Feb 05, 2021

Expression analysis about cell lines in CCLE

PLOS One

Keita Fukuyama[1], iwaisako [1]

[1]Division of Hepato-Biliary-Pancreatic and Transplant Surgery, Department of Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan

1 *Works for me* dx.doi.org/10.17504/protocols.io.bq5vmy66

K Keita Fukuyama

ABSTRACT

Cancer cell lines are widely used in basic research to study cancer development, growth, invasion, or metastasis. They are also used for the development and screening of anticancer drugs. However, there are no clear criteria for choosing the most suitable cell lines among the wide variety of cancer cell lines commercially available for research, and the choice is often based on previously published reports. Here, we investigated the characteristics of liver cancer cell lines by analyzing the gene expression data available in the Cancer Cell Line Encyclopedia. Unsupervised clustering analysis of 28 liver cancer cell lines yielded two main clusters. One cluster showed a gene expression pattern similar to that of hepatocytes, and the other showed a pattern similar to that of fibroblasts. Analysis of hepatocellular carcinoma gene expression profiles available in The Cancer Genome Atlas showed that the gene expression patterns in most hepatoma tissues were similar to those in the hepatocyte-like cluster. With respect to liver cancer research, our findings may be useful for selecting an appropriate cell line for a specific study objective. Furthermore, our approach of utilizing a public database for comparing the properties of cell lines could be an attractive cell line selection strategy that can be applied to other fields of research.

EXTERNAL LINK

https://github.com/fk506cni/cell_line_comparison

DOI

dx.doi.org/10.17504/protocols.io.bq5vmy66

EXTERNAL LINK

https://github.com/fk506cni/cell_line_comparison

PROTOCOL CITATION

Keita Fukuyama, iwaisako 2021. Expression analysis about cell lines in CCLE. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bq5vmy66

KEYWORDS

cell line, microarray, liver cancer, expression analysis

LICENSE

CREATED

Jan 04, 2021

LAST MODIFIED

Feb 05, 2021

PROTOCOL INTEGER ID

45973

Cell line comparison

1 1 Data acquisition

mRNA expression data (CCLE_Expression.Arrays_2013-03-18.tar.gz) and annotation files (CCLE_Expression.Arrays.sif_2012-10-18.txt) were downloaded from the CCLE (https://portals.broadinstitute.org/ccle). The Catalogue of Somatic Mutations in Cancer (COSMIC) data were downloaded from the ENBL-EBI website (https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-3610/). The microarray data were normalized using the MAS5 or RMA method, and log2 conversion was performed for the signal intensity of each probe set ID. RNA-seq data and annotation data were downloaded from the NCI website (ftp://caftpd.nci.nih.gov/pub/OCG-DCC/CTD2/TGen/CCLE_RNA-seq_Analysis/).

RNA-seq data and clinical information regarding HCC from TCGA, including the normalized transcripts per million (TPM) data for clustering and visualization (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/LIHC/20160128/gdac.broadinstitute.org_LIHC.Merge_rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.Level_3.2016012800.0.0.tar.gz), raw count data for identification of DEGs (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/LIHC/20160128/gdac.broadinstitute.org_LIHC.Merge_rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes__data.Level_3.2016012800.0.0.tar.gz), and clinical information for annotation (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/LIHC/20160128/gdac.broadinstitute.org_LIHC.Clinical_Pick_Tier1.Level_4.2016012800.0.0.tar.gz) were obtained from the Broad Institute website (https://gdac.broadinstitute.org/).

Clinical TCGA information was obtained from the GDC portal (https://portal.gdc.cancer.gov/repository, S1 Table).

Gene expression data of various primary cells were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/gds/?term=GSE49910[Accession]).

2 Clustering

We performed unsupervised hierarchical clustering analysis for mRNA expression data of cancer cell lines, HCC samples, or primary cells using Spearman correlation and average linkage or Ward's method. We uploaded the codes used in this study to GitHub (https://github.com/fk506cni/cell_line_comparison/tree/master/analysis/77GEarray_re) .

3 Literature review of liver cancer cell lines

Original articles on the establishment of liver cancer cell lines were searched [13-25], and the age of the patients was determined using the Brunner-Munzel test.

4 Identification of differentially expressed genes (DEGs)

We examined the difference in the expression data of the two liver-cancer cell-line clusters. The corrected p-values (obtained by moderated t-statistics using the eBayes function of the Bioconductor package "limma") and the q-values (adjusted by the Benjamini-Hochberg method) were calculated for each probe set ID. DEGs were defined as probe set IDs with q-values <0.0001. Additionally, we classified DEGs by the difference in the log2 average between the two clusters. DEGs with higher expression of cluster A cell lines than those of cluster B were defined as A genes, and those with higher expression of cluster B cell lines were defined as B genes.

5 Principal Component Analysis (PCA)

Principal component analysis (PCA) for DEGs was performed to confirm the ability of our approach to distinguish the cell types in clusters A and B. Further, we carried out a discriminant analysis of the principal components and evaluated the validity of the clusters using the adegenet package in R.

6 Comparison of DEGs from RNA-seq data

We evaluated the DEGs in the HCC RNA-seq data acquired from TCGA. For visualization, we used log-transformed TPM as the expression level [specifically log2(TPM + 1)]. For clinical samples from TCGA, the fold change between the two clusters was calculated. Genes were then sorted according to fold change after subtracting the average expression level between two clusters.

7 TCGA clinical data analysis

We analyzed the clinical data of patients classified into two clusters defined based on cell line analysis. The weight of the surgical specimen, a continuous variable, was assessed using the Brunner-Munzel test. The operating procedure, a categorical variable, was assessed using Fisher's exact test. Cox proportional-hazards model and log-rank test were used for survival analysis. A p-value of less than 0.05 was considered statistically significant.

Citation: Keita Fukuyama, iwaisako (02/05/2021). Expression analysis about cell lines in CCLE. https://dx.doi.org/10.17504/protocols.io.bq5vmy66

8 Gene Ontology (GO) enrichment analysis

We performed GO enrichment analysis of the A and B genes through DAVID (https://david.ncifcrf.gov/) via the Bioconductor “RDAVIDWebService” package, with a Benjamini-Hochberg-adjusted p-value < 0.2. We used “GOTERM_BP_FAT” level annotation. The top ten GO terms were visualized using bar plots, with the q-value and the number of enriched genes.

9 Mapping to primary cell atlas

We normalized all CEL files, including liver cancer cell lines and primary cells, and integrated them. Sample categories were annotated from the metadata. Expression levels were visualized using a heatmap.

10 Calculation of representative vectors for each cell type

The vector of the expression pattern for each cell type was generated as the log2 median of each probe set ID value (S1 Document).

11 Statistical Analysis

Statistical analysis and data visualization of the expression array and RNA-seq were performed on the R studio (version 1.2.5042) server and R studio (1.3.959) (https://rstudio.com/) using several R (version 3.4.1, 3.5.3, and 3.6.3) and Bioconductor (3.5 and 3.6) packages (S1 Document). Analysis codes were deposited to GitHub (https://github.com/fk506cni/cell_line_comparison). Comparison of a continuous variable in 2 groups was assessed using the Brunner-Munzel test. Qualitative data were compared using the Fisher’s exact test to compare categorical variables. A p-value of less than 0.05 was considered statistically significant.

Citation: Keita Fukuyama, iwaisako (02/05/2021). Expression analysis about cell lines in CCLE. https://dx.doi.org/10.17504/protocols.io.bq5vmy66