UNIVERSIDADE D
COIMBRA

Xavier Sá Castro Pinho

Adverse outcome pathway for benzene induced toxicity through reverse causal reasoning and network perturbation analysis

Dissertação no âmbito do Mestrado Integrado em Engenharia Biomédica, orientada pelo Doutor Shaji Krishnan, pelo Doutor Rob Stierum e pela Professora Doutora Irina Moreira e apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Dezembro de 2020

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Xavier Sá Castro Pinho

Adverse outcome pathway for benzene induced toxicity through reverse causal reasoning and network perturbation analysis

University of Coimbra
Master in Biomedical Engineering

Internship Institute
TNO
Risk Analysis for Products in Development (RAPID)
Princetonlaan 6, 3584 CB Utrecht
Daily supervisor: Dr. Shaji Krishnan
On-site supervisor: Dr. Rob Stierum

UC Supervisor
Dr. Irina Moreira
Center for Neuroscience and Cell Biology (CNC.IBILI)
irm2223@gmail.com

Coimbra, 2020

This work was developed in collaboration with:

University of Coimbra

UNIVERSIDADE D
COIMBRA

Netherlands Organisation for Applied Scientific Research

Acknowledgments

"I would like to recognize the invaluable assistance that you all provided during my study."

"Desenrasca-te. Para grandes males, água e sabão." (João César Monteiro)

Abstract

The increase and improvement in molecular profiling technologies have enabled the acquisition of large datasets consisting of measurements for many molecular entities. These datasets allow an understanding of molecular profiles of, for example, a disease, drug and compounds action, or toxicity. Furthermore, gene expression profiling experiments usually produce extensive lists of differential expressed genes that characterize the comparison between the two states in the study, such as disease versus healthy or treatment versus control. In this study two approaches are used to interpret these lists, take out relevant and reliable hypotheses and quantify biological network perturbations: Reverse Causal Reasoning (RCR) and Network Perturbation Analysis (NPA); towards exploring the full potential of these datasets. The RCR and NPA methods are implemented and tested on the transcriptome of benzene-exposed individuals to propose a hypothesis of biological processes alterations. Several proposed altered biological mechanisms are in agreement with literature evidence, meaning that this approach can be a valuable tool for understanding mechanisms associated with benzene exposure. While some of them have not been studied and false positives are a possibility, this approach indicates possible candidates, that have not been verified by the literature as potential future directions in research.

List of abbreviations

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ANLL	Acute Non-Lymphocytic Leukemia
BEL	Biological Expression Language
CML	Chronic Myeloid Leukemia
CTD	Comparative Toxicogenomics Database
DNA	Deoxyribonucleic Acid
FDR	False Discovery Rate
GO	Gene Ontology
GPI	Geometric Perturbation Index
GSA	Gene Set Analysis
GSEA	Gene Set Enrichment Analysis
HGNC	HUGO Gene Nomenclature Committee
IARC	International Agency for Research on Cancer
INDRA	Integrated Network and Dynamical Reasoning Assembler
KEGG	Kyoto Encyclopedia of Genes and Genomes
LFDR	Local False Discovery Rate
log2FC	Logarithmic Fold Change
MSigDB	Molecular Signatures Databases
NHL	Non-Hodgkin Lymphoma
NLP	Natural Language Processing
NPA	Network Perturbation Analysis
ORA	Over-Representation Analysis
PMBC	Peripheral Mono Nuclear Blood Cells
PPM	Parts Per Million
RCR	Reverse Causal Reasoning
RMA	Robust Multi-array Average
RNA	Ribonucleic Acid
SD	Standard Deviation
SIF	Simple Interaction Format

List of Figures

List of Tables

Contents

1

Introduction

1.1 Motivation

The exposure of organisms to some biologically active compounds may have potential health effects. Some of these effects take years to manifest, at a point that there is not a way to prevent disease onset. Thus the solution to fight diseases is based on prevention. In the last years, the amount of available data produced by high-throughput measurement technologies has increased. These datasets are a valuable key to an understanding of the molecular profiles of diseases and the way these compounds influence these. They disclose genome-wide modifications induced by toxic agents that can provide insight into possible mechanisms of toxicity and the inference of potential effects that have not been reflected by phenotypic changes. However, these datasets are not human-readable, so methods to filter and extract relevant information from extensive lists of differential expressed genes are necessary, in order to assess the biological impact in a qualitative and quantitative manner.

1.2 Benzene

Benzene, an aromatic hydrocarbon, is a clear, colorless, volatile, and highly flammable liquid, at room temperature. It is also a ubiquitous chemical in our environment that is known to cause serious health problems, such as leukemia (Loomis D et al., 2017).

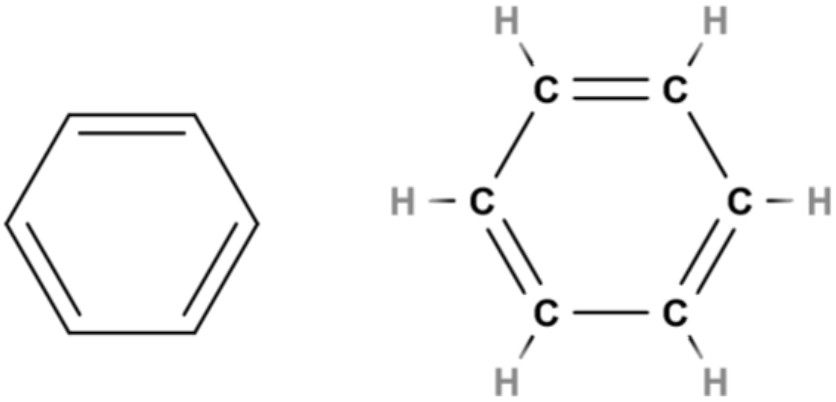

Figure 1.1: Structure of Benzene (C6H6).

