Genomics 103 (2014) 48–55

Contents lists available at ScienceDirect

Genomics

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

Gene expression profile based classification models of psoriasis

Pi Guo [a,b,c,1], Youxi Luo [a,b,1], Guoqin Mai [a,b], Ming Zhang [f,g], Guoqing Wang [d,e,**], Miaomiao Zhao [a,b], Liming Gao [a,b], Fan Li [d,e], Fengfeng Zhou [a,b,*]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, PR China
[b] Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, PR China
[c] Department of Public Health, Shantou University Medical College, No. 22 Xinling Road, Shantou, Guangdong 515041, PR China
[d] Department of Pathogeny Biology, Norman Bethune Medical College, Jilin University, Changchun, Jilin 130021, PR China
[e] Key Laboratory of Zoonosis Research, Ministry of Education, Jilin University, Changchun, Jilin 130021, PR China
[f] Department of Epidemiology and Biostatistics, Faculty of Infectious Diseases, University of Georgia, Athens, GA 30605, USA
[g] Institute of Bioinformatics, University of Georgia, Athens, GA 30605, USA

ARTICLE INFO

Article history:
Received 25 March 2013
Accepted 1 November 2013
Available online 13 November 2013

Keywords:
Gene expression profiles
Classification
Psoriasis

ABSTRACT

Psoriasis is an autoimmune disease, which symptoms can significantly impair the patient's life quality. It is mainly diagnosed through the visual inspection of the lesion skin by experienced dermatologists. Currently no cure for psoriasis is available due to limited knowledge about its pathogenesis and development mechanisms. Previous studies have profiled hundreds of differentially expressed genes related to psoriasis, however with no robust psoriasis prediction model available. This study integrated the knowledge of three feature selection algorithms that revealed 21 features belonging to 18 genes as candidate markers. The final psoriasis classification model was established using the novel Incremental Feature Selection algorithm that utilizes only 3 features from 2 unique genes, IGFL1 and C10orf99. This model has demonstrated highly stable prediction accuracy (averaged at 99.81%) over three independent validation strategies. The two marker genes, IGFL1 and C10orf99, were revealed as the upstream components of growth signal transduction pathway of psoriatic pathogenesis.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Psoriasis is a widely spread chronic autoimmune disease with major symptoms on skins and joints [1,2]. Psoriasis may increase the risks of stroke and myocardial infarction [3,4] and shares the same effective drug, p40-neutralizing antibodies, with the neurodegenerative Alzheimer's disease [5]. The prevalence of psoriasis varies considerably in different geographic regions and ethnic groups, with higher rate in the Caucasian population (e.g. 2.2% in the United States) and lower rate in the Asian (e.g. less than 1% in China) [6,7]. Along with the cutaneous manifestations, psoriasis is accompanied by inflammatory arthritis in up to 40% cases [8]. Its phenotypic symptoms may also include epidermal hyperplasia, keratinocyte differentiations, angiogenesis, infiltration of T lymphocytes, dendritic cells, neutrophils, cytokines as well as chemokines [9–11]. The current diagnosis of psoriasis heavily relies on the visual observation of the skin lesion biopsy by an experienced dermatologist [12], which procedure is inefficient and labor intensive.

