Genomics Data 3 (2015) 8–14

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Global gene expression profiling data analysis reveals key gene families and biological processes inhibited by Mithramycin in sarcoma cell lines

Kirti K. Kulkarni [a], Kiran Gopinath Bankar [a], Rohit Nandan Shukla [a], Chandrima Das [b], Amrita Banerjee [b], Dipak Dasgupta [b], Madavan Vasudevan [a,*]

[a] Genome Informatics Research Group, Bionivid Technology Pvt Ltd., Bangalore 560043, India
[b] Biophysics and Structural Genomics Division, Saha Institute of Nuclear Physics, Kolkata 700064, India

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form 31 October 2014
Accepted 3 November 2014
Available online 8 November 2014

Keywords:
Mithramycin
Sarcoma
Microarray
Global gene expression

ABSTRACT

The role of Mithramycin as an anticancer drug has been well studied. Sarcoma is a type of cancer arising from cells of mesenchymal origin. Though incidence of sarcoma is not of significant percentage, it becomes vital to understand the role of Mithramycin in controlling tumor progression of sarcoma. In this article, we have analyzed the global gene expression profile changes induced by Mithramycin in two different sarcoma lines from whole genome gene expression profiling microarray data. We have found that the primary mode of action of Mithramycin is by global repression of key cellular processes and gene families like phosphoproteins, kinases, alternative splicing, regulation of transcription, DNA binding, regulation of histone acetylation, negative regulation of gene expression, chromosome organization or chromatin assembly and cytoskeleton.

Specifications	
GEO accession	GSE25127
Organism	*Homo sapiens*
Cell line	Ewing sarcoma cell lines (TC71 and TC32)
Sex	–
Array type	Expression profiling by array
Platform	GPL570 [HG-U133_Plus_2] Affymetrix human genome U133 plus 2.0 array
Data format	CEL files
Experimental factors	The data consist of 12 arrays. Two cell lines, TC71 and TC32, were treated with solvent control or with Mithramycin, and RNA was extracted at 6 h. Three biological replicates per cell line/treatment
Experimental features	The study aims to define gene expression changes associated with Mithramycin treatment of Ewing sarcoma cell lines
Consent	–
Sample source location	Bethesda, MD — 20892, USA

Data files

Accession	Title	Source name	Cell line	Treatment
GSM617274	TC32-M1	TC32 cell line, Mithramycin	TC32	Mithramycin
GSM617275	TC32-M2	TC32 cell line, Mithramycin	TC32	Mithramycin
GSM617276	TC32-M3	TC32 cell line, Mithramycin	TC32	Mithramycin
GSM617277	TC32-S1	TC32 cell line, control	TC32	Control
GSM617278	TC32-S2	TC32 cell line, control	TC32	Control
GSM617279	TC32-S3	TC32 cell line, control	TC32	Control
GSM617280	TC71-M1	TC71 cell line, Mithramycin	TC71	Mithramycin
GSM617281	TC71-M2	TC71 cell line, Mithramycin	TC71	Mithramycin
GSM617282	TC71-M3	TC71 cell line, Mithramycin	TC71	Mithramycin
GSM617283	TC71-C1	TC71 cell line, control	TC71	Control
GSM617284	TC71-C2	TC71 cell line, control	TC71	Control
GSM617285	TC71-C3	TC71 cell line, control	TC71	Control

Material and methods

Gene expression data for reanalysis was obtained from Gene Expression Omnibus (GEO) database NCBI with the link. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25127. The raw data (CEL file) was normalized and processed using GeneSpring GX V 12.5 (Agilent Technologies Inc., Santa Clara, USA).

* Corresponding author.
E-mail address: madavan@bionivid.com (M. Vasudevan).

http://dx.doi.org/10.1016/j.gdata.2014.11.001

Raw data summarization

All the samples raw data were summarized using the Robust Multi Array Average (RMA) method. RMA is a background correction method that is based on the distribution of Perfect Match (PM) values among probes on an Affymetrix array. It can be used attaching a standard error (SE) to the quantity using a linear model that removes probe-specific affinities [1]. Background corrected, log transformed and Quantile normalized arrays were used and to protect from outliers robust procedures like median Polish are used [2]. Median Polish is an iterative process which operates on a matrix by alternately extracting row and column medians. The convention followed is that the iteration starts with extracting medians for arrays (across probes). Iteration continues until convergence or until a limit on the number of iterations is reached. The limit is of 50 iterations [2].