The general population is widely exposed to low-level benzene from tobacco smoke, vehicle exhaust, gasoline stations, and contaminated water and food. Millions of workers are daily exposed to benzene in the manufacturing of chemicals, transport, or construction or others employed at workplaces with exposure to exhaust gases from motor vehicles (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2012). Human exposure to benzene is unavoidable and the possible adverse health effects associated with benzene chronic or acute exposure remain a matter of great concern (Snyder et al., 2012).

Since 1974, in International Agency for Research on Cancer (IARC) Monographs Volume 29 (IARC, 1974), benzene was classified as having "sufficient evidence that benzene is carcinogenic to man". Benzene occupational exposure is linked to a set of chronic diseases such as Acute Myeloid Leukemia (AML), Acute Non-Lymphocytic Leukemia (ANLL), Non-Hodgkin Lymphoma (NHL), Chronic Myeloid Leukemia (CML), and Acute Lymphoblastic Leukemia (ALL). Other studies also reported data for several other cancers in adults, including cancer of the: lung, nasal cavity, pharynx, larynx, and related sites; esophagus; stomach; colon, rectum, and anus; pancreas; kidney; liver and biliary tract; prostate; bladder, brain, and central nervous system; and skin.

At the cell level, there is scientific evidence both in experimental and epidemiological studies, that benzene exposure, even in low concentrations have a high hemolytic potential (R. Hosseinzadeh et al., 2016), that can lead to a decrease in white blood cell count, lymphocytes, platelet counts and B cells (Q. Lan et al., 2004). Regarding the suppression of components and functions of the immune system, is noted in subjects exposed to benzene (B. Li et al., 2009, N. Uzma et al., 2010).

It is well established that benzene and/or its metabolites cause chromosomal aberrations in the peripheral blood lymphocytes (V. Kašuba et al., 2000, Forni et al.,

1979, Smith et al., 2010). These chromosomal rearrangements and mutations are on the causal pathway to malignancies such as AML and ALL. During the metabolism of benzene, oxygen radicals are produced and can induce toxic effects. This active oxygen, known as oxidative stress, can damage cellular DNA, reported in mouse bone marrow in vivo studies (P. Kolachana et al., 1993). Studies have also shown that benzene can provoke epigenetic marks including histone modification, DNA methylation, and microRNA expression (S. Guil et al., 2009). Recent progress in the field of epigenetics has highlighted the fundamental role of epigenetic mechanisms in ensuring the proper control of key biological processes. It is now known that both genetic and epigenetic mechanisms are responsible for the establishment and progression of cancer (M. Esteller et al., 2008).

There are also less-known effects associated with benzene exposure. Other studies report effects in reproductive dysfunction in female workers, exposure to benzene could interrupt the function of the hypothalamic-pituitary-ovarian axis and affect their normal levels of follicle-stimulating hormone, urine pregnandiol-3-glucuronide, luteinizing hormone, and estrone conjugate (H. Chen et al., 2001, S. R. Reutman et al., 2002), another study indicated a higher incidence rate of menstrual disorder in the exposed group (X. Y. Huang et al., 1991). Besides affecting the reproductive health of females, effects in sperm total count and motility, and an increased incidence of chromosomally defective sperm were also noted in industrial workers (V. Katukam et al., 2012, F. Marchetti et al., 2012), which means that benzene exposure can be associated with male infertility.

Benzene metabolism is intrinsically complex (R. Snyder et al., 1996), it first occurs in the liver and lungs, where it is metabolized to a variety of products that are transported to the bone marrow where the secondary metabolism occurs (P. Sheets et al., 2004). The study of the relationship between the metabolism and toxicity of benzene indicates that several metabolites of benzene play significant roles in generating benzene toxicity. In Figure 1.2 is represented a very simplified scheme for benzene metabolism (C. McHale et al., 2012), including pathways and enzymes that lead to toxicity.

