Data in Brief 11 (2017) 131–135

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Fibronectin and androgen receptor expression data in prostate cancer obtained from a RNA-sequencing bioinformatics analysis

Dibash K. Das [a,b,c], Thahmina Ali [a], Konstantinos Krampis [a,d], Olorunseun O. Ogunwobi [a,b,c,*]

[a] Department of Biological Sciences, Hunter College of The City University of New York, New York, NY 10065, USA
[b] The Graduate Center Departments of Biology and Biochemistry, The City University of New York, New York, NY 10016, USA
[c] Department of Medicine, Weill Cornell Medicine, Cornell University, New York, NY 10065, USA
[d] Department of Physiology and Biophysics, Institute for Computational Biomedicine, Weill Cornell Medicine, Cornell University, New York, NY 10065, USA

ARTICLE INFO

Article history:
Received 7 October 2016
Received in revised form
30 December 2016
Accepted 27 January 2017
Available online 3 February 2017

Keywords:
Fibronectin
Androgen receptor

ABSTRACT

Prostate cancer is the second most commonly diagnosed male cancer in the world. The molecular mechanisms underlying its development and progression are still unclear. Here we show analysis of a prostate cancer RNA-sequencing dataset that was originally generated by Ren et al. [3] from the prostate tumor and adjacent normal tissues of 14 patients. The data presented here was analyzed using our RNA-sequencing bioinformatics analysis pipeline implemented on the bioinformatics web platform, Galaxy. The relative expression of fibronectin (FN1) and the androgen receptor (AR) were calculated in fragments per kilobase of transcript per million mapped reads, and represented in FPKM unit. A subanalysis is also shown for data from three patients, that includes the relative expression of FN1 and AR and their fold change. For interpretation and discussion, please refer to the article, "miR-1207-3p regulates the androgen receptor in prostate cancer via FNDC1/fibronectin" [1] by Das et al.

DOI of original article: http://dx.doi.org/10.1016/j.yexcr.2016.09.021
* Correspondence to: Belfer Research Building, 413 E 69th Street, Room 426, New York, NY 10021, USA.
E-mail address: ogunwobi@genectr.hunter.cuny.edu (O.O. Ogunwobi).

http://dx.doi.org/10.1016/j.dib.2017.01.016

Specifications Table

Subject area	*Biology*
More specific subject area	*Bioinformatics*
Type of data	*Paired-end RNA-sequencing data (fastq format)*
How data was acquired	*Array Express database of the EMBL European Bioinformatics Institute*
Data format	*Pre-processed RNA-sequencing data*
Experimental factors	*The primary study performed by Ren et al. applied preprocessing steps to the RNA-sequencing reads, which were, removal of sequencing adaptors and low-quality reads with specific data preprocessing filters.*
Experimental features	*Prostate tumor and adjacent normal tissues from 14 patients were sampled for RNA sequencing and sequencing libraries were constructed using the Illumina kit following the manufacturer's standard protocol, which was diluted to 2.5 pM for sequencing on a single lane of an Illumina HiSeq2000 flowcell.*
Data source location	*Prostate Cancer Discovery cohort at Shanghai Changhai Hospital, Shanghai, China*
Data accessibility	*Data available in the article and at:* https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-567/

Value of the data

- Performing a Bioinformatics RNA-sequencing analysis on published transcriptome data allows exploration and opportunities for discovery of new or not previously known biological implications.
- Allowing reproducibility of the analysis by using the method of the automated RNA sequencing pipeline on the Galaxy platform to analyze the same data and reusability of the pipeline to analyze other cancer transcriptome data.
- Promoting transparency of the analysis by allowing the data and methods used in the analysis accessible on databases and platforms.

1. Data

(Fig. 1).

2. Experimental design, materials and methods

We designed and implemented a comprehensive, standardized, and scalable RNA-sequencing bioinformatics analysis pipeline as a workflow on the Galaxy platform [2] (http://galaxy.hunter.cuny.edu:8080/u/bioitcore/w/ted-transcriptome-data-analysis) to analyze prostate cancer RNA-sequencing datasets from the Array Express archive of the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/arrayexpress/ experiments/E-MTAB-567/). As described in the primary study by Ren et al., the samples comprised poly-A containing RNA sequencing paired-end reads and replicates from fourteen prostate cancer patients [3]. The poly(A) random primed containing RNA were sequenced using Illumina HiSeq 2000 at a read length of 200-250nt producing on average 400 million reads for each library. The workflow requires eight input read files, one file of the human reference genome (UCSC hg19), as well as one file of the gene annotations of the reference genome. The workflow in total performs forty-four steps, using thirteen bioinformatics tools and requires approximately 84 h on a 4 core processor server, with four stages: