USING THE COMPARATIVE TOXICOGENOMICS DATABASE TO UNDERSTAND THE IMPACT OF HIGH-RISK TOXICANTS ON BIOLOGICAL PATHWAYS

Samantha E. Tulenko

A technical report submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Masters of Science in Public Health in the Department of Environmental Sciences and Engineering in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Rebecca Fry

Jacqueline MacDonald Gibson

Leena Nylander-French

ABSTRACT

Samantha E. Tulenko: Using the Comparative Toxicogenomics Database to understand the impact of high-risk toxicants on biological pathways
(Under the direction of Rebecca C. Fry)

The role of environmental contaminant exposure in the development of a broad suite of human diseases has been well established. An understanding of the biological mechanisms through which these diverse and prevalent toxicants cause toxicity would be beneficial to both public health and clinical medicine. Here, we used a systematic analysis of data available from the Comparative Toxicogenomics Database to explore known toxicant-gene-disease interactions for 83 high-priority toxicants, as ranked by the Agency for Toxic Substances and Disease Research. Rankings of toxicant-pathway associations identified specific biological pathways enriched for disruption by numerous toxicants such as the Mitogen-Activated Protein Kinase (MAPK) pathway. These results present a novel approach to identify and prioritize biological pathways of particular relevance to high-priority environmental contaminants.

TABLE OF CONTENTS

LIST OF TABLES & FIGURES

INTRODUCTION

Major public health disasters with widespread impact such as those in Love Canal, New York and Times Beach, Missouri brought attention to industry's role in releasing hazardous toxicants in the environment (Beck, 1979; EPA, 2010). Pressure from these events and others resulted in the 1980 passage of the federal government's Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), also referred to as Superfund (EPA, 2013). Superfund is the current federal regulation regarding mitigation at and cleanup of hazardous waste sites in the United States. After assessing current conditions, the EPA ranks abandoned hazardous waste sites on a National Priority List (NPL).

Under the CERCLA, the Agency for Toxic Substances and Disease Registry (ATSDR) maintains a Substance Priority List for use in the cleanup process. Toxicants found at sites on the NPL are ranked based on prevalence and potential for exposure as well as the potential for causing adverse human health effects. To be included on the Substance Priority List, toxicants must be found at more than three NPL sites. Exposure potential is based jointly on the concentration of the toxicant at NPL sites as well as the number of recorded exposures to the toxicant (ATSDR, 2014). Toxicants included on the list range from metals and degradation by-products to industrial solvents and pesticides.

There is increasing information relating exposure to environmental contaminants and development of human disease from both epidemiological and toxicological studies demonstrating the substantial impacts of these toxicants on human health (Smith et al., 1992; Berk et al., 1976). However, the precise mechanisms of action by which these toxicants cause disease is often not well understood (Toscano & Oehlke, 2005). It is

recognized that interactions between the environment and genes impact biological processes in ways that can manifest in disease (Davis et al., 2011). Elucidating the mechanisms through which contaminant exposure may contribute to different states of disease development can have significant implications for prevention and treatment of chronic health problems. Biological pathways offer a mechanism for understanding how genes interact to manifest in various disease states (Li & Agarwal, 2009). Knowledge about the pathways through which toxicants act will enable targeted disease prevention strategies (Ahir, Sanders, Rager, & Fry, 2013).

In the present study, we set out to test whether high priority toxicants, many of which are located at Superfund sites, impact common biological pathways through a systematic analysis of the data available from the Comparative Toxicogenomics Database (CTD). Toxicant-gene-pathway interactions were analyzed for more than 80 high-priority toxicants. Using biological pathway information available for top-ranking toxicants on the ATSDR's Substance Priority List, computational analysis was used to compare known biological pathways associated with each of these toxicants. Trends that emerged were interpreted as a way to characterize the perturbation of biological pathways upon exposure to high-priority toxicants.