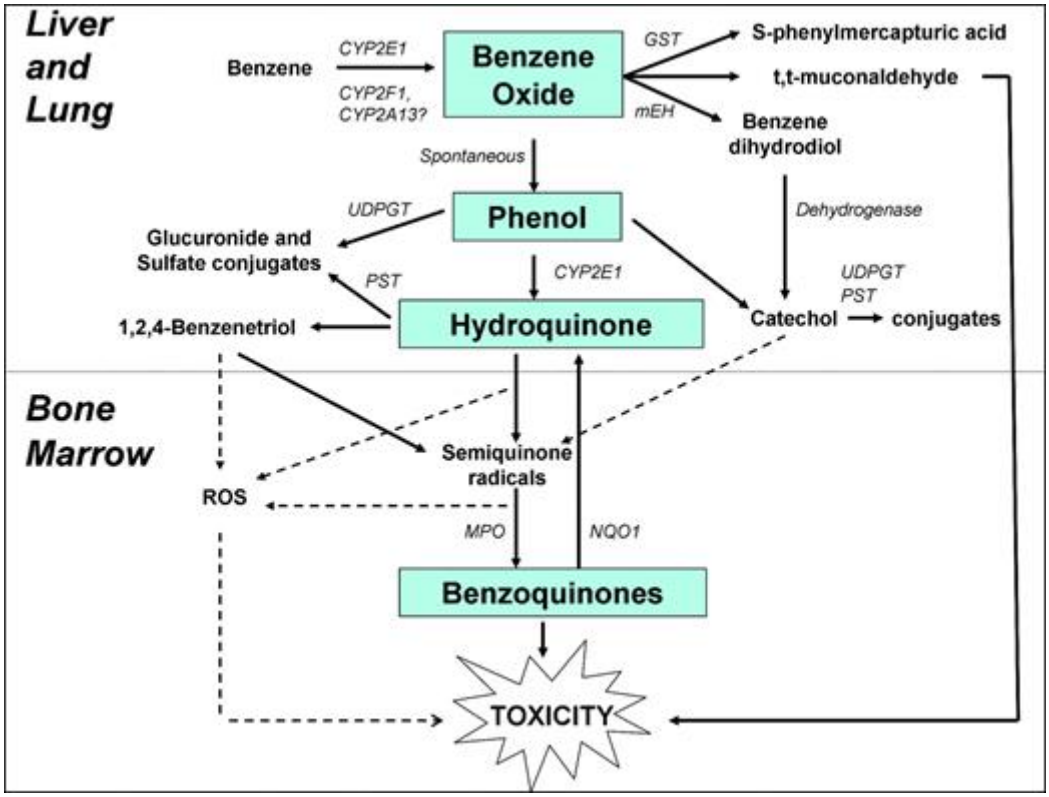

Figure 1.2: Simplified scheme for benzene metabolism (C. McHale et al., 2012).

Benzene is a much-studied chemical in respect of blood components, but its broad spectrum of incidence in the human body and the role in other chronic disorders is still unclear. Further research of the mechanisms through which benzene alters gene expression is needed to better comprehend the toxic potential and to develop appropriate preventative measures, particularly for occupationally-exposed subjects. A disease phenotype is rarely a consequence of an irregularity in a single gene product, but a reflection of several pathobiological processes that interact in a complex network. Network-based approaches have potential biological and clinical applications, from the identification of disease genes to better drug targets (A. L. Barabási et al., 2011). Transcriptomic studies measure gene expression in different conditions (such as exposed vs unexposed), genes that are differentially regulated between these conditions are of great interest, which can identify biological mechanisms affected by benzene exposure and also their mutual relationships.

The data sets generated in these gene analysis studies are not easily interpreted due to a large amount of information. The development of methods that can interpret these sets of differential expressed genes allows us to identify and quantify changes in biological mechanisms.

Gene Set Analysis (GSA) summarizes the extensive list of differentially expressed

genes in terms of biological relevant sets, based on shared biological or functional properties as defined by a reference knowledge base. The most well-known knowledge base sources are gene ontology (GO) based on functional annotation (M. Ashburner et al., 2000), Kyoto Encyclopedia of Genes and Genomes (KEEG) based on pathways (M. Kanehisa et al., 2010), and Molecular Signatures Databases (MSigDB), a collection of annotated gene sets (A. Liberzon et al., 2011).
The Over-Representation Analysis (ORA) tests the overlap of a predefined group of genes and the set of differentially expressed genes assuming the hypergeometrical distribution under the null hypothesis (R. Breitling et al., 2004). Gene Set Enrichment Analysis (GSEA) overtakes two of the major problems of ORA as it uses a valid sampling procedure and computes over the whole extension of genes (A. Subramanian et al., 2005).
Neither ORA nor GSEA consider that genes are dependently expressed; genes are found to be correlated due to mechanisms of co-regulation and co-expression (L. Geistlinger et al., 2011).
All of these methods consider a high correlation between RNA abundance and protein expression. However, protein expression levels are highly variable due to translational and post-translational events such as protein modification or binding. The output of these methods consists of a list of p-values quantifying the association of each gene set with the experimentally derived data. The drawback of these methods is that they do not take into account any direction of regulation and only capture general biological phenomena (e.g. cell differentiation) without any regard to the mechanistic details of the process (Chindelevitch et al., 2012).
The methods present in this study to extract mechanistic insights from the system response associated with a perturbation, in a qualitative and quantitative manner, not only rely on experimental measures but also on prior biological knowledge in the form of cause-and-effect relationships. This *a priori* knowledge is used in the form of directed causal graphs since they are a common method to illustrate relationships in the data and allow to display interactions between biological entities (A. Gebharter et al., 2014). The causal biological networks are capable of capturing the relationships between the biological mechanisms, allowing an assessment of the impact of exposures to active substances (T. M. Thomson et al., 2013).

1.3 Objectives

The aim of this work is to elaborate an approach to study a stimulus-response behavior in a qualitative and quantitative sense. The workflow includes building a

biological causal network related to benzene, processing the transcriptomic data, and implementing two algorithms for analyzing how the human body responds to benzene occupational exposure. The case-study stimulus used in this study is an mRNA signature corresponding to *in vivo* benzene exposure and the response is the internal biological processes response relating to a given disease end-point. The undertaken technologies place the measured experimental data in the context of a derived causal biological network consisting of prior knowledge and apply a set of algorithms to assess the overall biological response to the given exposure. Additionally, the results from these models need to be explored and compared to prior knowledge about the clinical effects of benzene on the human body, in order to evaluate the efficiency of these methods.

