Mutation Research 791 (2016) 27–34

Contents lists available at ScienceDirect

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

journal homepage: www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Mining potential biomarkers associated with space flight in *Caenorhabditis elegans* experienced Shenzhou-8 mission with multiple feature selection techniques

Lei Zhao[a], Ying Gao[b], Dong Mi[c,*], Yeqing Sun[a,*]

[a] *Institute of Environmental Systems Biology, College of Environmental Science and Engineering, Dalian Maritime University, Dalian 116026, People's Republic of China*
[b] *Center of Medical Physics and Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Shushanhu Road 350, Hefei 230031, People's Republic of China*
[c] *Department of Physics, Dalian Maritime University, Dalian 116026, People's Republic of China*

ARTICLE INFO

Article history:
Received 19 June 2016
Accepted 15 August 2016
Available online 17 August 2016

Keywords:
Spaceflight
Space radiation
Biomarkers
Microarray
Caenorhabditis elegans
Feature selection

ABSTRACT

To identify the potential biomarkers associated with space flight, a combined algorithm, which integrates the feature selection techniques, was used to deal with the microarray datasets of *Caenorhabditis elegans* obtained in the Shenzhou-8 mission. Compared with the ground control treatment, a total of 86 differentially expressed (DE) genes in responses to space synthetic environment or space radiation environment were identified by two filter methods. And then the top 30 ranking genes were selected by the random forest algorithm. Gene Ontology annotation and functional enrichment analyses showed that these genes were mainly associated with metabolism process. Furthermore, clustering analysis showed that 17 genes among these are positive, including 9 for space synthetic environment and 8 for space radiation environment only. These genes could be used as the biomarkers to reflect the space environment stresses. In addition, we also found that microgravity is the main stress factor to change the expression patterns of biomarkers for the short-duration spaceflight.

1. Introduction

The "biomarker" was defined as any detectable changes at the molecular, cellular, and physiological levels in exposed organisms [1], which could be divided into three different classes: biomarkers of exposure, sensitivity and disease [2]. Biomarkers of exposure refer to the biological alterations for which a dose-response relationship can be validated [3]. During space flight, many biomarkers have been used to indicate the radiation dose that the astronauts received [4]. Chromosome aberrations [5] and micronuclei [6], for example, which are usually considered to be two such important biomarkers. However, a large interindividual variability was also found in the actual measurements, and this variability had no correspondence with the measured dose [7]. This is mainly due to the differences of individual radiosensitivity with genetic and/or physiological background [7], and the large statistical errors associated with each measurement [8].

With recent developments in high throughput screening, it is found that many other biomarkers, including changes in gene expression [9], protein expression and post-translational modifications [10], and metabolic profile [11], also have the potential to be applied to estimate the radiation dose and even space radiation risk [1,7]. Notably, flight experiments are expensive, restricted to samples, and hard to repeat. And, the biological effects are usually small and often below the detection threshold except the "omics" technology, e.g., microarray. Microarray technology is now extensively used to mine potential biomarkers (or key genes) in space biology researches. The characteristics of microarray datasets obtained directly after spaceflight was always far smaller than the features (genes) to be investigated, which can be termed as small samples with big data.

In such cases, the key question of the high-throughput profiles is how to analyze large amounts of data to detect a panel of biomarkers reflecting the exposure to the environment. At present, at least two different strategies can be adopted to deal with this situation [12]. One strategy attempts to use relevant biological knowledge to reduce the set of genes to a manageable number, while the other ignores the dependencies between genes and analyses the data gene-by-gene. Our previous studies mainly focused on the first strategy in order to select the special genes involved in the biological responses (DNA damage responses, apoptotic gene expression, and miRNAs expression, etc.) of spaceflight environments for Shenzhou-8 mission [13,14]. By integrated analysis of miRNA and mRNA, we have found that microgravity probably enhanced the biological responses in the presence of space radiation, and suggested a possible synergistic interaction between space radiation and microgravity [13–15]. However, we do not know what kinds of genes are important and should be studied due to the lack of understanding of the function or pathway of these genes.