MATERIALS & METHODS

Environmental Toxicants

The present study first focused on the top 100 toxicants listed on the ATSDR 2013 Substance Priority List (http://www.atsdr.cdc.gov/spl/). Toxicants on the list are ranked by both toxicity and potential for exposure due to presence at NPL sites. These one hundred toxicants include toxic metals, pesticides, and industrial by-products, among others. The ATSDR Substance Priority List provides a name and Chemical Abstracts Society (CAS) Registry Number (RN) for each ranked toxicant (http://www.cas.org/). To confirm the toxicants included on the ATSDR list, CAS RNs were cross-referenced using SciFinder, a research tool developed by the Chemical Abstracts Society (https://scifinder.cas.org/). SciFinder provides a "Substance Identifier" tool for searching toxicants by registry number and name. CAS RNs on the Substance Priority List were considered correct if the toxicant name on the priority list was included as a synonym for the toxicant identified using the SciFinder substance identifier.

Biological Pathways of Interest

All information relating the toxicants to genes and diseases for each of the top one hundred toxicants was obtained from the CTD (http://ctdbase.org/). The CTD is a manually curated database that associates toxicants with genes, diseases, and pathways (Davis, Murphy, Rosenstein, Wiegers, & Mattingly, 2008). At the time of our analysis, the CTD included over 1,000,000 toxicant-gene interactions and over 15,000,00 gene-disease associations. P-values are assigned to the relationships based on curated

interactions between toxicants and genes in a given pathway. Pathways are included in the database if the relationship between toxicant and pathway has a p-value < 0.01.

Toxicants were identified in the CTD using a query of CAS registry numbers. Once each toxicant was identified in the database, the list of curated associated pathways was compiled. The CTD site includes pathways that have been enriched significantly among genes that interact with each toxicant. Pathways are provided with a pathway name and KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg/pathway.html) or REACT (Reactome) (http://www.reactome.org/) identifier.

Relationships between toxicants and associated pathways

Data available for each of the toxicants was compiled and organized by common pathways. For each pathway, the proportion of associated toxicants relative to all toxicants was calculated. Pathways were ranked according to the proportion of toxicants associated with each pathway, and enriched pathways were identified. After determining this pathway of interest, each toxicant associated with the pathway was queried in the CTD site by CAS RN or toxicant name. For each toxicant, the specific genes involved in the pathway were compiled from the CTD.

Hierarchical clustering analysis was performed to determine similarities of ranking based on both toxicant and pathway information. Data was clustered on both pathways and toxicants using Euclidean distance and average linkage measurements. For each toxicant, each pathway was assigned a binary value of 0 or 1 to indicate whether or not that toxicant was associated with the pathway. Toxicants were considered associated

with a pathway if they were given a p-value < 0.01 in CTD based on curated genes associations.

Correlation analysis

For an individual toxicant, the number of references was correlated with the number of pathways using the Spearman Rank Correlation test (TIBCO® Spotfire®, v5.0.0).

RESULTS

Using a combination of computational analyses, the top 100 toxicants on the ATSDR 2013 Substance Priority List were examined for their known effects on genes within biological pathways. The analyses aimed to elucidate underlying trends or similarities between the mechanisms of action of these priority toxicants found at U.S. Superfund sites. Of the first 100 priority toxicants, pathway information was only available for 83 of the toxicants (Table 1). The remaining 17 toxicants were either not included in the CTD by name or CAS registry number (11), or the pathways had not yet been curated for the toxicant (6). Significantly associated biological pathways ($p < 0.01$) were compiled for each of the 83 toxicants, resulting in 271 pathways across all toxicants (Table 1). Among these pathways, there were a total of 225 KEGG pathways identified.

Interestingly, the data highlight that some chemicals impact many pathways. The number of significantly-associated biological pathways for a toxicant ranged from one for heptachlor epoxide and ethion to 266 pathways for 2,3,7,8-tetrachlorodibenzo-p-dioxin (Table 1). For each toxicant, the proportion of associated pathways identified relative to the total number of possible pathways ranged from 0.4% to 98.2% with a median of 21.0%. When analyzed separately, a similar distribution was observed for the KEGG and REACT pathways. Heptachlor epoxide and ethion both had the lowest percentages of total KEGG pathways (0%), while both polycyclic aromatic hydrocarbons (PAHs) and 2,3,7,8-tetrachlorodibenzo-p-dioxin were associated with the highest percentage of KEGG pathways (97.8%). Three toxicants—1,2-dibromo-3-chloropropane, endosulfan sulfate, and methane—were not associated with any of the REACT pathways. PAHs and

2,3,7,8-tetrachlorodibenzo-p-dioxin were again associated with the highest proportion of REACT pathways, at 97.8% of pathways.