Majority of the large-scale psoriasis studies have employed microarray or qPCR techniques to compare the gene expression pattern in psoriatic lesion samples with healthy controls. Gudjonsson et al. systematically screened ~54,000 probe sets in the Affymetrix HG-U133 Plus 2 platform, and detected 179 unique genes for 223 probe sets with differential expression in the uninvolved psoriatic skin samples [13]. The lipid metabolism was demonstrated to be most associated with psoriasis, and three lipid metabolic related transcription factors, peroxisome proliferator-activator receptor alpha (PPARA), sterol regulatory element-binding protein (SREBF), and estrogen receptor 2 (ESR2), seem to regulate the dysregulated genes in the uninvolved psoriatic samples. Reischl et al. chose the Affymetrix HG-U133A platform to profile the expression patterns of ~22,000 probe sets in the human genome, and detected 179 genes for 203 probe sets with more than 2-fold expression changes in the psoriatic lesion skins [14]. Their data strongly suggested that the stem cell proliferation regulation Wnt pathway is associated with the psoriasis development. But it still remains elusive whether the Wnt pathway plays a causative role or merely the human immune response to the psoriatic lesion. The human immune response to the environmental stimulus, the type I interferon system, was screened in both the psoriatic lesion and uninvolved skin samples on the Affymetrix HG-U133 Plus 2 platform [15]. There were 1408 up-regulated and 1465 down-regulated probe sets detected in the psoriatic lesion skin samples, among which multiple T cell marker genes and two type I interferon inducible genes (STAT1 and ISG15) had ~3 fold or higher in expression changes. Swindell et al.

* Correspondence to: Fengfeng Zhou, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, PR China. Fax: +86-755-86392299.

** Correspondence to: Guoqing Wang, Department of Pathogeny Biology, Norman Bethune Medical College, Jilin University, Changchun, Jilin 130021, PR China.

E-mail addresses: ff.zhou@siat.ac.cn, FengfengZhou@gmail.com (F. Zhou).
URL: http://www.healthinformaticslab.org/ffzhou/ (F. Zhou).

[1] These authors contribute equally to this work.

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.11.001

[16,17] compared the gene expression files of psoriasiform versus normal samples between human and mouse. The calculated lists of top 5000 fold-change ranked up-regulated and down-regulated probe sets showed consistent overlaps among different mouse psoriasiform phenotypes. Unfortunately, the lists of marker genes barely overlapped between different studies, leaving little information to construct a psoriatic classification model.

We hypothesized that a robust disease classification model can be constructed based on marker genes. This study employed the microarray based gene expression profiles, in which three widely-used feature selection algorithms were applied to screen the psoriasis associated features (probe sets). Majority of the 21 features were known to be associated with psoriasis. We further conducted a comprehensive performance evaluation of the neural network based binary classifiers trained with the Incremental Feature Selection strategy. The classifier based on the three features of IGFL1 and C10orf99 outperformed the rest of models, with over 99.43% accuracy in all three validation tests.

2. Methods

2.1. Data collection and preprocessing

Two microarray data sets with accession numbers GSE14905 [15] and GSE13355 [16,17] were retrieved from the Gene Expression Omnibus (GEO) Database [18]. The two data sets provide gene expression profiles for the same disease type (psoriasis) and the corresponding controls, but have no overlap of samples. This makes them an ideal paired set for detecting the consistent psoriasis biomarkers. There are 176 participating subjects in total, including 91 psoriatic patients and 85 healthy controls, denoted by $P = \{P_1, P_2,\ldots, P_{91}\}$ and $N = \{N_1, N_2,\ldots, N_{85}\}$. Samples from both data sets were processed by the standard protocol for the Affymetrix Human Genome U133 Plus 2.0 platform. The detailed information of the samples used in this study is listed in the Supplementary Table S1.

The raw fluorescence intensity data within CEL files were processed by background correction, log2 transformation, and quantile normalization using the Robust Multichip Analysis (RMA) algorithm [19], as implemented in the program Affymetrix Expression Console™, which was downloaded from http://www.affymetrix.com/estore/index.jsp. The systematic batch biases were corrected by the Distance Weighted Discrimination (DWD) algorithm [20]. DWD eliminates source effects across different studies by finding a hyper-plane that separates the two systematic biases and adjusts the microarray data by projecting them on the hyper-plane through subtracting out the DWD plane multiplied by the batch mean.

We removed the probe sets with less discriminating power by measuring the overall variance. This procedure was performed in varFilter of the R package Bioconductor genefilter. We chose the default setting of the Inter-Quartile Range function IQR for the parameter var.func, and .5 as the var.cutoff. The function IQR was chosen for its robust outlier detection. This filtering step was applied on the DWD-processed data sets, and 27,336 probe sets of the 176 participating subjects were retained for further analysis.

