Pathology – Research and Practice 213 (2017) 394–399

Contents lists available at ScienceDirect

Pathology – Research and Practice

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

Original article

Predicting novel salivary biomarkers for the detection of pancreatic cancer using biological feature-based classification

Huan-Jun Liu[a,1], Yuan-Ying Guo[b,1], Du-Jun Li[c,*]

[a] Department of Hepatobiliary Surgery, Yantai Affiliated Hospital of Binzhou Medical University, Yantai, 264000, China
[b] Department of Preventive Medicine, School of Public Health, Jilin University, Jilin 130000, China
[c] Department of Clinical Laboratory, Yantai Yeda Hospital, No. 11 Taishan Road, Yantai, Shandong, 264000, China

ARTICLE INFO

Article history:
Received 9 December 2015

Keywords:
Pancreatic cancer
Salivary biomarkers
Early detection
Gene ontology

ABSTRACT

Aim: The use of saliva as a diagnostic fluid enables non-invasive sampling and thus is a prospective sample for disease tests. This study fully utilized the information from the salivary transcriptome to characterize pancreatic cancer related genes and predict novel salivary biomarkers.
Methods: We calculated the enrichment scores of gene ontology (GO) and pathways annotated in Kyoto Encyclopedia of Genes and Genomes database (KEGG) for pancreatic cancer-related genes. Annotation of GO and KEGG pathway characterize the molecular features of genes. We employed Random Forest classification and incremental feature selection to identify the optimal features among them and predicted novel pancreatic cancer-related genes.
Results: A total of 2175 gene ontology and 79 KEGG pathway terms were identified as the optimal features to identify pancreatic cancer-related genes. A total of 516 novel genes were predicted using these features. We discovered 29 novel biomarkers based on the expression of these 516 genes in saliva. Using our new biomarkers, we achieved a higher accuracy (92%) for the detection of pancreatic cancer. Another independent expression dataset confirmed that these novel biomarkers performed better than the previously described markers alone.
Conclusion: By analyzing the information of the salivary transcriptome, we predict pancreatic cancer-related genes and novel salivary gene markers for detection.

1. Introduction

Pancreatic cancer is one of the most lethal human cancers and is the fourth most common cause of cancer-related deaths in the United States. The American Cancer Society has estimated that there would be 48,960 new cases and 40,560 deaths from pancreatic cancer in 2015. Many patients experience abdominal or back pain that broadly localizes to the tumor area, followed by obstructive jaundice. Other signs can include asthenia, anorexia, and weight loss. Less common symptoms include venous thrombosis, panniculitis, liver disorders, gastric-outlet obstruction, increased abdominal girth, and depression [1,2].

Although the causes of pancreatic cancer are currently poorly understood, many studies have demonstrated that age, tobacco use and some genetic disorders are the main risk factors. Approximately 20% of pancreatic cancers are associated with tobacco use; indeed, the risk in smokers is 2 times greater than the risk for nonsmokers [3]. Individuals who have a family history of pancreatic cancer may have a nine-fold increased risk for the development of this tumor over the general population [4]. However, most pancreatic cancers represent sporadic cases. The pathogenesis of most pancreatic cancers is due to the accumulation of mutations in three types of genes: oncogenes, tumor-suppressor genes, and genomic maintenance genes [5]. Many early-stage pancreatic cancer patients have activating mutations in the *KRAS* gene. This gene might participate in the pathogenesis of pancreatic cancer via the autocrine epidermal growth-factor (EGF)-family signaling pathway [6]. Additionally, carriers of mutations in *p16* [7], *TP53* [8], or *DPC4* [9] are at high risk for the development of pancreatic cancer.

Patients with pancreatic cancer usually develop fatal metastases. Overall, the 5-year survival rate is less than 26%. Moreover, pancreatic cancer is resistant to most forms of current treatment. Less than 20% of patients are suitable for surgery due to dissemination to other tissues prior to detection [3]. Therefore, the early detection of the disease is urgently required for the prevention

* Corresponding author.
E-mail address: lidujun336@163.com (D.-J. Li).
[1] Co-first authors.

http://dx.doi.org/10.1016/j.prp.2016.09.017
0344-0338/

and management of pancreatic cancer. Traditional blood tests are generally nonspecific, and the results may be influenced by hyperglycemia, anemia or other disorders [2,10]. Novel strategies that use distinguishing salivary biomarkers for the detection of systemic cancers have emerged in recent decades. Saliva is an important biological fluid in the oral cavity. Because it is a filtration of blood, it can reflect bodily conditions. Moreover, the collection of saliva is convenient, cost-effective and noninvasive. Thus, saliva is an ideal sample for monitoring physiological statuses and predicting systemic cancers.

Different categories of biomarkers in saliva have been evaluated to detect pancreatic cancer. Eight salivary metabolites, including leucine, phenylalanine and aspartic acid, were previously identified as specific pancreatic cancer biomarkers [11]. Variations in salivary microbiota were observed between pancreatic cancer patients and healthy subjects, indicating that the salivary microbiota may serve as a biomarker for the disease [12]. One previous study identified four salivary transcriptomic biomarkers (*KRAS, MBD3L2, ACRV1* and *DPM1*) that could be used for the detection of pancreatic cancer [13]. However, these biomarkers still suffer from low sensitivity and limited utility. In this study, we employed an effective computational method to fully utilize the information from the salivary transcriptome. We characterized the pancreatic cancer-related genes through a biological feature-based classification method and predicted 29 salivary mRNA biomarkers. These novel identified mRNA biomarkers expand the choice for saliva tests of pancreatic cancer and may provide more accurate and specific early detection for the disease.

