Taiwanese Journal of Obstetrics & Gynecology 55 (2016) 552–557

Contents lists available at ScienceDirect

Taiwanese Journal of Obstetrics & Gynecology

journal homepage: www.tjog-online.com

Original Article

Gene set-based integrative analysis of ovarian clear cell carcinoma

Chia-Ming Chang [a, b, c], Shih-Hwa Chiou [b, d], Ming-Jie Yang [a, c], Ming-Shyen Yen [a, c], Peng-Hui Wang [a, c, e, *]

[a] Department of Obstetrics and Gynecology, Taipei Veterans General Hospital, Taipei, Taiwan
[b] Institute of Oral Biology, National Yang-Ming University, Taipei, Taiwan
[c] Department of Obstetrics and Gynecology, National Yang-Ming University, Taipei, Taiwan
[d] Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan
[e] Department of Medical Research, China Medical University Hospital, Taichung, Taiwan

ARTICLE INFO

Article history:
Accepted 20 May 2016

Keywords:
clear cell adenocarcinoma
functionome
gene expression
integrative analysis

ABSTRACT

Objective: The pathogenesis of ovarian clear cell carcinoma is still poorly understood; therefore, we conducted a gene set-based analysis by integrating datasets downloaded from publicly available microarray gene expression databases to investigate the pathogenesis of clear cell carcinoma, which was based on the regularity of functions defined by gene ontology or canonical pathway databases.
Materials and Methods: The gene expression profiles of 80 clear cell carcinomas and 136 normal ovarian controls were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus database. The gene expression profiles were converted to the gene set regularity (GSR) indexes computed using the modified differential rank conservation, an algorithm measuring the degree of gene expression ranking change in a gene set. Then the differences of GSR indexes between clear cell carcinomas and normal ovarian controls were analyzed.
Results: Machine learning can accurately recognize and classify the patterns of functional regularities containing the GSR indexes between the clear cell carcinomas and normal controls with an accuracy of 99.3%. The significant aberrations included oxidoreductase activity, binding, transport, channel activity, cell adhesion, immune response, chromosome assembly, and the deregulated signaling molecules, such as guanyl nucleotide exchange factors, phosphoinositide 3-kinase-activating kinase, receptor tyrosine kinase B, and protein tyrosine kinase.
Conclusion: Our pioneering works using the functionome, which was converted from microarray gene expression profiles for integrative analysis, showed a clear distinction of functional changes between the clear cell carcinomas and normal ovarian controls. This approach might provide a comprehensive view of the deregulated functions of clear cell carcinomas for further investigation.

Introduction

Ovarian carcinoma (OC), especially epithelial ovarian cancer, is one of the most lethal gynecological malignancies [1]. OC is a heterogeneous disease and consists of several molecularly and clinicopathologically distinct subtypes [2–4]. Clear cell carcinoma may be the most common secondary subtype, especially in Oriental countries such as Taiwan [5,6]. The prognosis of clear cell carcinoma is relatively poor, and the recurrence and 5-year survival rates were 27% and 60%, respectively [7]. The pathogenesis of clear cell carcinoma is unknown, but it is postulated to involve oxidative stress, genomic alterations, inflammatory processes, and estrogens [8,9].

Deoxyribonucleic acid (DNA) microarray gene expression is the primary tool to investigate the pathogenesis of various kinds of diseases, including clear cell carcinoma. The workflow of analyzing gene expression profiles usually consists of detecting the differentially expressed genes, and then mapping them to the gene ontology (GO) terms or signaling pathways for annotating the deregulated biological functions. However, this methodology focuses only on the statistically significant genes or pathways; the

* Corresponding author. Department of Obstetrics and Gynecology, Taipei Veterans General Hospital, 201, Section 2, Shih-Pai Road, Taipei 112, Taiwan.
E-mail addresses: phwang@vghtpe.gov.tw, phwang@ym.edu.tw (P.-H. Wang).

http://dx.doi.org/10.1016/j.tjog.2016.06.004

complete information about the regulation of the functions, for example, the functionome of clear cell carcinoma, is not provided.

Most of the microarray datasets investigating the pathogenesis of clear cell carcinoma were derived from a relatively small sample size. To overcome these limitations, we used a gene set based model to investigate the pathogenesis of clear cell carcinoma with the functionome. This model converts and quantizes the biological function defined by a gene set with the gene expression profiles downloaded from publicly available databases to a gene set regularity (GSR) index computed by the modified differential rank conservation (DIRAC) algorithm [10], which measured the matching degree of gene expression rankings in a given gene set between two different phenotypes, i.e., clear cell carcinomas and normal ovarian tissue controls in this study. This model utilized the gene set definitions from the GO term [11] and canonical pathways databases downloaded from the Molecular Signatures Database [12]. These two-gene set definitions collect relatively comprehensive biological functions, processes, or signaling pathways, so we used them to establish the human functionomes. The GO database contains 1454 gene sets, defining biological functions, processes, and cellular components. The canonical pathway database defines 1330 canonical signaling pathways, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome pathways. The pathogenesis of clear cell carcinoma was investigated with the functionomes using statistical methodologies.

