Bioinformatics, 32(15), 2016, 2399–2401
doi: 10.1093/bioinformatics/btw162
Advance Access Publication Date: 25 March 2016
Applications Note

Systems biology

CellMiner Companion: an interactive web application to explore CellMiner NCI-60 data

Sufang Wang[1], Michael Gribskov[1], Tony R. Hazbun[2,3] and Pete E. Pascuzzi[3,4,5,*]

[1]Department of Biological Sciences, [2]Department of Medicinal Chemistry and Molecular Pharmacology, [3]Purdue University Center for Cancer Research, [4]Department of Biochemistry and [5]Purdue University Libraries, Purdue University, West Lafayette, IN 47907, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 11, 2015; revised on March 1, 2016; accepted on March 18, 2016

Abstract

Summary: The NCI-60 human tumor cell line panel is an invaluable resource for cancer researchers, providing drug sensitivity, molecular and phenotypic data for a range of cancer types. CellMiner is a web resource that provides tools for the acquisition and analysis of quality-controlled NCI-60 data. CellMiner supports queries of up to 150 drugs or genes, but the output is an Excel file for each drug or gene. This output format makes it difficult for researchers to explore the data from large queries. CellMiner Companion is a web application that facilitates the exploration and visualization of output from CellMiner, further increasing the accessibility of NCI-60 data.

Availability and Implementation: The web application is freely accessible at https://pul-bioinformatics.shinyapps.io/CellMinerCompanion. The R source code can be downloaded at https://github.com/pepascuzzi/CellMinerCompanion.git.

Contact: ppascuzz@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The National Cancer Institute Developmental Therapeutics Program initiated a 60 human tumor cell line (NCI-60) anticancer drug screen over two decades ago (Shoemaker, 2006). The NCI-60 database includes drug sensitivity data from more than 50 000 compounds and molecular data such as gene expression, miRNA expression, protein expression, genetic variation and DNA copy number. The NCI-60 data have been used to determine drug mode of action and potential molecular targets (Shoemaker, 2006). The NCI-60 data have also been used to develop predictive models for drug sensitivity (Abaan *et al.*, 2013).

The CellMiner database (http://discover.nci.nih.gov/cellminer/) provides access to quality-controlled NCI-60 pharmacological and molecular data (Reinhold *et al.*, 2012). CellMiner also provides analysis tools to explore the NCI-60 data. The Pattern Comparison Tool enables the discovery of drugs or genes with similar patterns of sensitivity or expression across the NCI-60 panel. The query output is a spreadsheet that includes a Pearson correlation coefficient between the query drug(s) or gene(s) and every other drug and gene in the database. Researchers can then submit up to 150 of these identifiers to CellMiner to obtain additional data in a single query. The result is multiple spreadsheets making integrative analysis of the results difficult for researchers that lack computer programming skills.

The recent R package, rcellminer (Luna *et al.* 2015; Reinhold *et al.*, 2012), provides a powerful programmatic interface to the NCI-60 data, duplicating and enhancing the tools on the CellMiner website. However, rcellminer is designed for experienced R users.

Here, we present CellMiner Companion, a web application that enables researchers to explore the output of CellMiner queries. The data from multiple files is summarized, assembled into a single data matrix, *z*-score normalized, clustered and visualized both as a heatmap and dendrogram. Users can interactively change many parameters to explore the relationship among drugs or genes across the

NCI-60 cell lines. Importantly, this tool increases the accessibility of the NCI-60 data to those lacking programming skills. Further, CellMiner Companion could be incorporated easily into workflows that use rcellminer.

2 Implementation

CellMiner Companion is implemented in R as a Shiny application (Chang *et al.*, 2016; R Core Team, 2015). Data can be uploaded as multiple Excel files returned by CellMiner or as a single Excel file generated from a template. The data is parsed from the files and reformatted as a single data matrix. A data summary is produced that includes the number of replicate experiments that underlie the dataset, the range of values, and both the number of repeated values and missing values in each dataset. This report serves as a useful quality check of the individual datasets.

To aid comparisons between drugs or genes, the data is *z*-score normalized. *Z*-scores can be calculated by drug or gene or across the entire dataset. The latter method can aid the absolute comparison of drugs or genes, but this method must be used cautiously because concentration units can differ between drugs. Specific cell lines can be excluded from the analysis.

Hierarchical clustering is performed on the *z*-scores columnwise, i.e. to cluster gene or drug patterns across the cell lines. The cell lines are grouped by tissue of origin, and the order of cell lines is fixed according to the CellMiner standard (Supplementary Table S1). Three distance choices are implemented: Euclidean, Manhattan and the Pearson correlation distance. The Pearson correlation distance is a good metric to explore trends in drug activity or gene expression across the NCI-60 panel. Euclidean and Manhattan distances can reveal quantitative differences.

The columns of the *z*-score matrix can be reordered based on the hierarchical clustering, and a heatmap is generated with breakpoints determined by the deciles of the *z*-score matrix, or set *z*-scores of 0, ± 1.65, ± 1.96, ± 2.57, ± 3.30, ± 5, grouping data into bins with *P*-values of > 0.1, ≤ 0.05, ≤ 0.01 and ≤ 0.001 (*P*-values are not adjusted for multiple comparisons). The upper and lower breakpoints are extended to cover the range of the dataset.