Normalization

RMA summarized raw data was Quantile normalized to calculate probe level expression values. Quantile is most widely used preprocessing technique designed to remove technological noise in genomic data. It makes the empirical distribution of all the gene expressions same in the whole experiment [3]. Thus after normalization, all statistical parameters of the sample, i.e., mean, median and percentiles of all samples will be identical. With Quantile normalization (QUANT), a reference array of empirical quantiles, denoted as $q = (q_1,q_2,\ldots,q_m)$, is first computed by taking the average across all ordered arrays. Let $y^c_{(1),j} \leq y^c_{(2),j} \leq \cdots y^c_{(m),j}$ denote the ordered gene expression observations in the jth array ($j = 1,2,\ldots,n$) of the cth ($c = A,B$) group, the rth ($r = 1,2,\ldots,m$) element of this reference array is as follows [3].

$$\boldsymbol{q_r} = \frac{1}{2\boldsymbol{n}}\left(\sum_{k=1}^{n} \boldsymbol{y^A_{(r),k}} + \sum_{l=1}^{n} \boldsymbol{y^B_{(r),l}}\right).$$

Baseline transformation

In order to improve the sensitivity of the measurement, baseline transformation of the normalized data is done. This step includes subtraction of an estimated background signal, subtracting the reference signal. Variance ratios were computed for the data set after shifting all measurements upwards by a number of medians for the channel, and subsequently taking the algorithm [4].

Quality control analysis

Quality control of normalized data is critical to identify inliers and outliers and multiple testing methods are applied for critical evaluation of the data quality.

Box-Whisker plot is a visualization method that requires a sample size of only 5 for analysis [5]. It characterizes a sample using the 25th-lower quartile (Q1), 50th-median (m or Q2) and 75th percentiles-upper quartile (Q3) and the interquartile range (IQR = Q3 − Q1), that covers the central 50% of the data. Quartiles are insensitive to outliers and preserve information about the center and spread. The core element that gives the box plot its name is a box whose length is the IQR and its width is arbitrary [5]. A line inside the box shows the median, which is not necessarily central. Whiskers are conventionally extended to the most extreme data point that is no more than 1.5 × IQR from the edge of the box or all the way to minimum and maximum of the data values.

Analysis of hybridization controls in the microarray

The hybridization controls show the signal value profiles of the transcripts (only 3′ probe sets are taken) where a line graph is plotted with X axis representing Biotin labeled cRNA transcripts and the Y axis represents the log of the normalized signal values. Typical quality observation is indicated by all samples adhere to the same trend line of internal controls.

Principal component analysis

(PCA) is a statistical technique for determining the key variables in a multidimensional data set which explains the differences in the observations [6]. PCA is computed by considering the n eigenvalues and their corresponding eigenvectors that are calculated from the n × n covariance matrix of conditions. Each eigenvector defines a principal component. A component can be viewed as a weighted sum of the conditions, where the coefficients of the eigenvectors are the weights. The projection of gene i along the axis defined by the jth principal component is:

$$\boldsymbol{a'_{ij}} = \sum_{t=1}^{n} a_{it}\, v_{tj}$$

where v^{tj} is the tth coefficient for the jth principal component; a^{it} is the expression measurement for gene i under the tth condition. A′ is the data in terms of principal components. Since V is an orthonormal matrix, A′ is a rotation of the data from the original space of observations to a new space with principal component axes. The variance for each of the components is associated with its eigenvalue; it is the variance of a component over all probes [6]. Consequently, the eigenvectors with large eigenvalues are the ones that contain most of the information; eigenvectors with small eigenvalues are uninformative.

Correlation-Coefficient analysis reveals the correlation analysis across arrays. It is calculated using Pearson Correlation coefficient as follows:

$$\boldsymbol{\rho_{X,Y}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where σ_X is the standard deviation of X, μ_X is the mean of X, and E is the expectation.

Condition tree is a hierarchical clustering method where a tree of genes is built by successively finding the two most similar gene expression patterns from the complete data set [7]. It makes use of Distance metric and linkage rule. Distance metric used is Pearson uncentered which is similar to Pearson Correlation coefficient except that the entities are not mean-centered. It is calculated by the following formula

$$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}.$$

Average-linkage rule was used for clustering. This algorithm computes a dendrogram that assembles all elements into a single tree. For any set of n genes, an upper-diagonal similarity matrix is computed that contains similarity scores for all pairs of genes. This matrix is scanned to identify the highest value. A node is created to join these two genes, and a gene expression profile is computed for the node by averaging observation for the joined elements. The similarity matrix is updated with the new node replacing the two joined elements, and this process is repeated until only a single element remains [8].

Identification of differentially expressed genes

The volcano plot method is one of the most widely used method to identify statistically significant differentially expressed genes between two conditions. Each point in volcano plot represents a probe set or a gene, and the x-coordinate represents the (log) fold-change (FC) and y represents the t-statistic or −log10 of the p-value from a *t*-test. The log (FC) is the unstandardized measure of differential expression, but t-statistic is a noise-level-adjusted standardized measure [9]. In the