For further analysis, KEGG- and REACT-identified pathways were considered separately, because these are two separate classifications of identifiers and redundantly cover cellular processes. Of the 225 KEGG-identified pathways, 26 represent canonical pathways (Table 2). Considering the relevance of biological pathways in disease phenotypes, these 26 KEGG pathways were analyzed separately. Of the original 83 toxicants, 62 were associated with at least one of the more specific 26 biological pathways (Table 2).

When analyzed at the pathway level, some pathways were shown to be associated with numerous toxicants. Hierarchical clustering performed on the 26 biological pathways revealed trends between toxicants in the biological pathways they impact (Figure 1). Many interesting or expected groupings are apparent in the heat map when clustering toxicants based on perturbed pathways. PAHs, polychlorinated biphenyls (PCBs), and di-n-butyl phthalate are associated with the greatest number of pathways. Specifically, these three contaminants are associated with all 26 specific pathways (Figure 1).

The proportion of toxicants associated with each pathway was used to highlight the pathways most affected by these toxicants. Certain pathways were associated with a majority of toxicants. Of the significant KEGG-identified specific pathways, the MAPK signaling pathway was associated with the greatest number of toxicants (54/62 = 87.1%). The Notch signaling pathway was associated with the lowest number (7) of toxicants (11.3%). Other pathways associated with over 70% of the toxicants include T-cell

receptor signaling pathway (79.0%), toll-like receptor signaling pathway (77.4%), p53 signaling pathway (74.2%), neurotrophin signaling pathway (74.2%), and the nucleotide-binding oligomerization domain (NOD)-like receptor signaling pathway (71.0%). Table 2 presents all pathways and associated toxicants.

The MAPK signaling pathway was identified to have the greatest number of associated toxicants. Using gene data compiled from the CTD, genes in the MAPK pathway that were perturbed by toxicants were identified. Table 3 presents the 273 unique genes associated with at least one of the toxicants found to perturb the MAPK pathway. The genes associated with the greatest number of toxicants were *CASP3* and *JUN* (associated with 74.1% of the 54 toxicants that were associated with the MAPK pathway). Twelve of the genes in the MAPK pathway were associated with only one toxicant (1.9%) (Table 3).

Certain biological pathways were associated with numerous high-priority toxicants, which is relevant knowledge in the context of environmental mixtures. Of the 26 specific pathways, 18 pathways were associated with ≥ 50% of the toxicants (Table 2). These other pathways included those involved in metabolic disorders such as the insulin signaling pathway (associated with 62.9% of toxicants) and the adipocytokine signaling pathway (58.1%). Other pathways associated with a majority of the toxicants included pathways regulating genes that can play key roles in cancer development when perturbed. These pathways include the p53 signaling pathway (associated with 74.2% of toxicants), the Wnt signaling pathway (59.7%), and the Vascular Endothelial Growth Factor (VEGF) signaling pathway (61.2%).

Table 1 also includes the number of references curated by the CTD for each toxicant. Reflective of the breadth of existing toxicogenomics data published for each toxicant, the number of associated references ranges from 2 (endosulfan sulfate) to 11,459 (PAHs), with a median of 49 references for all toxicants. PAHs had almost 10,000 more references than the toxicant with the next highest number of references, 2,3,7,8-tetrachlorodibenzo-p-dioxin (1,537 references). A Spearman Rank Correlation test was used to evaluate the relationship between the number of references and the number of pathways identified, giving $R = 0.87$ with p-value = 5.53E-26.

