Gene xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Gene

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

Research paper

Carcinogenic effects of oil dispersants: A KEGG pathway-based RNA-seq study of human airway epithelial cells

Yao-Zhong Liu [a,*], Lei Zhang [b,1], Astrid M Roy-Engel [c], Shigeki Saito [d], Joseph A Lasky [d], Guangdi Wang [e], He Wang [f]

[a] Dept. of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA
[b] Center for Genetic Epidemiology and Genomics, School of Public Health, Medical College, Soochow University, Su Zhou, China
[c] Dept. of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA
[d] Dept. of Medicine, Tulane University School of Medicine, New Orleans, LA, USA
[e] Dept. of Chemistry, RCMI Cancer Research Center, Xavier University of Louisiana, New Orleans, LA, USA
[f] Dept. of Chronic Respiratory Diseases, School of Health Sciences, University of Newcastle, Callaghan, Australia

ARTICLE INFO

Article history:
Received 16 October 2016
Accepted 14 November 2016
Available online xxxx

Keywords:
BP oil spill
RNA-seq
Airway epithelial cells
Dispersants
Oil
Carcinogenesis
GAGE
GSNCA
KEGG

ABSTRACT

The health impacts of the BP oil spill are yet to be further revealed as the toxicological effects of oil products and dispersants on human respiratory system may be latent and complex, and hence difficult to study and follow up. Here we performed RNA-seq analyses of a system of human airway epithelial cells treated with the BP crude oil and/or dispersants Corexit 9500 and Corexit 9527 that were used to help break up the oil spill. Based on the RNA-seq data, we then systemically analyzed the transcriptomic perturbations of the cells at the KEGG pathway level using two pathway-based analysis tools, GAGE (generally applicable gene set enrichment) and GSNCA (Gene Sets Net Correlations Analysis). Our results suggested a pattern of change towards carcinogenesis for the treated cells marked by upregulation of ribosomal biosynthesis (hsa03008) ($p = 1.97E-13$), protein processing (hsa04141) ($p = 4.09E-7$), Wnt signaling (hsa04310) ($p = 6.76E-3$), neurotrophin signaling (hsa04722) ($p = 7.73E-3$) and insulin signaling (hsa04910) ($p = 1.16E-2$) pathways under the dispersant Corexit 9527 treatment, as identified by GAGE analysis. Furthermore, through GSNCA analysis, we identified gene co-expression changes for several KEGG cancer pathways, including small cell lung cancer pathway (hsa05222, $p = 9.99E-5$), under various treatments of oil/dispersant, especially the mixture of oil and Corexit 9527. Overall, our results suggested carcinogenic effects of dispersants (in particular Corexit 9527) and their mixtures with the BP crude oil, and provided further support for more stringent safety precautions and regulations for operations involving long-term respiratory exposure to oil and dispersants.

1. Introduction

The year 2016 marked the sixth anniversary of the Deepwater Horizon oil spill, the largest man-made disaster in the history of petroleum industry. The sheer scale of this disaster is marked by its long duration (lasting for ~3 months), the gigantic volume of crude oil (210 million gal) spilled and dispersants (1.8 million gal) applied and the large number of workers (>50.000) involved in the cleaning operation (Hayworth and Clement, 2012; Kujawinski et al., 2011).

The full impacts of the disaster to the environment and marine and human lives have yet to be fully unveiled. Specifically, the long term health impacts of the BP oil spill to the >50.000 workers involved in the cleaning operation have not been well characterized and followed up, although limited data on other smaller scale oil spills (e.g., the Prestige oil spill) did suggest that involvement in oil spill cleaning operations may cause persistent respiratory symptoms (Zock et al., 2012), long-lasting airway oxidative stress (Rodriguez-Trigo et al., 2010), and systemic genetic effects (Laffon et al., 2006; Perez-Cadahia et al., 2007; Perez-Cadahia et al., 2008a; Perez-Cadahia et al., 2008b). In addition, the oil-dispersant mixtures may contain potentially mutagenic/carcinogenic chemicals including PAH, benzene, and benzene derivatives (Rodrigues et al., 2010; Saeed and Al-Mutairi, 1999). More importantly, chemical components in the mixtures may enhance each other to induce harmful effects synergistically. The mechanism for inhalation of hazardous substances during the oil spill was also proposed through models where inhalable aerosols that contain dispersed oil can be

Abbreviations: GAGE, generally applicable gene set enrichment; GSNCA, Gene Sets Net Correlations Analysis; WAF, water accommodated fraction; DAVID, database for annotation, visualization, and integrated discovery.

