종양 내 프로모터 DNA 메틸화 이질성과 약물 반응성의 관계 분석을 위한 웹 기반 탐색적 데이터마이닝 시스템

Web-based Exploratory Data Mining System for Analyzing the Gene-level Relationship between Intratumoral Heterogeneity of Promoter DNA Methylation and Drug Response

Tae Hoon Kweon[O1], Bonil Koo[2,3], Sungjoon Park[4], Thibaud Southiratn[1], Sun Kim[1,2,3,5]

[1]Department of Computer Science and Engineering, Seoul National University, [2]Interdisciplinary Program in Bioinformatics Seoul National University, [3]AIGENDRUG Co., Ltd., [4]Bioinformatics Institute, Seoul National University,[5]Interdisciplinary Program in Artificial Intelligence, Seoul National University

rnjsxogns25@snu.ac.kr, bikoo95@snu.ac.kr, stj@snu.ac.kr, tsouthiratn@snu.ac.kr, sunkim.bioinfo@snu.ac.kr

ABSTRACT

Understanding the complex relationship between genetic data and drug response is critical for advancing precision medicine. Selecting a subset of genes from genetic data is required due to the high cost of genome-wide approach in clinical practice. However, integrating multiple high dimensional data is a computationally difficult problem due to the vast search space. In this study, we present a web-based data mining system for exploring the relationship between gene-level promoter DNA methylation heterogeneity (mITH) and drug response. Leveraging a dataset of 928 cell lines, 545 drugs, and 679 cancer-related genes, our system allows users to identify gene-drug pairs with significant correlations between promoter mITH and drug response. Through the system, we identified that the local pairwise methylation discordance (LPMD) value of the promoter region of *CCND1* showed a significant correlation with decitabine response, which is a hypomethylating agent. This observation is supported by the literature and we propose promoter LPMD value of *CCND1* as a predictive marker for decitabine response. This system serves as a valuable tool for hypothesis generation and biomarker discovery, advancing our understanding of drug response mechanisms. The system is available at http://biohealth.snu.ac.kr/software/mITHDrugViz/.

1. Introduction

1.1 Motivation

Understanding drug response is a complex yet crucial task for stratifying patients when considering different medications in precision medicine. Genetic data is used to explain drug response. Among genetic information, DNA methylation at promoter regions plays vital roles in regulating gene expression as a significant epigenetic mechanism, affecting drug responses. An existing work examined relationship between intratumor heterogeneity of DNA methylation (mITH) and drug response [1].

1.2 Challenges

The existing work analyzed drug response with genome-wide mITH values [1], but genome-wide approach is costly in clinical practice. Accordingly, selecting a set of small number of genes is required. Genetic and epigenetic profiles are high dimensional data. In human, the number of protein-coding genes is about 20,000. Assume we have a drug to investigate. When considering relationship between the response of the drug and protein-coding genes at gene-level, about 2^{20000} cases should be considered, which is a very large search space. In other words, it is a computationally difficult problem, because integrating multiple high-dimensional data to understand drug response presents a formidable challenge due to the vast search space.

1.3 Approach

We provide a web-based exploratory data mining system where a user can explore relationship between each gene's promoter mITH value and drug response to identify putative gene-level biomarkers.

The system provides the user with options to choose from for three variables:

- 545 drugs are available from Cancer Therapeutics Response Portal (CTRP) v2.
- For each drug, 679 cancer-related genes from Catalogue Of Somatic Mutations In Cancer (COSMIC) are available.
- For each 5 mITH metric, the gene-drug pair mentioned above can mine the relationship between promoter mITH values and drug response values in up to maximum 928 cell lines.

Through the system, we aim to facilitate the discovery of gene-drug pairs with significant correlations, serving not only as a powerful tool for hypothesis generation and biomarker identification but also advancing our comprehension of drug response mechanisms.