2. Materials and methods

2.1. Dataset

The analysis of salivary gene expression profiles employed microarrays and qPCR to examine differentially expressed genes in saliva samples from pancreatic cancer patients and healthy controls [13]. In the study, 114 saliva samples derived from the saliva bank of pancreatic disease at the University of California-Los Angeles (UCLA) Dental Research Institute were selected. While there was no significant difference in total RNA quantity between the pancreatic cancer and healthy controls, 49 up-regulated and 21 down-regulated transcripts were identified in the pancreatic cancer samples using a more stringent cutoff p-value (<0.01). Quantitative PCR (qPCR) validation verified 23 up-regulated and 12 down-regulated transcripts that were consistent with the microarray data. These 35 genes were regarded as positive samples (PC-related genes, Table 1) in this study, while $35 \times 60 = 2100$ background genes in the Ensemble database were randomly selected as the negative samples (non-PC-related genes, data not shown). Because the number of negative samples was much larger than the positive samples, the negative samples were split into 10 groups to release the imbalance. Then, the positive samples were mixed with each group of negative samples to construct 10 datasets for use in the subsequent procedures.

2.2. Encoding the genes using gene ontology and KEGG pathway enrichment scores

We encoded the positive and negative genes to disguise the positive genes (PC-related genes). Gene properties can be depicted using gene ontology and KEGG pathway annotation. If two genes have a similar function, they will have similar annotation in the gene ontology and KEGG pathway analyses. Therefore, we encoded each gene as a numeric vector consisting of an enrichment score for each of the GO and KEGG terms. The enrichment score was calculated as: where N is the total number of genes, M is the number of genes annotated to gene ontology item j in the gene-gene interaction network in humans, gene i has number n direct neighbors in the sub-network of gene i, and m is the number of genes in the gene i neighborhood network that are annotated to gene ontology or KEGG item j. The gene-gene interaction network was constructed based on the STRING protein network. Larger enrichment scores for a GO item or KEGG pathway indicate a greater overrepresentation of that item or pathway. In total, a vector having 6242 GO and 214 KEGG features was built for each gene.

Table 1
Genes that are regarded as positive samples.

Ensembl Gene ID	Gene name	Different expression
ENSG00000134940	ACRV1	up-regualted
ENSG00000081377	CDC14B	up-regualted
ENSG00000129691	ASH2L	up-regualted
ENSG00000109689	STIM2	up-regualted
ENSG00000020181	GPR124	up-regualted
ENSG00000239998	LILRA2	up-regualted
ENSG00000106991	ENG	up-regualted
ENSG00000112183	RBM24	up-regualted
ENSG00000154237	LRRK1	up-regualted
ENSG00000104093	DMXL2	up-regualted
ENSG00000196812	ZSCAN16	up-regualted
ENSG00000230522	MBD3L2	up-regualted
ENSG00000211445	GPX3	up-regualted
ENSG00000005961	ITGA2B	up-regualted
ENSG00000179242	CDH4	up-regualted
ENSG00000163993	S100P	up-regualted
ENSG00000122515	ZMIZ2	up-regualted
ENSG00000215301	DDX3X	up-regualted
ENSG00000171794	UTF1	up-regualted
ENSG00000133703	KRAS	up-regualted
ENSG00000198947	DMD	up-regualted
ENSG00000134508	CABLES1	up-regualted
ENSG00000230204	FTH1P5	up-regualted
ENSG00000133112	TPT1	down-regualted
ENSG00000277443	MARCKS	down-regualted
ENSG00000130066	SAT1	down-regualted
ENSG00000135241	PNPLA8	down-regualted
ENSG00000000419	DPM1	down-regualted
ENSG00000173762	CD7	down-regualted
ENSG00000140479	PCSK6	down-regualted
ENSG00000166548	TK2	down-regualted
ENSG00000167996	FTH1	down-regualted
ENSG00000167552	TUBA1A	down-regualted
ENSG00000105373	GLTSCR2	down-regualted
ENSG00000006837	CDKL3	down-regualted

2.3. Removing features that have weaker effects

To release the heavy computational load, we simplified the features of the 10 datasets by removing the low-related features. We transformed all feature enrichment scores to a standard scale using the following formula:in which and are the standard deviation and mean value, respectively, of the jth feature, and are the original value and transformed standardized value, respectively, of the ith gene of the jth feature. We calculated each feature vector's correlation coefficient with the classification vector (vector containing 1 and 0, where 1 indicates the positive sample and 0 indicates the negative sample). The corresponding feature was removed when the correlation coefficient was less than 0.1. Finally, the corresponding remaining features were used in the 10 datasets.

2.4. Ranking features based on the minimum redundancy maximum relevance algorithm

We employed the minimum redundancy maximum relevance (mRMR) algorithm to rank the features used for selecting the optimal features. The mRMR method, which was first proposed by Peng