2

Methods

2.1 Resources and tools

This project is implemented in R 3.6.3 and Python 3.7.6. The *numpy* (S. Van Der Walt et al., 2011) and *pandas* (McKinney, W., 2010) Python libraries were used for computing operations and for building data structures. Several R packages were also exploited: *ggplot2* (H. Wickham, 2016) for data visualization; *affy* (Gautier et al., 2004) in the RNA expression data analysis and *causalR* (G. Bradley et al., 2017) for computing RCR methods. The Integrated Network and Dynamical Reasoning Assembler (INDRA) software (B.M. Gyori et al., 2017), was used to build the causal networks. It assembles mechanistic information through text mining techniques to generate several different kinds of predictive and explanatory models.

2.2 Algorithms

Both the following algorithms RCR and NPA have a similar workflow and combine the same two inputs: experimental data and a biological causal network. These methods can identify and quantify changes in biological mechanisms considering the measured genes from the experimental data when applied to a causal network.
There is a common structure for both algorithms, for scoring network models called HYP. A HYP is a specific type of network consisting of a single upstream node, connected to a set of downstream nodes. Some of these downstream entities could be genes present in the experimental data, which means it is possible to deduce information about the activity of the upstream node. Each edge represents a qualitative interaction of increases, decreases, or ambiguous. All of the downstream nodes are assumed to be independent of each other, resembling a qualitative Bayesian network.

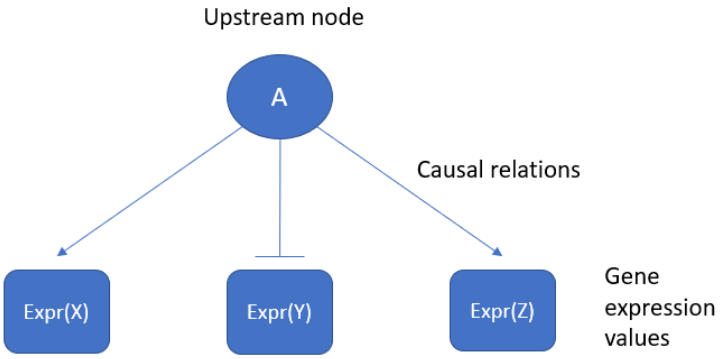

Figure 2.1: Schematic representation of a HYP network.

In the example in figure 2.1 entity A regulates the expressions of genes X, Y, and Z following the specific regulation signs "→" represents a positive regulation and "−|" represents a negative regulation.
The causal network is composed of several nodes connected by direct edges. To generate a HYP, a reference node must first be selected within the network. The selected node will be the upstream node and the downstream nodes will be all the neighbors connected to it. Next, the causal relationships between the nodes are set based on the original causal network. An increased edge connecting two nodes means that the nodes are connected by a "activates" interaction, in the same way, a decreasing edge connecting two nodes is derived from a "inhibits" interaction. When the same pair of nodes are connected by two different interactions, their relationship is set as ambiguous.

2.2.1 RCR: Reverse Causal Reasoning

This algorithm identifies biological mechanisms that are statistically significant for differential measurements in molecular profiling data when applied to a causal network (N. L. Catlett et al., 2013). The methods of this algorithm handle non-numerical data as input (here, the list of differential expressed genes).

The implementation of RCR is performed with the support of the open-source causal network analysis platform CausalR.
The reverse causal reasoning is applied to each HYP individually, assigning a direction of increase or decrease. This direction represents the deducted state of the upstream node, taking into account the states of the downstream nodes. The

direction is set based on the majority of the significantly increased or decreased downstream nodes of the HYP. For each one of the downstream nodes classified as up- or down-regulated, the interaction with the upstream node determines if the observed state is consistent with the assigned direction of increase or decrease of the upstream node.

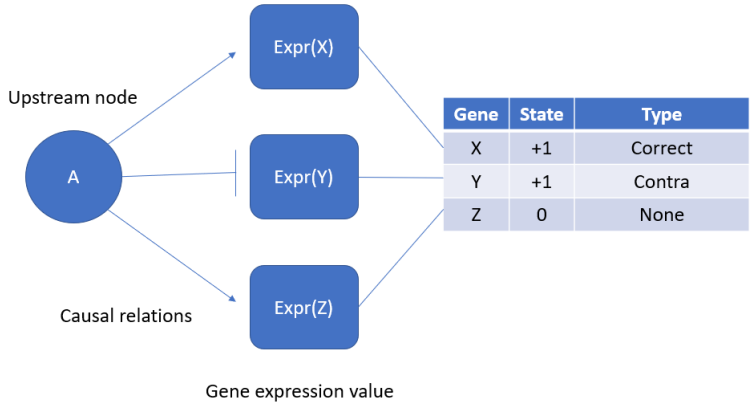

Figure 2.2: Mapping of the downstream nodes to a HYP network.

For each downstream node is assigned one of the different predictions: *correct*, *contra*, *ambiguous* and *none* . In the example in figure 2.2, if the upstream node A is connected to downstream node X by a causal increase, an increase in X would be consistent with an increase in A, *correct*. In the other case, if upstream node A is connected to downstream node Y by a causal decrease, an increase in Y would be consistent with a decrease in A, *contra*. If the downstream node has not a value of the expression is assigned as *none* and if, for example, upstream node A is connected both by a causal increase and causal decrease to downstream node W, it is set as *ambiguous*.