Different techniques have been proposed and implemented to select the differentially expressed (DE) genes based on the second strategy

* Corresponding authors at: No. 1 Linghai Road, Dalian, Liaoning, 116026, People's Republic of China.
E-mail addresses: mid@dlmu.edu.cn (D. Mi), yqsun@dlmu.edu.cn (Y. Sun).

http://dx.doi.org/10.1016/j.mrfmmm.2016.08.002
0027-5107/

[16]. Historically, the first method used to identify DE genes was the "fold change", which is performed through a simple fold change cutoff, typically between 1.8 and 3.0 [17,18]. However, the choice of threshold has a certain degree of arbitrariness, which may give rise to both false negative and false positive results [16]. We have ever used the traditional method of fold change was used to detect the DE genes in the microarray datasets in our previous studies, while the defects in this method are obvious. For example, some genes with small fold-change may have some important biological functions, such as transcription factors [16–18]. In addition, the traditional method does not consider the background noise and variability of microarray datasets [16].

In order to overcome the problem of "fold change", the feature selection techniques have been proposed to meet the challenges of biomarkers screening [19]. These techniques can be divided into three categories: filter methods, wrapper methods and embedded methods. Each of them possesses advantages as well as disadvantages. Briefly, the filter methods (such as Inter Quantile Range (IQR) [20], *t*-test [21], analysis of variance (ANOVA) [22], Wilcoxon rank sum [23], etc.) identify the relevance of features by looking only at the intrinsic properties of the data, i.e., the method of "gene-by-gene", which are computationally simple, fast and independent of the classification algorithm [19]. However, the flaws of filter methods are mainly ignoring the feature dependencies, which may lead to worse classification performance. Wrapper methods (such as sequential search [24], genetic algorithms [25], etc.) embed the model hypothesis search within the feature subset search, while the common drawback of these methods is that they all have a higher risk of over-fitting and computational cost than filter techniques [19]. In addition, the embedded methods (such as Random forest (RF) [26], Support vector machine (SVM) [27], etc.) take an optimal subset of features to build into the classifier construction, while at the same time being far less computationally intensive than wrapper methods [19].

To effectively obtain the DE genes induced by spaceflight environments, it is necessary that more precise comparisons between limited samples should be made by considering the joint distribution of gene expressions. In addition, to overcome the deficiency of the traditional method of fold change, in the present work, we propose a combination algorithm of feature selection techniques to mine the potential biomarkers of space environment exposures. This algorithm takes into account both the advantages of filter and embedded methods with less computation time and importance ranking. And then it was used to clarify the microgravity effects on biological responses to space radiation.

2. Materials and methods

The processes of data collection and analysis are shown in Fig. 1, and the details can be found in the following subsections. All steps of microarray data analysis were performed using R software (R version 2.15.3). Several R programs, including algorithm implementations, statistical calculations and plot artworks used in this study, are freely available upon request to corresponding authors.

2.1. Sources of data

The microarray dataset used in this study was obtained from the dauer larvae of *Caenorhabditis elegans* experienced the Shenzhou-8 mission, which belonged to the "International Space Biological Experiments" cooperative projects between China and Germany. The Shenzhou-8 mission was flown from 1 to 17 November 2011, with a total mission time of 16.5 days. The experimental containers used in this project were special SIMBOX devices, which were manufactured by EADS Astrium Friedrichshafen. More details can be found in previous publication [28].