To help researchers assess the hiearchical clustering, a silhouette analysis is performed (Rousseeuw, 1987). Silhouette values can range from -1 to 1. A value of 1 indicates that a sample is perfectly placed in a cluster, a value of 0 indicates that a sample falls between two clusters and negative values suggest that a sample is placed in the wrong cluster. The average silhouette value for a cluster indicates how tight samples group within that cluster. Similarly, the mean silhouette for a group of clusters can be used to assess the overall quality of a clustering analysis. CellMiner Companion produces a plot that shows the mean silhouette values from $k = 2$ to $k = n$ clusters, where n is the number of input files. Researchers are encouraged to choose k so that the mean silhouette value is maximized. When researchers choose a specific value for k, a table of silhouette values for each sample and each cluster is produced. This helps researchers identify tight clusters as well as possible outliers in their data.

3 Example

Cancer patients can exhibit multidrug resistance (MDR) due to the overexpression of genes encoding transporters that can pump drugs out of cells, e.g. the resistance of cells to Doxorubicin by the overexpression of ABCB1 (Reinhold *et al.*, 2012; Szakacs *et al.*, 2006). CellMiner and CellMiner Companion were used to find other possible instances of MDR in the NCI-60 data.

The gene expression pattern for 21 transporter genes was retrieved from CellMiner and visualized with CellMiner Companion (Supplementary Fig. S1 and Supplementary Table S2). Five genes with distinct patterns of overexpression across the NCI-60 cell lines were selected for further analysis (Fig. 1A and Supplementary Table S3). For each of these genes, there were one or two cell lines that showed strong evidence of overexpression, with *z*-scores ≥ 3.30. These seven cell line—gene combinations are highlighted in Figure 1A. The five genes were submitted to CellMiner for Pattern Comparisons.

For Pattern Comparisons between gene expression and drug sensitivity, a significant negative Pearson correlation suggests that, in the simplest model, a cell line has become resistant to the drug due to overexpression of a gene. Conversely, a significant positive Pearson correlation suggests that a cell line has become sensitive to the drug. To find putative instances of MDR, drugs were selected from the Pattern Comparison data based on the following criteria: (i) all FDA-approved drugs with a Pearson correlation ≤ -0.334, (ii) the twenty drugs with the most negative Pearson correlation ≤ -0.334 and (iii) for drugs with multiple National Service Center (NSC) numbers and similar Pearson corrleations, all NSC numbers were included for evaluation with CellMiner Companion. Collectively, 140 drugs met these criteria (Supplementary Table S4).

The drug sensitivity data was retrieved from CellMiner and visualized with CellMiner Companion (Supplementary Figs S2–S6). For each gene, three representative drugs were chosen using the data summary produced by CellMiner Companion for quality control. The pattern of activity for the resulting fifteen drugs was then analyzed with CellMiner Companion (Fig. 1B and Supplementary Fig. S7). Not surprisingly, the maximum silhouette value of 0.32 occurred with five clusters, and the mean silhouette values for the individual clusters ranged from 0.19 to 0.57 (Supplementary Table S5). With the exception of drug 690 757, the drug sensitivity patterns clustered according to the expression patterns of the five underlying query genes.

Importantly, the seven cell line—gene combinations highlighted in Figure 1A now reveal 21 cell line—drug sensitivities (Fig. 1B). In nineteen instances, $z \leq -1.65$ corresponding to an unadjusted p-value ≤ 0.05 (Supplementary Table S6). In fourteen instances, $z \leq -3.39$ corresponding to an unadjusted $P \leq 0.001$. This is strong evidence to support that overexpression of these transporter genes can lead to MDR. However, MDR is a complicated process (Szakacs, 2006), so it is not surprising that there is not a perfect correlation between gene expression and drug sensitivity. Furthermore, careful experimental validation would be required to confirm that the observed MDR is caused by the overexpression of these specific genes.

This example illustrates how researchers can use CellMiner and CellMiner Companion to explore NCI-60 data. In the example, 141 datasets, 21 genes and 120 drugs, were visualized at various steps to arrive at Figure 1A and B. Integrating these datasets from the original 141 CellMiner Excel files is beyond the capability of many researchers, but quite painless with CellMiner Companion, making it a useful tool for researchers to confirm, or generate, hypotheses related to gene expression and drug sensitivity. The tool also has the capacity to visualize microRNA expression data (Supplemental Fig. S8). In the future, we intend to develop methods to enable users to visualize the genetic variation and DNA copy number data for the NCI-60 cell lines. For those with programming experience, the underlying code is easily adaptable to other similar datasets, e.g. the Broad-Novartis Cancer Cell Line Encyclopedia.