* Corresponding author at: Dept. of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, Suite 2001, New Orleans, LA 70112, USA.
E-mail address: yliu8@tulane.edu (Y.-Z. Liu).
[1] Co-first author.

http://dx.doi.org/10.1016/j.gene.2016.11.028
0378-1119/

formed on the sea surface (Ehrenhauser et al., 2014; Middlebrook et al., 2012).

To characterize the effects and mechanisms of oil spill to human lung health at the molecular level, we hypothesize that oil spill chemicals (i.e., oil, dispersant, or their mixtures) may have significant effects on respiratory cells, which can be detected at the transcriptomic level. To test this hypothesis, we performed an RNA-seq study of human airway epithelial cells treated with BP crude oil, oil dispersants (Corexit 9500 and 9527) and their mixtures (Liu et al., 2016). Through that study we identified a large number of genes differentially expressed due to the treatments, suggesting significant transcriptomic perturbations of the cells caused by the toxicological effects of the oil and oil cleaning chemicals. The findings provided a strong support to our hypothesis. Furthermore, by annotating the differentially expressed genes using DAVID analysis (Dennis et al., 2003), our study (Liu et al., 2016) suggested several key biological processes affected by the chemicals, including degradation of the cell junction, enhanced immune response, decreased local steroid biosynthesis and enhanced angiogenesis. These identified biological processes are consistent with some of the pathological features for several common lung diseases, such as COPD (Faner et al., 2013; Holtzman et al., 2014), asthma (Ribatti et al., 2009) and cystic fibrosis (Georas and Rezaee, 2014; Heijink et al., 2014; Rezaee and Georas, 2014). Therefore our study (Liu et al., 2016) not only detected the existence but also preliminarily characterized potential molecular mechanisms for the toxicological effects of oil and oil cleaning chemicals. Overall our findings (Liu et al., 2016) provided compelling evidence for the potential lung health impact of the BP oil spill on those workers involved in the cleaning operation.

One of the key limitations of our previous study (Liu et al., 2016) is that the analysis was largely performed at single gene level. Although we did perform analysis at the level of GO functional terms, the analysis was still based on the results from single gene differential expression analysis. For example, only those genes that achieved a p value of <0.05 in single gene differential expression analysis were submitted to GO analysis. Such a p value threshold, although commonly used, may be arbitrary in a genomic study as those genes that did not achieve the p value <0.05 may also contribute to the effects of biological significance. Hence annotation based only on those genes with a small p value may lose some sensitivity to capture the key signatures of the toxicological effects on the cells. Furthermore, genes often work correlatively and collaboratively in pathways and functional modules. Differential expression analysis at the single gene level as in our previous study (Liu et al., 2016) ignored such correlative relationship, which again may have missed some important transcriptomics signatures and failed to capture some key functional variations of the transcriptome.

To alleviate the problems/limitations of our previous study and take full advantage of this valuable RNA-seq dataset (Liu et al., 2016), here we performed a KEGG pathway-based analysis using the two well-developed software packages, GAGE (generally applicable gene set enrichment) (Luo et al., 2009) and GSNCA (Gene Sets Net Correlations Analysis) (Rahmatallah et al., 2014). While the findings from our new analysis agree with some of the previous study (Liu et al., 2016), the findings provided distinct new clues. A number of pathways, including ribosomal biosynthesis, protein processing, Wnt signaling, neurotrophin signaling and insulin signaling pathways, were all upregulated at the whole pathway level (mainly by the dispersant 9527). Importantly, upregulation of these pathways was all closely related with cancers, including lung cancer, in previous molecular and epidemiological studies (Mazieres et al., 2005; Poloz and Stambolic, 2015; Prakash et al., 2010; Zhou et al., 2015). Moreover, by GSNCA analysis (Rahmatallah et al., 2014), we further identified a number of cancer-related pathways, including the small cell lung cancer pathway, whose gene co-expression was changed due to the exposure to the oil/dispersant treatment, especially the mixture of dispersant 9527 and oil. Overall, our study here by analyzing the transcriptomics signals at the KEGG pathway level has revealed potential carcinogenic effects of dispersant 9527 and its mixture with crude oil. The findings provided further evidence for the health hazards of oil and oil dispersants to the respiratory system.