Dauer *C. elegans* larvae were considered as an ideal model organism to avoid interference between different generations and development stages. Dauer larvae of *C. elegans*, including *dys-1* mutant, *ced-1* mutant, and wild-type, were divided into nine groups and put into the devices. The mutation in *cx18* results in the expression of truncated DYS-1 protein products, in which the C-terminus, a region essential for the protein's function, is removed. And this mutation does not result in any obvious muscular degeneration phenotypes but lead to hyperactivity of the body wall muscle [29,30]. The mutation in *ced-1* displays a relatively weak defect in engulfment of apoptotic cells, because *ced-1* as checkpoint and apoptosis related genes in *C. elegans* could mediate the engulfment of cell corpses [31]. In spite of the different responses of three genetypes on the space environment stresses, the common DE genes in *C. elegans* with different genetypes can reflect the widespread mechanism of biological responses. And, different genetypes in *C. elegans* can show different endogenous sensitivities with physiological background to space environment stresses, and can reflect a large interindividual variability in space biology.

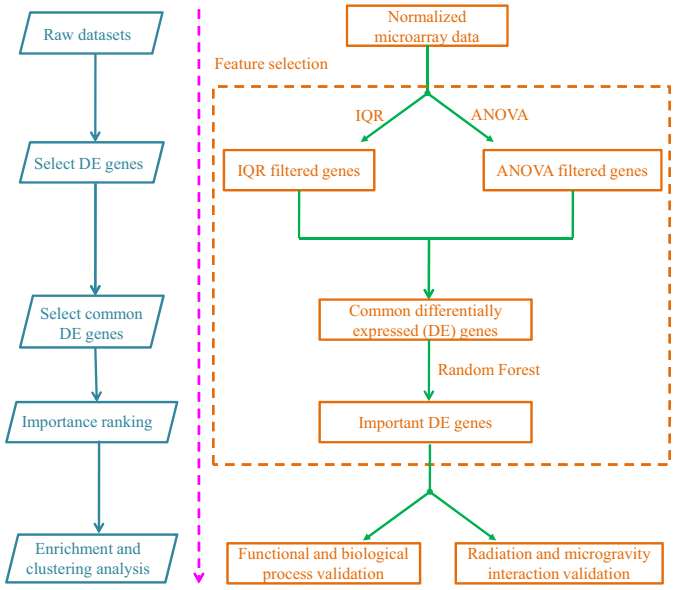

Fig. 1. Flow-chart of data analysis in this study. The process of data processing mainly includes four steps: data normalization, common differentially expressed genes selection, importance genes ranking, and enrichment and clustering analysis.

The devices were exposed under three exposure conditions: spaceflight (SF), spaceflight control (SC), and ground control (GC) (see Supplementary Table S1). SF samples were put in a fixed device on the Shenzhou-8 spacecraft, which were affected by space synthetic environments including space radiation (1.92 mGy) and microgravity (±0.005 g); SC samples were put in a centrifugal device with 74.4 rpm on the Shenzhou-8 spacecraft, which were mainly affected by space radiation environment (2.27 mGy) without microgravity; GC samples placed on the ground in parallel at the Payload Integration Test Center in Beijing, used as a control, were mainly affected by 1 g gravity.

During 7 h after the landing of the re-entry vehicle, all samples were fixed with liquid nitrogen and maintained until microarray detection. About 2000 worms from each sample were collected and total RNA was isolated and assessed by the ratio of $OD_{260}/OD_{280} > 1.9$. The NimbleGen Gene Expression Profiling service was performed by KangChen Bio-tech Inc. (Shanghai, China). Microarray data from mRNA expression profiling were validated by qRT-PCR from independently-isolated RNA samples. More details can be found in our previous publications [13–15].

2.2. Data pre-processing

The raw dataset of microarray data obtained in the Shenzhou-8 mission were normalized to have a common mean over a set of whole genes by estimating scaling factors [32]. Such normalization scheme is to correct for systematic differences between genes or arrays in different conditions. Moreover, the *i*th sample is represented by $N = 18{,}186$ features in such form $x_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$.

2.3. Feature selection

2.3.1. IQR algorithm

IQR algorithm is a commonly used filtering approach, which can remove the genes without significant change in expression across all samples, and tend to provide little discriminatory power [12]. In addi-