OXFORD

Databases and ontologies

GSCALite: a web server for gene set cancer analysis

Chun-Jie Liu[1,†], Fei-Fei Hu[1,†], Meng-Xuan Xia[1], Leng Han[2], Qiong Zhang[1],* and An-Yuan Guo[1],*

[1]Department of Bioinformatics and Systems Biology, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, People's Republic of China and [2]Department of Biochemistry and Molecular Biology, McGovern Medical School at The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Abstract

Summary: The availability of cancer genomic data makes it possible to analyze genes related to cancer. Cancer is usually the result of a set of genes and the signal of a single gene could be covered by background noise. Here, we present a web server named Gene Set Cancer Analysis (GSCALite) to analyze a set of genes in cancers with the following functional modules. (i) Differential expression in tumor versus normal, and the survival analysis; (ii) Genomic variations and their survival analysis; (iii) Gene expression associated cancer pathway activity; (iv) miRNA regulatory network for genes; (v) Drug sensitivity for genes; (vi) Normal tissue expression and eQTL for genes. GSCALite is a user-friendly web server for dynamic analysis and visualization of gene set in cancer and drug sensitivity correlation, which will be of broad utilities to cancer researchers.

Availability and implementation: GSCALite is available on http://bioinfo.life.hust.edu.cn/web/GSCALite/.

Contact: guoay@hust.edu.cn or zhangqiong@hust.edu.cn

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

1 Introduction

Next generation sequencing (NGS) technology has emerged as a powerful method for cancer genomics analysis (Ding *et al.*, 2014). The Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013), Genotype-Tissue Expression (GTEx) (GTEx Consortium, 2015) and other projects have generated a large amount of complex, multi-omics data for cancer and normal samples. These publicly available datasets provide unprecedented opportunities to understand cancer causal genes and mechanism, find candidate drug targets, and screen genes associated with phenotypes. Recently, a few excellent web servers such as cBioPortal focusing on the genomic variations based on multi-omics (Cerami1 *et al.*, 2012), GEPIA (Tang *et al.*, 2017) and Oncomine (Rhodes *et al.*, 2007) providing analysis for single

gene expression and survival. However, cancer initiation, progression and metastasis are inclined to the result of mutation and/or expression alterations of a set of genes or pathways (Harvey *et al.*, 2013). Thus, performing gene set association analysis with big data of cancer multi-omics and drug sensitivity is imperative and very useful for cancer research. Therefore, we developed an interactive web-based application named GSCALite for Gene Set Cancer Analysis to analyze and visualize the expression/variation/correlation of a gene set in cancers with flexible manner. GSCALite offers analyses including gene differential expression, overall survival, single nucleotide variation, copy number variation, methylation, pathway activity, miRNA regulation, normal tissue expression and drug sensitivity. GSCALite provided various publication-ready figures

and tables for users and the workflow was used in our recent paper (Gong *et al.*, 2017). In brief, we integrated big multi-omics and drug data to provide all-in-one analysis for a set of genes in cancers.

2 Methods and functions

The user-interface and back-end of GSCALite were written in Shiny. GSCALite consists of analytic modules for data from three major sources including multi-omics data from TCGA 11 160 samples across 33 cancer types (TCGA Cancer), 746 drug data from Genomics of Drug Sensitivity in Cancer (GDSC) (Yang *et al.*, 2013) and Cancer Therapeutics Response Portal (CTRP) (Basu *et al.*, 2013) (Drug Sensitivity), and normal tissue expression data of 11 688 samples from GTEx (GTEx Normal Tissue). We used R scripts and packages (ggplot2, visNetwork, survival and maftools) to generate figures and tables (details refer to the web site help pages). Analysis results are returned to the web page and can be downloaded in PDF, PNG, EPS, TXT as well as HTML formats. The workflow and typical output schema are shown in Supplementary Figure S1. Detailed functions and operations for each module are described below.

