Data in Brief 15 (2017) 933–940

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Application of bi-clustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials

Andrew Williams *, Sabina Halappanavar

Environmental Health Science and Research Bureau, Health Canada, Ottawa, Ontario, Canada K1A 0K9

ARTICLE INFO

Article history:
Received 19 September 2017
Accepted 23 October 2017
Available online 26 October 2017

Keywords:
Nanomaterials
Toxicogenomics
Predictive toxicology
Bi-clustering

ABSTRACT

This article contains data related to the research article '*Application of bi-clustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials*' (Williams and Halappanavar, 2015) [1]. The presence of diverse types of nanomaterials (NMs) in commerce has grown significantly in the past decade and as a result, human exposure to these materials in the environment is inevitable. The traditional toxicity testing approaches that are reliant on animals are both time- and cost- intensive; employing which, it is not possible to complete the challenging task of safety assessment of NMs currently on the market in a timely manner. Thus, there is an urgent need for comprehensive understanding of the biological behavior of NMs, and efficient toxicity screening tools that will enable the development of predictive toxicology paradigms suited to rapidly assessing the human health impacts of exposure to NMs. In an effort to predict the long term health impacts of acute exposure to NMs, in Williams and Halappanavar (2015) [1], we applied bi-clustering and gene set enrichment analysis methods to derive essential features of altered lung transcriptome following exposure to NMs that are associated with lung-specific diseases. Several datasets from public microarray repositories describing pulmonary diseases in mouse models following exposure to a variety of substances were examined and functionally related bi-clusters showing similar gene expression profiles were identified. The identified bi-clusters were then used to conduct a gene set enrichment analysis on lung gene expression profiles derived from mice exposed to nano-titanium dioxide, carbon black or carbon nanotubes (nano-TiO2, CB and CNTs) to determine the disease significance of these

* Corresponding author.
E-mail address: Andrew.Williams@canada.ca (A. Williams).

https://doi.org/10.1016/j.dib.2017.10.060

data-driven gene sets. The results of the analysis correctly identified all NMs to be inflammogenic, and only CB and CNTs as potentially fibrogenic. Here, we elaborate on the details of the statistical methods and algorithms used to derive the disease relevant gene signatures. These details will enable other investigators to use the gene signature in future Gene Set Enrichment Analysis studies involving NMs or as features for clustering and classifying NMs of diverse properties.

Specifications Table

Organism/cell line/ tissue	*Mus Musculus/Lung*
Sequencer or array type	*Agilent-028005 SurePrint G3 Mouse GE 8×60K Microarray*
Data format	*Raw: TXT files; normalized data: TXT files*
Experimental factors	*Exposures to a variety of nanomaterials (nano-titanium dioxide, carbon black, carbon nanotubes)*
Experimental features	*Bi-cluster analysis on publically available data obtained from Gene Expression Omnibus (GEO) describing specific lung diseases was conducted to identify functionally related gene sets. DAVID analysis was conducted on each of the gene sets from this analysis to identify functional representation of each gene set. Gene set enrichment analysis was then conducted on nine toxicogenomic gene expression studies examining the toxicity induced by a variety of nanomaterials to determine the disease significance of the altered gene expression profiles following exposure to NMs.*
Sample source location	*Ottawa, Ontario, Canada*
Data accessibility	*National Centre for Biotechnology Information, GEO database Accession:* GSE35193, GSE41041, GSE47000, GSE60801, GSE61366

Value of the data

- The results enabled deeper mechanistic understanding of NM-induced lung toxicity.
- The data enabled the development of a database with toxicity fingerprints that are specific to lung diseases.
- Using the statistical tools and algorithms established, it may be possible to predict the toxicities of new NMs that have yet to undergo experimental testing.
- The data was integral in identifying new gene sets associated with lung pathology that were previously not known.
- The gene sets identified could serve as features for clustering and classifying NMs of diverse properties.

1. Data description

We anticipate that the importance of toxicogenomics studies in chemical risk assessment will continue to increase in the coming years. However, its success will depend on 1) accurate and prompt reporting of the data, 2) ensuring public availability of the datasets and 3) sharing of the training sets