Research Paper

Identification of novel targets for multiple myeloma through integrative approach with Monte Carlo cross-validation analysis

Congjian Liu*, Xiang Gu, Zhenxian Jiang

Department of Orthopaedics, People's Hospital of Ri Zhao, No. 126 Tai-An Road, Ri Zhao 276826, Shandong, China

A R T I C L E   I N F O

A B S T R A C T

More than one pathway is involved in disease development and progression, and two or more pathways may be interconnected to further affect the disease onset, as functional proteins participate in multiple pathways. Thus, identifying cross-talk among pathways is necessary to understand the molecular mechanisms of multiple myeloma (MM). Based on this, this paper looked at extracting potential pathway cross-talk in MM through an integrative approach using Monte Carlo cross-validation analysis. The gene expression library of MM (accession number: GSE6477) was downloaded from the Gene Expression Omnibus (GEO) database. The integrative approach was then used to identify potential pathway cross-talk, and included four steps: Firstly, differential expression analysis was conducted to identify differentially expressed genes (DEGs). Secondly, the DEGs obtained were mapped to the pathways downloaded from an ingenuity pathways analysis (IPA), to reveal the underlying relationship between the DEGs and pathways enriched by these DEGs. A subset of pathways enriched by the DEGs was then obtained. Thirdly, a discriminating score (DS) value for each paired pathway was computed. Lastly, random forest (RF) classification was used to identify the paired pathways based on area under the curve (AUC) and Monte Carlo cross-validation, which was repeated 50 times to explore the best paired pathways. These paired pathways were tested with another independently published MM microarray data (GSE85837), using in silico validation. Overall, 60 DEGs and 19 differential pathways enriched by DEGs were extracted. Each pathway was sorted based on their AUC values. The paired pathways, inhibition of matrix metalloproteases and EIF2 signaling pathway, indicated the best AUC value of 1.000. Paired pathways consisting of IL-8 and EIF2 signaling pathways with higher AUC of 0.975, were involved in 7 runs. Furthermore, it was validated consistently in separate microarray data sets (GSE85837). Paired pathways (inhibition of matrix metalloproteases and EIF2 signaling, IL-8 signaling and EIF2 signaling) exhibited the best AUC values and higher frequency of validation. Two paired pathways (inhibition of matrix metalloproteases and EIF2 signaling, IL-8 signaling and EIF2 signaling) were used to accurately classify MM and control samples. These paired pathways may be potential bio-signatures for diagnosis and management of MM.

1. Introduction

Multiple myeloma (MM) is an incurable cancer of plasma cells, caused by aberrant expansion of monoclonal plasma B cells in the bone marrow [1]. MM accounts for 10% of all hematological cancers, and is characterized by wide clinical and pathophysiologic heterogeneities, with lethal outcomes. The median survival time of MM patients is 7–8 years [2]. Thus, a better understanding of MM biology will aid in developing new therapeutic modalities that could potentially cure MM.

Many advances in the understanding of MM pathogenesis have been the result of major developments in genomic technologies [3, 4]. In recent years, genomic technologies identified certain disease-related biomarkers [5,6]. For example, Leone et al. [7] have demonstrated that

CDKN2C plays an important role in the progression and clinical outcome of MM. Jagani et al. [8] have indicated that *Bmi-1* is crucial for MM growth. However, the reproducibility and overlap of the extracted genes are poor. Generally, gene biomarkers obtained from gene-based classification methods are often produced independently. Due to this, the gene signatures might not synergistically improve the overall classification ability.

In an attempt to overcome these shortcomings, it is important to understand the complicated interactions between genes, to help elucidate essential principles of cellular systems and the disease machinery [9]. To obtain a clear interpretation of genomic results, pathway analysis is the first criteria to identify abnormal pathways, to shed light on the potential biology of genes, thereby decreasing complexity and

promoting explanatory power [10]. Moreover, several studies have reported that pathways-based classifiers are more reproducible and usually achieve better results, as compared to the single gene bio-markers-based classifier [11].

Remarkably, more than one pathway is involved in the development of cancer and its progression, due to the complex characterization of biological systems. Two or more pathways may be interconnected to further affect the disease onset, as functional proteins might participate in multiple pathways [12]. Therefore, identifying cross-talk between pathways is important to understand the molecular mechanisms of MM. Intuitively, different pathways influence each other, but at present, there is no reliable method to quantify the amount of cross-talk be-tween paired pathways [13]. An integrative approach using Monte Carlo cross-validation has been created, to quantify the cross-talk be-tween paired pathways.

Therefore, this study considers gene expression profile and biolo-gical pathway data as study objects, and utilizes Monte Carlo cross-validation analysis to detect pathway cross-talk in MM. The pathway cross-talk may be potential signatures for early detection and treatment of MM.