2.1 Gene set based multi-omics cancer analysis

GSCALite provides the following six analysis modules for a gene set based on TCGA multi-omics cancer data:

i. mRNA Expression module calculates the gene set differential expression between tumor and paired normal samples, the impact of gene expression to overall survival and expression difference between subtypes in each selected cancer type.

ii. Single Nucleotide Variation module uses maftools (Mayakonda and Koeffler, 2016) to present the SNV frequency and variant types of the gene set in selected cancer types. The effects of mutations to overall survival are given by means of the log-rank test which facilitate to evaluate the prognosis of the gene set mutations.

iii. On Copy Number Variation module, the statistics of heterozygous and homozygous CNV of each cancer type are displayed as pie chat for gene set, and Pearson correlation is performed between gene expression and CNV of each gene in each cancer to help to analyze the gene expression significantly affected by CNV.

iv. Methylation module explores the differential methylation between tumor and paired normal samples, the correlation of methylation and expression, and the survival affected by methylation level for selected cancer types.

v. Pathway Activity module presents the correlation of genes expression with pathway activity groups (activation and inhibition) that defined by pathway scores (Akbani *et al.*, 2014).

vi. For miRNA regulations, miRNA Network module combines miRNA targeting data from verified target databases and prediction methods as our previous studies (Zhang *et al.*, 2015, 2016), and the negative correlation with gene expressions to explore the miRNA-gene regulatory network for gene set in all cancer types.

2.2 The analysis of drug sensitivity and resistance to genes

Genomic aberrations influence clinical responses to treatment and are potential biomarkers for drug screening. Drug sensitivity and gene expression profiling data of cancer cell lines in GDSC and CTRP were integrated into GSCALite. The expression of each gene in the gene set was performed by Spearman correlation analysis with the small molecule/drug sensitivity (IC50). Correlations with false discovery rate (FDR < 0.05) were filtered as significant ones.

2.3 Expression profiling and eQTL in normal tissues

GSCALite provides GTEx Normal Tissue module for gene set tissue specificity analysis. This analysis offers a comprehensive display of expression profiling and eQTL information for gene set in selected normal tissues. After analysis of this module, GSCALite provides a heatmap plot for selected tissues with expression value of each gene normalized by the median.

3 Discussion

The GSCALite provides foundational tools and workflows in an all-in-one platform for cancer genomics analysis for a set of genes. GSCALite is a time-saving and intuitive tool for unleashing the value of the cancer genomics big data which enables experimental biologists without any computational programming skills to test hypothesis. It is based on gene set analysis with multi-omics data which complements the analysis with mRNA expression alone or single gene analysis. We will maintain the GSCALite web server for at least 5 years and update it with cancer genomics data increasing and new methods development. We anticipate GSCALite to help cancer research community and aid discovery of cancer pathways and drugs.

References

Akbani,R. *et al.* (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.*, 5, 3887.

Basu,A. *et al.* (2013) An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154, 1151–1161.

Cerami1,E. (2012) The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, 2, 401–404.

Ding,L. *et al.* (2014) Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.*, 15, 556–570.

Gong,J. *et al.* (2017) A pan-cancer analysis of the expression and clinical relevance of small nucleolar RNAs in human cancer. *Cell Rep.*, 21, 1968–1981.

GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348, 648–660.

Harvey,K.F. *et al.* (2013) The Hippo pathway and human cancer. *Nat. Rev. Cancer*, 13, 246–257.

Mayakonda,A. and Koeffler,H.P. (2016) Maftools: efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. bioRxiv, 052662.

Rhodes,D.R. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9, 166–180.

Tang,Z. *et al.* (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.*, 45, W98–W102.

Weinstein,J.N. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, 45, 1113–1120.

Yang,W. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41, D955–D961.

Zhang,H.-M. *et al.* (2016) miR-146b-5p within BCR-ABL1–positive microvesicles promotes leukemic transformation of hematopoietic cells. *Cancer Res.*, 76, 2901–2911.

Zhang,H.-M. *et al.* (2015) Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief. Bioinform.*, 16, 45–58.