2. Materials and Methods

2.1 Data Preprocessing

For our study, we utilized a following comprehensive set of materials:

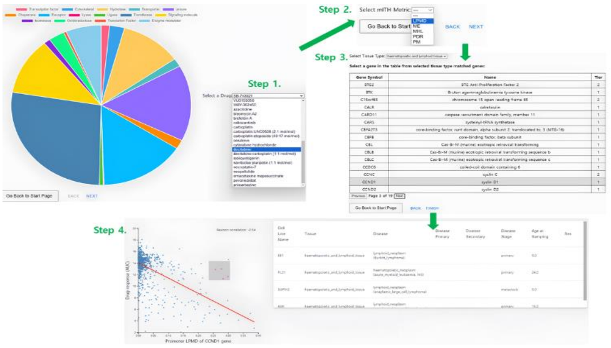
- Reduced representation bisulfite sequencing (RRBS) data from 928 cell lines: Cancer Cell Line Encyclopedia (CCLE)
- Drug response data for 545 drugs (of 361,908 drug-cell line pairs): CTRP v2
- 679 cancer-related genes: COSMIC

RRBS data for cell lines were mapped to hg38 reference genome using Bismark. Transcripton start sites (TSSs) were obtained GENECODE v38 and the promoter region for each gene is defined as 2kbp centered at the TSS. Subsequently, mITH values of each gene's promoter region were calculated using Metheor [2].

2.2 System description and operation guide

We used *D3.js* and *React* to allow users to explore interactive web visualization interface in different variable settings for drug, mITH metric, and gene.

Our system provides in total of four operation steps to explore relationship between drug response and gene-level promoter mITH. User should first press "Try Me" button from the manual page. The details of four operation steps follow as Figure 1:



Select a gene in the table from selected tissue type matched genes:

Gene Symbol	Name	Tier
BTG2	BTG Anti-Proliferation Factor 2	2
BTK	Bruton agammaglobulinemia tyrosine kinase	1
C15orf65	chromosome 15 open reading frame 65	2
CALR	calreticulin	1
CARD11	caspase recruitment domain family, member 11	1
CARS	cysteinyl-tRNA synthetase	1
CBFA2T3	core-binding factor, runt domain, alpha subunit 2; translocated to, 3 (MTG-16)	1
CBFB	core-binding factor, beta subunit	1
CBL	Cas-Br-M (murine) ecotropic retroviral transforming	1
CBLB	Cas-Br-M (murine) ecotropic retroviral transforming sequence b	1
CBLC	Cas-Br-M (murine) ecotropic retroviral transforming sequence c	1
CCDC6	coiled-coil domain containing 6	1
CCNC	cyclin C	2
CCND1	cyclin D1	1
CCND2	cyclin D2	1

Figure 1 System Operation Steps

- Step 1 (Drug selection): User can select among 545 drugs from using a pie chart interface that categorizes drugs into 15 types. Upon selecting a drug type, the corresponding filtered drugs become available.
- Step 2 (mITH metric selection): Users select one of five provided mITH metrics from a dropdown list interface.
- Step 3 (Tissue type and gene selection): Users first select a tissue type of interest from a dropdown list interface to filter the available cancer-related genes (679 genes from COSMIC). Then, users choose a gene from the filtered gene table interface.
- Step 4 (Drug response (AUC) vs promoter mITH value): Based on user selections in Steps 1 to 3, a corresponding scatter plot is generated. Users can interact with the plot, dragging elements to view cell-line details displayed in a table beside the plot.

Users have the flexibility to backtrack (using the "Back" button) to adjust drug, mITH metric, or gene selections within Steps 1 to 3 while maintaining continuity with the generated scatter plot. Incorporating a backtrack feature is essential as it enhances data analysis by allowing users to test hypotheses with different settings for drugs, mITH metrics, or genes. This capability promotes an iterative approach to exploration, enabling users to refine their selections based on insights gained from the visualization output. We further used this system to explore and analyze the complex relationship between gene-level promoter mITH and drug response.

3. Results

3.1 Comprehensive exploration of relationship between drug response and gene-level promoter mITH

Among mITH metrics the system provides, local pairwise methylation discordance (LPMD) was selected to investigate the relationship between gene-level promoter LPMD values and drug response. Our decision was motivated by the statistically significant correlation observed between genome-wide promoter LPMD values and drug response outcomes as illustrated in Figure 2.