Based on the number of *correct*, *contra*, *ambiguous* and *none*, two evaluating metrics are calculated for each HYP: concordance and richness. Both statistics are biased in favor of HYPs with a larger number of downstream nodes because bigger networks are associated with superior levels of significance than those with few downstream nodes.

These two metrics work as a filter to identify possible and viable explanations for the given data.

Concordance

The concordance statistic is calculated as a p-value that represents the coherence between the defined state of the downstream nodes with the inferred state assigned to

the upstream node of the HYP. To compute concordance, only downstream nodes classified as up or down-regulated are used. Besides, only the nodes classified as *correct* or *contra* are considered.
The concordance is the cumulative probability based on the area under the curve of a probability distribution function, where:

- k is the number of predictions set as *correct*,
- n is the number of significant measured downstream nodes,
- p is the probability of achieving a result, in this case the probability of getting the correct prediction is 0.5, and
- l is the number of downstream nodes set as *ambiguous*.
  The concordance p-value for a given HYP, x_i, is the sum of $prob_{ij}$ for all $j = k_i, k_i + 1,..., min(n_i l_i, m)$, and can be calculated as:

$$\text{conc}_i = \sum_{j=k_i}^{\min(n_i-1,m)} \begin{pmatrix} n_i - l_i \\ j \end{pmatrix} p^j (1-p)^{n_i-j-l_i} \tag{2.1}$$

Richness

The richness statistic is also computed as a p-value, but it represents the enrichment of the network. It compares the number of downstream nodes classified as up or down-regulated of the HYP to the total number of significant measured nodes in the experimental data.
Similarly, as concordance, the richness is a cumulative probability based on the area under the curve of a probability distribution function, where:

- k is the number of downstream nodes classified as up or down-regulated of the HYP,
- N is the total number of nodes in the experimental data,
- m is the number of measured genes classified either as up or down-regulated in the experimental data,
- n is the total number of downstream nodes of the HYP.

The richness p-value for a given HYP, x_i, is the sum of $prob_{ij}$ for all $j \leq k_i$, where $j = k_i, k_i + 1,..., min(n_i, m)$ and can be calculated as:

$$\text{rich}_i = \sum_{j=k_i}^{\min(n_i m)} \frac{\begin{pmatrix} m \\ j \end{pmatrix} \begin{pmatrix} N-m \\ n_i - j \end{pmatrix}}{\begin{pmatrix} N \\ n_i \end{pmatrix}} \tag{2.2}$$

2.2.2 NPA: Network Perturbation Amplitude

This algorithm combines experimental data and a causal network to quantify changes or perturbation in biological processes, based on the magnitude and direction of expression changes of the downstream nodes in each HYP (F. Martin et al., 2012). The methods of this algorithm handles numerical data (here, the expression values of the differential expressed genes).
In this study, two different NPA scores were computed to evaluate the activity of the biological process represented in the HYPs: Strength and Geometric Perturbation Index (GPI). These metrics are designed so that positive values denote the increased activity of the upstream node in the HYP and in the other way, negative values mean decreased activity when compared to the control.
The HYP mapping performed in NPA is similar to the one applied in RCR, but instead of using the "State" of the gene it takes the real value of the logarithmic expression. The same way that in RCR, directionality is fundamental for NPA scoring.

Strength

The Strength is calculated as the weighted mean of the logarithmic differential expressions of the HYP genes, where:

- β_i is the log2FC of the i^{th} gene in the HYP,
- s_i is the type of interaction between the upstream regulator of the HYP and the i^{th} gene (+1 for activation and -1 for inhibition), and
- N is the number of nodes in the HYP, present in the transcriptome.

$$Strength = \frac{1}{N} \sum_{i=1}^{N} s_i \cdot \beta_i \tag{2.3}$$

This scoring method is vulnerable to noise because it considers all the measured downstream genes independently of data quality (here, the p-value for expression), assuming that noise is evenly distributed.
A positive score for Strength suggests that the upstream node regulator is up-regulated in the exposed group compared to the unexposed group. A negative Strength score denotes that the process has a decrease in its activity.

GPI

GPI method is a modified version of the Strength method. GPI is normalized by $\sqrt{N}$ instead of N. The weighing in GPI adds another factor, the false non-discovery

rate.
The false discovery rate (FDR) described by Y. Benjamini and Y. Hochberg (Y. Benjamini & Y. Hochberg, 1995) is the expected proportion of type I errors (false positives). The FDR plays a prominent role in many high-dimensional testing and model selection procedures. The FDR is obtained from the raw p-values using the Benjamini-Hochberg multiple testing corrections. The local false non-discovery rate, $fndr_i$ is calculated as $fndr_i = 1 - fdr_i$. This way genes with low p-values will have a higher influence in the score than those with a high p-values.

$$GPI = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} s_i \cdot fndr_i \cdot \beta_i \tag{2.4}$$

A GPI positive score suggests that the upstream regulator process has an increase in its activity and a negative score suggests that the process is down-regulated.