2. Methods

2.1. Experimental methods

The methods for generating the RNA-seq data were detailed in our previous study (Liu et al., 2016). Briefly, human airway epithelial cells (BEAS-2B cells, ATCC® CRL-9609™) were grown under six treatments using WAF (water accommodated fraction) of the following, i.e., the BP crude oil (abbreviated as "oil"), dispersant Corexit 9500 (abbreviated as "9500"), dispersant Corexit 9527 (abbreviated as "9527"), the mixture of oil + 9500 (abbreviated as "oil + 9500"), the mixture of oil + 9527 (abbreviated as "oil + 9527"), and water (abbreviated as "control"). There were three cell samples (biological replicates) for each treatment and total RNA extraction followed by RNA-seq experiments was performed on each sample. The RNA-seq experiments were performed at Omega Bioservices (Norcross, GA). The RNA-seq data was submitted to GEO under the accession number (GSE70909). Differential expression analysis was performed between each of the first five treatment groups vs. the control group.

The BP crude oil was kindly provided by The Architecture, Engineering, Consulting, Operations and Management Company (AECOM, Los Angeles, CA). This oil was obtained from the site of the Macondo well during the BP Oil Spill disaster. Commercially available Corexit EC9500A and EC9527 dispersants were kindly provided by a contract between Nalco/Exxon Energy Chemicals, L.P. (Sugar Land, TX, USA) and Tulane University (New Orleans, USA). The dispersants were provided as liquid solutions ready for use.

2.2. GAGE analysis

Following the Bioconductor workflow http://www.bioconductor.org/help/workflows/rnaseqGene/, the raw RNA-seq data (GSE70909) was analyzed to generate the raw count matrix, based on which, we used DESeq2 (Love et al., 2014) to perform differential expression analysis between a treatment (e.g., oil) and the control.

DESeq2 package requires "raw" counts of sequencing reads as the starting point for differential expression analysis (Love et al., 2014). Therefore, before submitted to the program for analysis, the count matrix was not normalized (which is explicitly required by the software) (Love et al., 2014). However, during the analysis procedures of DESeq2, normalization did occur in the modeling process, where the read count for gene i and sample j was modeled as a negative binomial distribution with mean μ_{ij} and dispersion α_j, and $\mu_{ij} = s_{ij}q_{ij}$, where q_{ij} is the raw read count and s_{ij} is a size factor that normalizes differences in sequencing depth between samples and other sources of technical biases, such as GC content and gene length (Love et al., 2014). The size factor s_{ij} was estimated with the "estimateSizeFactor()" function and the dispersion α_j estimated with the "DESeq()" function.

We ran the "DESeqDataSetFromMatrix" function, with the count matrix for all 18 samples (3 for each of the 5 treatments and 3 for the control) and the design matrix as input data, which produced an R object for the downstream differential expression analysis. Including count data for all the 18 samples (rather than only 3 samples from a treatment and the 3 control samples) is more robust for estimating parameters (such as the size factor and dispersion). For differential expression analysis to compare a certain treatment with the control, we used "contrast()" function, e.g., by defining "contrast = c ("treatment", "oil", "control")". The result file that contain p values and fold changes for each gene was generated with the "results()" function.

The log2 fold changes for all the genes from the differential expression analysis (which is a column in the result file) were submitted as

Please cite this article as: Liu, Y.-Z., et al., Carcinogenic effects of oil dispersants: A KEGG pathway-based RNA-seq study of human airway epithelial cells, Gene (2016), http://dx.doi.org/10.1016/j.gene.2016.11.028