2. Materials and methods

The integrative approach using Monte Carlo cross-validation method comprised of four steps: Firstly, identification of differentially expressed genes (DEGs) between MM and healthy control samples was conducted. Secondly, the DEGs obtained were mapped to the pathways downloaded from the database of ingenuity pathways analysis (IPA), to reveal the relationship between DEGs and pathways enriched by these DEGs. A subset of pathways enriched by DEGs were also required. Thirdly, a discriminating score (DS) value for each pair of pathways was computed. Lastly, random forest (RF) classification was used to identify paired pathways with high cross-talk, and Monte Carlo cross-validation analysis was then repeated 50 times, to find the best paired pathways. During the analysis, all steps were repeated 50 times. After 50 runs, the top 10 paired pathways with the best AUC were extracted and were considered as significant paired pathways. These best paired pathways were tested with another independently published MM microarray data (GSE85837) using in silico validation.

2.1. Acquisition of gene expression profile

The gene expression library of MM (accession number: GSE6477) [14] was retrieved from Gene Expression Omnibus (GEO) database, based on the GPL96 platform of [HG-U133A]Affymetrix Human Genome U133A Array. There were 150 MM samples and 12 healthy donor plasma cell samples, in the GSE6477 profile. The repeated probes were first eliminated, after obtaining the microarray profile. The re-maining probes were then mapped to the genomics, to further obtain human gene symbols. Finally, 12,437 genes were identified for sub-sequent analysis.

2.2. DEGs identification

During the research, "normalized quantile" was used to conduct normalization on mRNAs, with an average value determined across 162 samples [15]. Genes with values greater than 0.25-fold quantile average across all samples were extracted. Compared to several other estima-tors, the quantile-adjusted conditional maximum likelihood (qCML) was the most reliable in terms of bias, on a wide range of conditions, and performed best in smaller samples with a common dispersion quotient [16]. The qCML of edgeR package from Bioconductor was used to verify if these genes were expressed in a differential manner. Next, raw P values were corrected using a Benjamini-Hochberg method, based on a false discovery rate (FDR) [17]. Genes were considered differentially expressed when FDR was less than 0.001 and |log fold change (FC)|was more than 2.

2.3. Pathway enrichment analysis

Ingenuity pathways analysis (IPA), is widely utilized as a pathway database to analyze gene expression profile in the context of known biological responses and higher-order response pathways. In the present study, pathway enrichment analysis for DEGs was implemented using Fisher's exact test based on IPA tool, with the goal to extract significant pathways enriched by DEGs between MM and control samples. Hence, 589 biological pathways deposited in the IPA database were first downloaded. After the Fisher's exact test was applied to the genes in the IPA pathways and DEGs, the pathways enriched with P-value less than 0.01 were extracted. Later, the Benjamini-Hochberg procedure was used to correct the P values. Differential pathways were extracted based on the significance of FDR < 0.05.

2.4. DS calculation for pathway cross-talk

DS is an index used to compare the expression levels in the subgroup of samples showing amplification and in samples without amplification [18]. Thus, DS was employed to analyze the pathway cross-talk, in this paper. The DS was counted by comparing the gene expression levels of each paired pathway enriched by DEG in each sample, based on the description in the study by Cava et al. [19]. Generally, DS score im-plicates the relationships between paired pathways, and a larger DS suggests higher difference of activity between pathways.

2.5. Extracting the best paired pathways

RF created by Breiman [20], is a statistical method used to handle two issues of variable selection. To classify this methodology, an RF classification model was applied on the paired pathways based on the DS values of each sample. This helped to classify both MM and the control samples. AUC was calculated by a 10-fold cross-validation method, based on the following indexes: mtry and ntree. The mtry (the number of variables randomly sampled as candidates at each split) was equivalent to sqrt (p); p was the number of variables in the data matrix; and ntree (the number of trees grown) was equal to 500. All AUC values were then ranked in descending order, and the top 10 paired pathways were selected.

As documented in the validation analysis, the sample size in the training set was generally larger than that in the testing set. Of note, is the ratio of 6 to 4, which is a common distribution proportion. For example, Zhang et al. [21] randomly selected 60% as the training set and the remaining 40% as the testing data. Thus, in this study, the Monte Carlo cross-validation method was employed to randomly select 60% of the original microarray data comprising of training data, and the remaining 40% was assigned to the testing data. This step was re-peated 50 times, randomly forming new training and test datasets each time. For each bootstrap, DEGs and pathway enrichment analysis for DEGs, and a DS for the paired pathways was conducted. For each bootstrap, a training set was used to detect a DS for the top 10 paired pathways with the best AUC value between the two groups. For each bootstrap, a testing partition was employed to confirm the top 10 paired pathways. At the end of 50 runs, the list of the top 10 paired pathways sorted by descending rank were selected, such that each pathway pair was extracted in 50 bootstraps. Ultimately, the top 10 paired pathways ranked for all 50 runs, were regarded as significant.

2.6. In silico validation with independent MM microarray data

To predict these best paired pathways, other MM data of 15 patients with MM and 13 control patients, was obtained from the publicly available microarray dataset GSE85837. For validation, all steps and selection criteria were the same as the above analysis.