] Statistician, MDS Bio-Analytics Pvt. Ltd., 127, Sakar Enclave, Shankar Nagar, Nagpur 440010, India
[e] Bioinformatics Scientist, In silico, Research Laboratory, Eminent Biosciences, Indore 452010, India

ARTICLE INFO

Article history:
Received 5 March 2013
Accepted 17 April 2013
Available online 21 May 2013

Keywords:
Microrrays
Prostate tumors
Univariate simulation
Multidimensional scaling

ABSTRACT

Aim: Prostate cancer is one of the most common types of cancer observed in older men globally. In United States, the incidence rate is at its increase over the years and it has been noticed through several studies that the risk of disease differs within population and is more associated with the racial origin. The present study aims at determining differentially expressed (DE) genes using microarray in tumor samples of American patients suffering from prostate cancer who are either from African or European origin.

Methods: The study included microarray data on 89 tumors of prostate cancer. Of these, 34 were African–American, 35 were European–American and 20 were benign samples. After preprocessing, moderated t-statistic was used to determine genes that are differentially expressed across the groups. Gene ontology analysis was performed to understand the functional relevance of differentially expressed genes.

Results: The analyses resulted into 82 significant genes, of which 24 genes were present in 60 of the 100 simulated datasets with a corresponding threshold δ_0 value of 0.00003. These genes discriminated two study groups with an average silhouette width of 0.3211. Gene ADI1 had higher expression in European–American samples, while CTNNB1, PSPH and CRYBB2 had higher expression in African–American samples.

Conclusion: The GO analysis considering significant genes revealed that terms like immune response, antigen processing showed relatively higher abundance of DE genes. The potential significance of CTNNB1 and ADI1 genes in prostate cancer biology was re-established. However, their expression patterns are different in two study groups, which is an interesting observation from the study.

* *Corresponding author*. Tel.: +91 9686676661.
E-mail address: venkeyk@gmail.com (K. Venkatesh).

0974-6943/$ – see front matter
http://dx.doi.org/10.1016/j.jopr.2013.04.036

1. Introduction

The prostate cancer is one of the leading cause of cancer in men over 40 in United States, with 186,000 new cases in 2008 and 28,600 deaths.[1,2] It is more common cause of cancer in Europe and least common in South and East Asia. In the United Kingdom, the prostate cancer is second cause of cancer death after lung cancer, with around 35,000 cases diagnosed every year, of which, there are 10,000 cases of death. In literature, specific causes of prostate cancer were not mentioned but the possible factors could be: age, genetics, lifestyle, and other factors. The prostate cancer is uncommon in men in their 40s and becomes more common in their 70s. In United States, the African men are having high risk of developing prostate cancer than European men due to genetic factor,[3,4] though the mortality rate remains controversial.[5,6]

The primary objective of any microarray data is to obtain differentially expressed genes in different conditions. In the present study, microarray data was used for identifying differentially expressed genes that distinguish the tumor-groups of African–American and European–American men and to obtain biological information based on differentially expressed genes. For this, a simple and meaningful approach of moderated *t*-statistic was used,[7] on both normalized dataset and simulated datasets that were generated based on univariate simulation at gene level, in order to detect the true significant genes that can separate African–American and European–American prostate tumors.

2. Methods

2.1. Gene expression data

The prostate cancer study contains 89 human samples, of which, 34 were African–American prostate tumor samples, 35 were European–American prostate tumor samples and 20 were cancer-free samples. The processed data, multi-array suite (MAS) expressions, were downloaded from ArrayExpress using Exp ID: E-GEOD-6956. All these samples were hybridized to Affymetrix GeneChip HG-U133A 2.0 arrays, with 22,283 probe sets.

2.2. Data normalization and statistical analysis

The intensity data requires an appropriate transformation and normalization. The data was log transformed and normalized with the median centering. The median absolute deviation scaling was also performed across samples in order to reduce the variation across samples. The moderated *t*-statistics was used on the normalized data to detect the differentially expressed genes between gene expressions profiles of 34 African–American and 35 European–American patients. In the present analysis, the *p*- value of moderated *t*-statistics was chosen to be $\delta_0 = (0.05 > 0.1 \times 10^{-5})$ and univariate simulated data was generated, nearly, 100 times. In each simulated data, the moderated *t*-statistics were obtained the significant genes at *p*-value threshold to detect the true significant genes. The univariate simulation procedure is given in detail in the following section.

2.3. Univariate simulation

The univariate normal distribution is determined by two parameters: mean and standard deviation. The probability density function of univariate distribution is given by:

$$f(x) = \frac{1}{S_{(i,1)}\sqrt{2\pi}} e^{-\left(x_{ij} - m_{(i,1)}\right)^2 / 2s^2_{(i,1)}} \quad (1)$$

$$f(y) = \frac{1}{S_{(i,2)}\sqrt{2\pi}} e^{-\left(y_{ik} - m_{(i,2)}\right)^2 / 2s^2_{(i,2)}} \quad (2)$$

Here, *x* represents the gene expression profiles of 34 African–American patients, x_{ij} is the expression value for the ith gene in the jth, ($j = 1, 2, 3 \ldots 34$) sample, *m* (i, 1) and *s* (i, 1) are the mean and standard deviation for the ith gene respectively, and similarly, *y* represents the gene expression profiles of 35 European–American patients, y_{ik} is the expression value for the ith gene in the kth, ($k = 1, 2, \ldots 35$) sample, *m* (i, 2) and *s* (i, 2) are the mean and standard deviation for this group for the ith Gene respectively. The above estimators were used for generating random realization of univariate data for respective conditions. The steps involved in the analysis are given as below:

Step1: Generate a simulated dataset using the estimated parameters from Equations (1) and (2) for all genes. Obtain moderated *t*-statistic values for the simulated dataset. Similarly, simulate 100 datasets and obtain moderated *t*-statistic values for the respective simulated dataset.
Step2: Obtain mean of *t*-statistic values across all simulated datasets.
Step3: Choose the threshold to be δ_0 and find the significant genes on normalized data. For the same threshold, find the significant genes on each simulated dataset and count the significant genes that are present in each possible datasets.[8]
Step4: With the significant genes at the selected threshold, perform multidimensional scaling (MDS) classification on normalized data to see whether group of samples are properly classified or not. If samples are poorly classified, select the significant genes that are present in more than 50 simulated datasets and perform MDS again and see if any improvement in the classification. If still classification is poor, choose another threshold and repeat step 4 to arrive at the set of significant genes that accurately classifies the group of samples.
Step5: If still classification is not much improved, remove the samples that are often misclassified in two groups and repeat the step 1 to step 4 in order to find the true significant genes.

3. Results and discussion

3.1. Differential expression analysis

Gene expression profiles of 89 *Homo sapiens* prostate samples were downloaded from a publicly available database, Array