2.3 Data Overview

2.3.1 Transcriptomic data

Toxicogenomics can help us understanding both the mechanism of toxicity and predict compound toxicity by using omics data. There are different types of omics data: genomics, proteomics, transcriptomics, epigenomics, metabolomics, and pharmacogenomics.
Available omics data regarding benzene exposure in humans is only found in the context of transcriptomics. The transcriptome is the complete set of all RNAs transcribed by certain tissue or cell at a specific stage of development or physiological condition (Z. Wang et al., 2010).
Analyzing the transcripts whose abundance is altered by the experimental conditions (e.g. exposed vs unexposed) can reveal the mechanisms of disease processes and the mode of action for toxicity and adverse effects on cellular responses induced by exposures to chemicals or drugs. A limitation of this microarray technology is that gene expression microarrays measure changes in mRNA abundance and not of proteins, thus interpretation of these data must be done with caution.
The volume of omics data is increasing rapidly every year (Y. Perez-Riverol et al., 2019). Most of these are performed in non-human organisms, yet in this study, human omics data is used. Different organisms can have very different biological responses to the same perturbation, thus for a deep understanding of how a compound affects the human system, human omics data can give a better insight of these

compounds toxicity. The use of human omics data needs a careful interpretation due to the high biological variation between subjects, therefore it is important to collect meta-data to match comparison groups strongly.
The data used in this study (C. McHale et al., 2009) consist of eight workers exposed to Benzene (mean air benzene level $\pm$ SD = 39.0 $\pm$ 25.5 ppm) and eight unexposed controls ($<$ 0.04 ppm). Two micro-array platforms, Affymetrix and Illumina, are used in the identification of genes induced by benzene exposure in the Peripheral Blood Mono Nuclear cells (PBMC).

2.3.2 Biological Causal Network

Causal networks can describe relationships between heterogeneous biological entities, including protein, genes, chemicals, and biological processes. The nodes are connected by direct edges representing a type of interaction. These relationships are supported by literature reference.
The biological causal network can be built using the Natural Language Processing (NLP) based software INDRA, from a collection of causal relationships existent in the literature. The resulting causal graph can be queried to suggest molecular hypotheses that explain the changes in a high-throughput gene expression analysis. INDRA allows us to use a set of queries, that are searched in a large database, to extract scientific pieces of evidence, which are then stored as statements. These statements are converted to causal relationships having the format "A increases B". After that, all these relationships are merged forming a comprehensive biological network describing the interactions of benzene with other biological entities the human body.

3

Results

3.1 Data processing

To process the data from the McHale et al. study of 2012 a set of steps are made to build the right input for RCR and NPA methods.
Illumina and Affymetrix plataforms were used for the identification of the genes present on the collected data. Both platforms yield highly comparable data, especially for genes predicted to be differentially expressed (M. Barnes et al., 2005). For the simplification process, only the genes in the Affymetrix platform are considered. Analyzing RNA expression data requires several steps, with numerous potential methods available for each step, the data processing in this analysis is performed with the aid of multiple R packages. Raw expression data is analyzed using the *affy* and *limma* (M.E. Ritchie et al., 2015) R packages. The first step is the Robust Multi-array Average (RMA), which is an algorithm used to generate probe set expression values (R. A. Irizarry et al., 2003). This algorithm is composed by several steps: i) background correction, to remove the local artifacts and noise, this way measurements are not affected by neighboring measurements; ii) quantile normalization, thus data from different arrays can be compared. The second step is the construction of the gene expression matrix, where each row represents a probe set and each column represents a sample. Each entry in the matrix represents the expression value of a particular probe set in a given sample. A linear model is fitted to the expression data for each probe set. The coefficients of the fitted models describe the differences between the RNA sources hybridized to the arrays. Multiple statistics are computed for the linear model, such as t-statistics and F-statistics. Ending with a conversion of Affymetrix probes into genes using the annotation "hgu133plus2" (M. Carlson et al., 2016). Probe sets that map multiple genes are discarded and when multiple probe sets map the same gene, only one of them is randomly selected. The initial 22283 probe sets are then mapped to 12402 distinct genes.

Table 3.1: Summary of the list of expressed genes.

Affymetrix platform (threshold: log2FC = 0.5)		
Up-regulated	**Down-regulated**	**Non-significant**
235	102	12065

In table 3.1 it is represented the number of identified genes using the R package *affy*, these genes are classified as up-regulated, down-regulated or non-significant, according to their log2FC values. Logarithmic-Fold Change (log2FC) is a measure that describes how much a quantity varies between two conditions (here, unexposed *vs* exposed). Genes with log2FC values higher than 0.5 are considered up-regulated, the ones with log2FC values lower than -0.5 are considered down-regulated and the remaining ones (between -0.5 and 0.5 log2FC values) are labeled as non-significant.

Table 3.2: Experimental data subset.

Gene Symbol	**log2FC**	**adj.P.Value**	**P.Value**	**State**
JUN	-1.754	0.063	5.479E-05	-1
CEP97	-0.143	0.753	8.380E-02	0
PLK2	0.937	0.101	1.000E-04	1
...				

The table 3.2, constructed after processing the benzene experimental data, display a subset of 3 from all the 12402 genes available. The columns of the table 3.2 are the "Gene Symbol", designated by the HUGO Gene Nomenclature Committee (HGNC), the "log2FC", the "adj.P.Value", adjusted p-value or the estimated FDR, the "P.value" column contains the p-values corresponding to the t-statistics obtained with t-student tests. These values are not adjusted for multiple testing. Finally, the column "state" corresponds to gene signatures: genes with values of log2FC greater than 0.5 are labeled with the state "1" denoting up-regulation, genes with values lower than -0.5 are labeled with the state "-1" denoting down-regulation and all the other are labeled with the state "0" denoting unchanged.