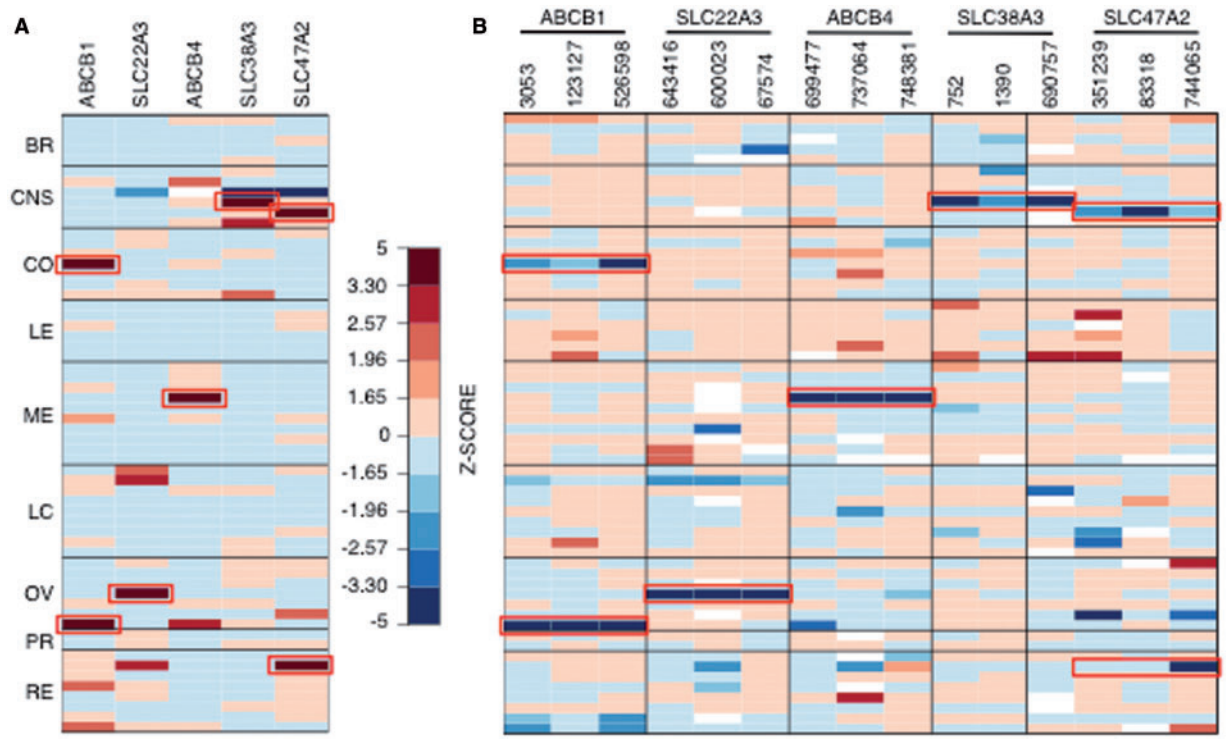

Fig. 1. Overexpression of transporter genes correlates with resistance to specific drugs. (**A**) Expression pattern for selected genes that encode transporters. Cell lines are grouped according to tissue of origin, BR—breast, CNS—central nervous system, CO—colon, LE—leukemia, ME—melanoma, LC—non-small cell lung, OV—ovarian, PR—prostate, RE—renal. The order of the cell lines is fixed to be consistent with the CellMiner standard (see Supplementary Table S1). Cell lines with strong evidence of gene overexpression are highlighted by boxes ($z \geq 3.30$, $P \leq 0.001$). Specific cell lines and *z*-scores can be determined from Supplementary Table S3. Missing values are colored white. (**B**) Pattern of sensitivity for 15 drugs with negative correlations to the gene expression patterns. Data were clustered using drug-specific *z*-scores and the Pearson correlation distance. NSC numbers are shown for the drugs. Vertical gridlines delineate five clusters that reflect the overexpression pattern of the genes in A. The dendrogram and silhouette data for the clustering are found in Supplementary Figure S7 and Supplementary Table S5, respectively. *Z*-scores are in Supplementary Table S6. Cell lines highlighted in panel A now highlight 21 cell line to drug combinations. Doxorubicin (NSC# 123127) sensitivity attributed to the overexpression of ABCB1 in cell line OV:NCI_ADR_RES is among the highlighted values (Color version of this figure is available at *Bioinformatics* online.)

Acknowledgements

The authors thank Ann Kirchmaier and Dave Zwicky for helpful comments on both the manuscript and CellMiner Companion.

Conflict of Interest: none declared.

References

Abaan,O.D. *et al.* (2013) The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.*, **73**, 4372–4382.

Chang,W. *et al.* (2016) shiny: Web Application Framework for R. R package version 0.13.0. http://CRAN.R-project.org/package=shiny.

Luna,A. *et al.* (2015) rcellminer: Exploring molecular profiles and drug response of the NCI-60 cell lines in R. *Bioinformatics.*, **32**, 1272–1274.

R Core Team. (2015) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Reinhold,W.C. *et al.* (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.*, **72**, 3499–3511.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*, **20**, 53–65.,

Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–818.

Szakacs,G. *et al.* (2006) Targeting multidrug resistance in cancer. *Nat. Rev. Drug Discov.*, **5**, 219–234.