This study investigated the binary classification problem between $N_1 = 91$ psoriatic patients and $N_2 = 85$ healthy controls. The total number of samples was $N = 91 + 85 = 176$. The expression level of each probe set was denoted as F_i, where $i \in \{1, 2,\ldots, 27{,}336\}$. The following sections aimed to detect m features F_j, where $j \in \{1, 2,\ldots, m\}$ from the $M = 27{,}336$ features of the binary classification model. We investigated various feature selection algorithms to detect psoriasis related features from 27,336 probe sets, with classification accuracy as the most important feature. The following three algorithms were chosen for their complementary performance features. ROC screens for the most significant phenotype-associated individual features, by investigating each feature's parameter-independent association [21]. SVM-RFE further optimizes the inter-feature associations, which can top-rank the important features that have weak individual phenotype-association [22]. Boruta removes the features with less contribution to classification accuracy by introducing random variables for the competition [23,24]. We believe, the consensus features derived from the three algorithms represent a reliable list of biomarker candidates. The R implemented version of the three algorithms was used in this study.

2.2. SVM-RFE feature selection algorithm

SVM-RFE is a Recursive Feature Elimination (RFE) strategy utilizing the weight vector produced by the classification model Support Vector Machine [25]. RFE is a widely used strategy that can recursively reduce the number of features by removing those features with the least phenotype correlation rankings [22,26,27]. The Support Vector Machine algorithm was invented by Vladimir N Vapnik in 1979 for the classification problem [28], and it produces a weight vector for the input features after being trained to optimize the classification accuracy between the psoriatic and healthy samples. The generated weight vector will be used to rank the features, and the RFE strategy will recursively eliminate the least ranked features. The R package mSVM-RFE was chosen as the implementation of this algorithm [22].

2.3. ROC feature selection algorithm

The receiver operating characteristic (ROC) is a graphical plot in the signal detection theory that illustrates the relationship between the true positive rate (TPR) versus the false positive rate (FPR) of a binary classifier [29]. The ROC curve is one of the best measurements to rank the features between multiple groups of tissues [30]. We ranked the features based on the ROC-based binary classification performance of the psoriatic and healthy samples. In brief, the two expression distributions of a given feature and/or gene were calculated for the psoriatic and healthy samples, respectively. Next, the area under the ROC curve (AUC) was calculated to rank the features. The R package rocc was used as the implementation of this algorithm [21].

2.4. Boruta feature selection algorithm

The Boruta feature selection algorithm is a random forest based wrapper strategy that removes features proven to be less informative than random probes [31]. The classification algorithm, random forest, was chosen because of its inexpensive calculation and parameter free for manual tuning. A random forest [22] collects votes over multiple decision tree based weak classifiers, which are independently trained over the binary classification data set between the psoriatic and healthy samples. The importance of a probe set is measured by how the classification performance is decreased due to the random permutation of the probe set among samples. The trained random forest classifier provides an importance estimate for all features [22]. The R package implementation of Boruta was used for this study [24].

2.5. Classification model

This study utilized a feed-forward neural network classifier with one hidden layer of nodes [32] to distinguish the psoriatic samples from the healthy controls. A neural network is a mathematical model with a function $f : X \rightarrow Y$, where X is a collection of input neurons and Y is a collection of output neurons. $X = \{F_j\}$, where $j \in \{1, 2,\ldots, m\}$, is the features selected by the aforementioned feature selection algorithms. There are two output nodes, $Y = \{0, 1\}$, representing the predicted results of a healthy or psoriatic sample, respectively. The input and output neurons are connected by the hidden layers of neurons. In order to avoid model over-fitting, the number of hidden layers was set to one in this study. The rest of the parameters used the default values.

Download English Version:

https://daneshyari.com/en/article/5907882

Download Persian Version:

https://daneshyari.com/article/5907882

Daneshyari.com