To verify if the most important genes affected by benzene were captured, a comparison between the experimental data and the Comparative Toxicogenomics Database (CTD), (http://ctdbase.org/, A. P. Davis et al., 2017) is performed. The CTD database is a collection of manually curated scientific literature on the molecular mechanisms. Chemicals in CTD can be associated with genes, phenotypes, and diseases. Querying in CTD for which genes are known to be somehow affected by benzene results in 625 referred genes. In figure 3.1, a Venn diagram is represented showing the intersection of the CTD genes and the experimental genes. The results

show that not all the genes from CTD are present in the experimental data used in this project (90 are missing), suggesting that this data is not complete regarding the benzene effects in the human body.

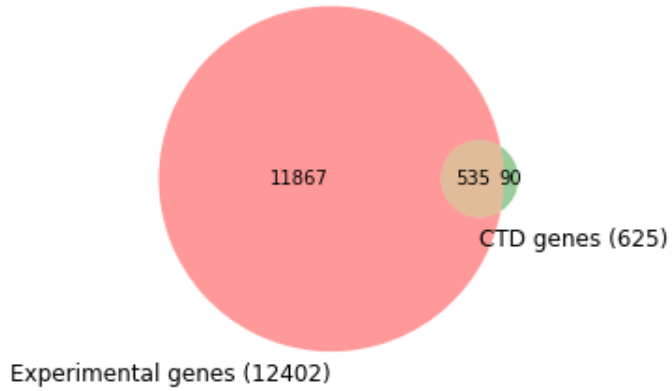

Figure 3.1: Venn diagram of genes that interact with benzene in CTD and genes in the experimental data.

3.2 Benzene Causal Network

The set of queries used to build the benzene biological network is represented in Appendix A, as well as the number of PMIDs, the unique identifier number used in PubMed (https://pubmed.ncbi.nlm.nih.gov/). The queries used represent the 10 key characteristics that are commonly exhibited by established human carcinogens (Smith, M. T. et al., 2016). These 10 key characteristics are: i) is electrophilic or can be metabolically activated, ii) is genotoxic, iii) alters DNA repair or causes genomic instability, iv) induces epigenetic alterations, v) induces oxidative stress, vi) induces chronic inflammation, vii) is immunosuppressive, viii) modulates receptor-mediated effects, ix) causes immortalization, and x) alters cell proliferation, cell death or nutrient supply.

The 10 key characteristics are grouped in 7 queries and introduced in INDRA software, and each one of these queries returns a network. These networks are then merged and the non-human genes are removed. The merged network has a total of 949 nodes and 1583 edges extracted from 1933 unique PMIDs.

In order to increase the complexity of the network, to a better understanding of benzene toxicity, more statements describing causal relationships between biological entities and processes at different levels need to be added. These statements are obtained using the Biological Expression Language (BEL), having a similar causal

format as referred above (https://bel.bio/). The set of BEL statements stored at "Selventa Large Corpus" contain about 80,000 statements, from 16,000 citations. BEL represents scientific findings from the literature in a computable form.
A neighborhood search is performed in the original network using the PyBEL python package (C. Hoyt et al., 2018) adding a set of new causal relationships, increasing the network complexity for 19786 nodes and 48490 edges.

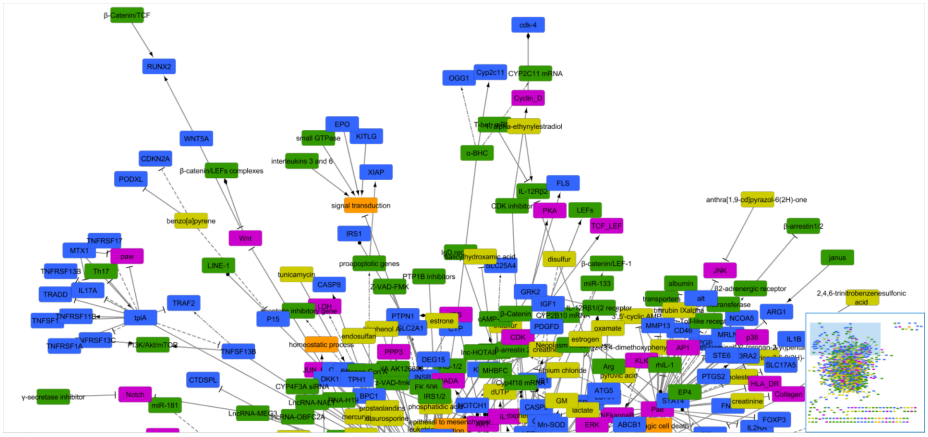

Figure 3.2: Representation of a part of the biological causal network in Cytoscape.

In figure 3.2 is represented a section of the network in Cytoscape software (P. Shannon et al., 2003). The different colors represent the type of molecular entity, e.g. blue represent genes. The only types of interactions present in this network are "activation" consisting of arrows and "inhibition" consisting of a line with a trace. This heterogeneous network has 7137 genes, 249 biological processes, and other types of entities, such as chemicals. For this study, only genes and biological processes are truly relevant, from a bioinformatics perspective, since we want to analyze and study the changes in the expression of the bio-processes related to benzene based on the gene expression measurements.
The network is then converted to CX file format using Cytoscape, an open-source software platform capable of visualizing complex networks and integrating them with any new type of data.
Finally, the enriched causal network is transformed to a Simple Interaction Format file (SIF), taking the form present in table 3.3.