DISCUSSION

As environmental contaminants are ubiquitous in the environment, there is a need for increased understanding whether and how contaminants influence disease. With this in mind, the present study aimed to identify whether the top 100 ATSDR high-priority substances target common biological pathways. *In silico* analyses querying the biological pathways associated with these substances elucidated three major findings. First, when analyzed individually, single contaminants have been shown to act on multiple biological pathways influencing genes/proteins within those pathways. Second, specific biological pathways are enriched for perturbation by numerous Superfund toxicants, for example the MAPK pathway. Third, this type of analysis is particularly suited for investigating environmental mixtures in the way that it highlights pathways and genes impacted by multiple toxicants.

An interesting finding was that individual toxicants impact multiple pathways. For example, di-n-butyl phthalate, PAHs, PCBs, and 2,3,6,7-tetrachlorodibenzo-p-dioxin were each associated with all 26 specific biological pathways. These data indicate that exposure to any single toxicant could have widespread effects throughout the body. Indeed, one disease may be linked to many different pathways while one pathway can be linked with numerous diseases (Li &Agarwal, 2009). Interfering with the activity of a pathway offers a possible mechanism for mitigating disease (Ahir et al., 2013). Therefore, understanding which pathways are impacted upon exposure to a toxicant is of potential use in developing therapies. The results of this analysis demonstrate that an individual toxicant has the potential to cause widespread biological effects by impacting numerous biological pathways.

We also found that individual pathways are perturbed by numerous toxicants. The MAPK pathway was identified as the highest ranking, as genes and proteins within this pathway were perturbed by 65.1% of the high-priority toxicants. Within the pathway, distinct genes were associated with up to 40 different toxicants (74.1%). These included genes such as RAS, a known oncogene, which was associated with 29 toxicants (53.7%) in the MAPK pathway. This putative conservation of biological mechanism is plausible when considering the function of MAPK. Through phosphorylation of substrates, kinases act as regulators of cellular activities. The MAPK pathway is activated in response to a number of different stresses, ranging from cellular exposure to DNA-damaging agents to heat shock as well as endogenous signals such as growth factors (Pearson et al., 2001). In this manner, MAPK is known to play a role in gene expression, mitosis, metabolism, apoptosis, embryogenesis and movement (Johnson & Lapadat, 2002). Ranking biological pathways by the proportion of toxicants they were associated with suggests that environmental toxicants may act through similar biological mechanisms. These data highlight the role of the MAPK pathway as a key responder.

The finding that numerous SRP toxicants are acting on common genes within common pathways is especially pertinent in the context of environmental mixtures. While the associations curated in the CTD were mostly based on studies that evaluated responses to a single contaminant, in the environment, individuals are exposed to many diverse toxicants (Carlin et al., 2013). It is imperative that toxicological research incorporates mixtures in order to better understand how these multichemical exposures impact human health. The presented findings suggest that diverse contaminants are acting

on the same biological pathways, highlighting a potential starting point for further research into the health effects of mixtures.

While this study was designed to utilize a comprehensive database of toxic substances, genes, and pathway most effectively, limitations do exist. Primarily, the CTD relies on existing published studies for toxicants and pathways. This limits the generalizability of the data when comparing across toxicants or pathways, as particular toxicants or pathways have been more extensively studied than others. PAHs have almost ten times more references in the CTD than any other toxicant ranked in the top 100 of ATSDR's Toxicant Priority List. In contrast, 15 toxicants have less than 10 curated references. It is possible that more associations may exist for toxicants and pathways that have been studied more extensively than others. This should be taken into consideration when examining which toxicants or pathways had the greatest proportions of associations. Nevertheless, the conclusions presented in this study should reflect general trends in current knowledge about these toxicants and are a springboard for future research.

In conclusion, understanding the mechanisms by which contaminants affect human health provides possible avenues for developing targeted therapies. Recognizing similarities in the way toxicants perturb cellular processes is a platform for understanding how these toxicants might interact in the body or for predicting how a similar toxicant might impact biological function. After determining that a high proportion of diverse toxicants act through the same mechanisms, it is compelling to target these specific pathways for future therapies. The results presented here provide a unique interpretation of compiled data about environmental contaminants and offer potential for future public

health initiatives. Widespread environmental contamination from diverse toxicants is well recognized in the United States; it is imperative to move forward with strategies that can improve the health of individuals who are exposed.