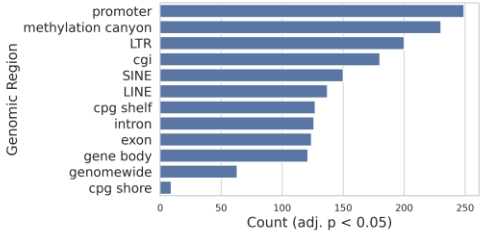


Figure 2 Correlation between 12 genome-wide genomic region LPMD values and drug response

Among 370k gene-drug pairs, *CCND1* gene had the greatest number of drugs showing statistically significant correlations with the promoter LPMD value as shown in Figure 3. For *CCND1* gene, our analysis identified top 10 drugs that correlates with drug response and LPMD value. These include four hypomethylating agents (HMA) such as decitabine and clofarabine. The Pearson correlation coefficients of each drug follow: decitabine: -0.54, decitabine+carboplatin (1:1): -0.51, decitabine+carboplatin (1:1): -0.51, decitabine+navitoclax (2:1): -0.45, and clofarabine: -0.45. The negative correlation coefficients indicate that higher LPMD values (mITH level) are associated with better drug response.

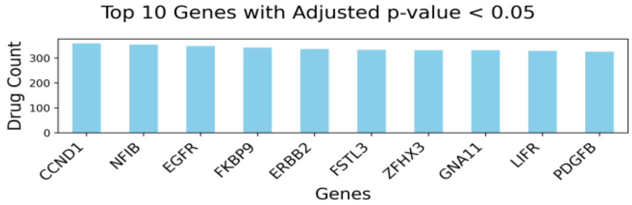


Figure 3 Top 10 genes with significant correlations

3.2 Insights from drug response and gene-level promoter LPMD Analysis

For our investigation into the relationship between drug response and gene-level promoter LPMD, we implemented the following experiment setup:

- Among 679 COSMICS genes, we obtained top 10 genes that correlates with decitabine drug response and LPMD value. Pearson correlation ranges -0.54 to -0.42.

The observed relationship between the top-ranked *CCND1* and decitabine is supported by literatures. In the presence of decitabine, up-regulation of *CCND1* expression was observed. [3]. Gene expression level is controlled by DNA methylation. Decitabine is a HMA, so it decreases methylation, consequently increasing expression levels. Up-regulation of CCND1 is associated with longer survival time [4]. Utilizing HMAs like decitabine, which decrease high ITH levels, can improve survival rate. Given these findings, promoter LPMD value of *CCND1* can be utilized as a predictive marker of drug response of decitabine.

4. Discussion and Conclusion

We developed an exploratory data mining system that integrates gene-level promoter mITH and drug response insights to explain the complex relationship between genetic variations, epigenetic factors, and drug response in cancer. As we showed promoter mITH values can be identified as predictive markers using the system, our study underscores the potential for personalized treatment strategies based on molecular signatures. Moving forward, this system serves as a powerful tool for hypothesis generation and biomarker discovery, advancing our understanding of drug response mechanisms and guiding precision oncology initiatives.

Reference

[1] Dohoon Lee et al., "AMLs harboring DNMT3A-destabilizing variants show increased intratumor DNA methylation heterogeneity at bivalent chromatin domains," *bioRxiv*, 2023.

[2] Dohoon Lee et al., "Metheor: Ultrafast DNA methylation heterogeneity calculation from bisulfite read alignments," *PLOS Computational Biology* 19.3, e1010946, 2023.

[3] Alicja Pawlak et al., "The contrasting delayed effects of transient exposure of colorectal cancer cells to decitabine or azacitidine," *Cancers* 14.6, 1530, 2022.

[4] Yijun Liu et al., "Upregulation of cyclin D1 can act as an independent prognostic marker for longer survival time in human nasopharyngeal carcinoma," *Journal of Clinical Laboratory Analysis* 34.8, e23298, 2020.

* This research was supported by the Bio & Medical Technology Development Program of the NRF funded by the Ministry of Science & ICT (NRF-2022M3E5F3085677, RS-2023-00257479, NRF-2019M3E5D3073375) and IITP grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]. This work was funded by AIGENDRUG CO., LTD. The ICT at Seoul National University provides research facilities for this study.