Materials and methods

Microarray datasets, gene set definitions, and data processing

Gene expression microarray datasets were downloaded in the SOFT format from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. The clear cell carcinoma and normal ovarian tissue control datasets were selected only when the samples originated from the ovarian tissue. The common genes, derived by intersecting of the genes from all datasets as well as the corresponding gene expression data, were used in this study. Datasets were discarded if the number of the common genes available was less than 8000 during cross-platform integration of these datasets. Gene sets were discarded if the number of gene elements in the gene set is less than three.

Computing GSR indexes by modified DIRAC

The algorithm for computing GSR indexes was modified from DIRAC, and the details are displayed in Figure 1. The GSR index measures the change of gene expression ranking between two phenotypes in a gene set. For this purpose, the GSR indexes of both the clear cell carcinoma and the normal control groups were computed by comparing the sample's gene expression ranking with a standard template, i.e., the baseline gene set ranking template derived from the most common gene expression ranking in a gene set among the entire normal ovarian tissue control samples. Then the subsequent analyses were carried out based on this same standard with the clear cell carcinoma and normal control GSR indexes. The baseline gene set raking template for each gene set is established by pairwise comparison between the expression levels of two genes for all possible combinations of gene pairs. Establishment of the baseline gene set expression-ranking template and measurement of GSR indexes were executed in (R code) environment; the code and the test datasets are available on the GitHub (https://github.com/carlzang/GSR-model.git).

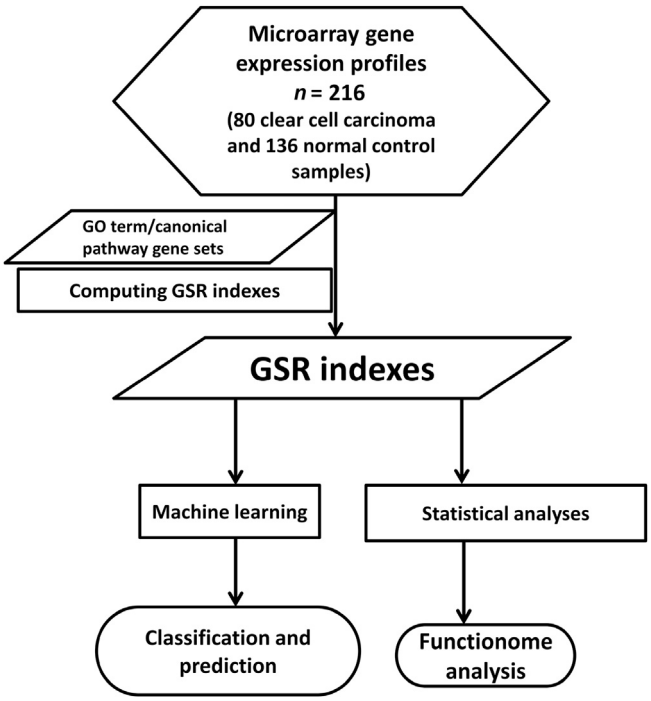

Figure 1. The gene set regularity (GSR) index was computed by converting the gene expression rankings of the clear cell carcinoma or normal ovarian control sample through each gene ontology (GO) term or canonical pathway gene set. Machine learning was trained to recognize the patterns consisting of GSR indexes, and then execute the binary (case-control) classifications. Functionome analysis was carried out by statistical methodologies to investigate the pathogenesis of clear cell carcinoma.

Statistical analysis

The differences between the clear cell carcinoma and control GSR indexes were tested using Mann–Whitney *U* test and corrected by multiple hypotheses using false discovery rate (Benjamini–Hochberg procedure). The significance level was set at $p < 0.001$.

Classification and prediction by machine learning

GSR index matrices computed through the GO term and canonical pathway gene sets were classified and predicted by "kernlab" [13], an R package for executing supporting vector machine (SVM) [14] with the setting of kernel = "rbfdot"(Radial Basis kernel "Gaussian"), type = "C-svc" (C classification). The performance of classification and prediction by SVM were measured by 5-fold cross-validation. The performance was assessed with the sensitivity, specificity, accuracy, and area under curve. The area under curve was computed using the R package "pROC" [15].

Establishment of GO tree

The tree of the deregulated GO terms was constructed and visualized by the RamiGO [16], an R package providing functions to interact with the AmiGO 2 web server (http://amigo2.berkeleybop.org/amigo) and retrieving GO trees.

Results

Sample information and means of the GSR indexes for clear cell carcinoma

DNA microarray gene expression datasets of the clear cell carcinoma samples were downloaded from the NCBI GEO database.