Table 3.3: Causal network in SIF format.

GeneA	Activates	ProteinB
ChemicalC	Inhibits	BioprocessD

Only two types of interactions are considered: "Activates" and "Inhibits". For

the remaining interactions present in the original network some are relabeled and others are removed. When the interaction is "DecreaseAmount" it is converted into "Inhibits" and "IncreaseAmount" is converted into "Activates". The other types of interactions are removed, such as "complex" and "phosphorylation".
Considering now the experimental data and the biological causal network, it is important to know how many genes were present in both inputs. Since it will be these genes that will contribute to the algorithms computations.

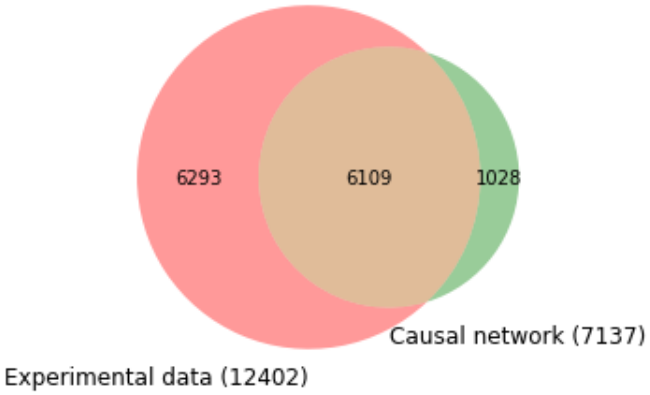

Figure 3.3: Venn diagram of genes in the experimental data and genes in the causal biological network.

In the figure 3.3, the Venn diagram illustrates the intersection between the two sets, showing a total of 6109 matching genes that will be used in the methods performed in this study.

Table 3.4: Differential expressed genes present both in the causal network and the experimental data, by state.

Up-regulated	Down-regulated	Non-significant
177	77	5855

In the table 3.4 it is possible to observe how many of the genes, present in the intersection, are considered as up-regulated, down-regulated or non-significant. These are the 6109 genes that are both present in the outcome of statistical analysis on the McHale dataset and the biological causal network.
All of the 249 HYPs have a biological process as an upstream node, in the table 3.5 it is represented a list of the 12 HYPs with more downstream nodes. However not all of the downstream nodes present in each HYP are of interest, so the number of neigbors are not representative of the real complexity of the HYP.

Table 3.5: List of top-12 bioprocesses with more downstream nodes.

Bioprocess	No. of neighbors
response to hypoxia	1367
response to heat	283
aging	265
myoblast differentiation	158
neuron differentiation	158
replicative cell aging	137
response to oxidative stress	133
angiogenesis	133
response to starvation	116
DNA methylation	102
response to osmotic stress	78
response to UV	77

3.3 RCR and NPA results

For a complete assessment of the biological impact across all the network, RCR and NPA procedures are applied to all of the 249 HYPs generated from the biological causal network using the gene expression data for occupational benzene exposure. Regarding the RCR methodology, of the 249 HYPS studied, 10 met the concordance and richness p-values significance thresholds of 0.1 (see table 3.6). According to the paper describing RCR by N. L. Catlett et al., 2013, p-values under 0.1 for both concordance and richness limits the number of false positives and false negatives to an acceptable level, in bold are represented the HYPs with even more stringent thresholds of 0.05.

Table 3.6: HYPs with both richness and concordance p-values thresholds of 0.1, ranked by richness and concordance

Name	Direction	Concordance	Richness
neutrophil chemotaxis	1	0.001	0.087
cell fate determination	**1**	**0.002**	**0.008**
histone deacetylation	**1**	**0.005**	**0.034**
stress-induced premature senescence	**1**	**0.008**	**0.017**
mast cell activation	**-1**	**0.013**	**0.043**
response to UV	**-1**	**0.015**	**0.046**
response to heat	**-1**	**0.017**	**0.000**
embryo implantation	**1**	**0.030**	**0.043**
angiogenesis	**1**	**0.042**	**0.001**
response to UV-B	-1	0.053	0.001

Of the 10 HYPs that met the significance thresholds, 4 of them are assigned with direction of "-1" meaning a decrease in the activity of these mechanisms: *mast cell activation*, *response to UV*, *response to heat* and *response to UV-B*. The remaining biological processes are set with a direction of "1", denoting an increase in their activity: *neutrophil chemotaxis*, *cell fate determination*, *histone deacetylation*, *stress-induced premature senescence*, *embryo implantation* and *angiogenesis*.
In figure 3.4 it is represented the HYP with the upstream node *angiogenesis*, coloured with blue. This sub-network contains 133 downstream nodes, where green represents genes significantly up-regulated (11), red represents genes significantly down-regulated (5), grey represents genes with no significant change (98) and black represents genes not present in the transcriptomic data (19). The two types of interactions are inhibition (colored with red) and activation (colored with green). It is possible to visualize inter-dependency between genes. This HYP is classified as having positive regulation by RCR with a concordance of 0.042 and with a richness of 0.001.

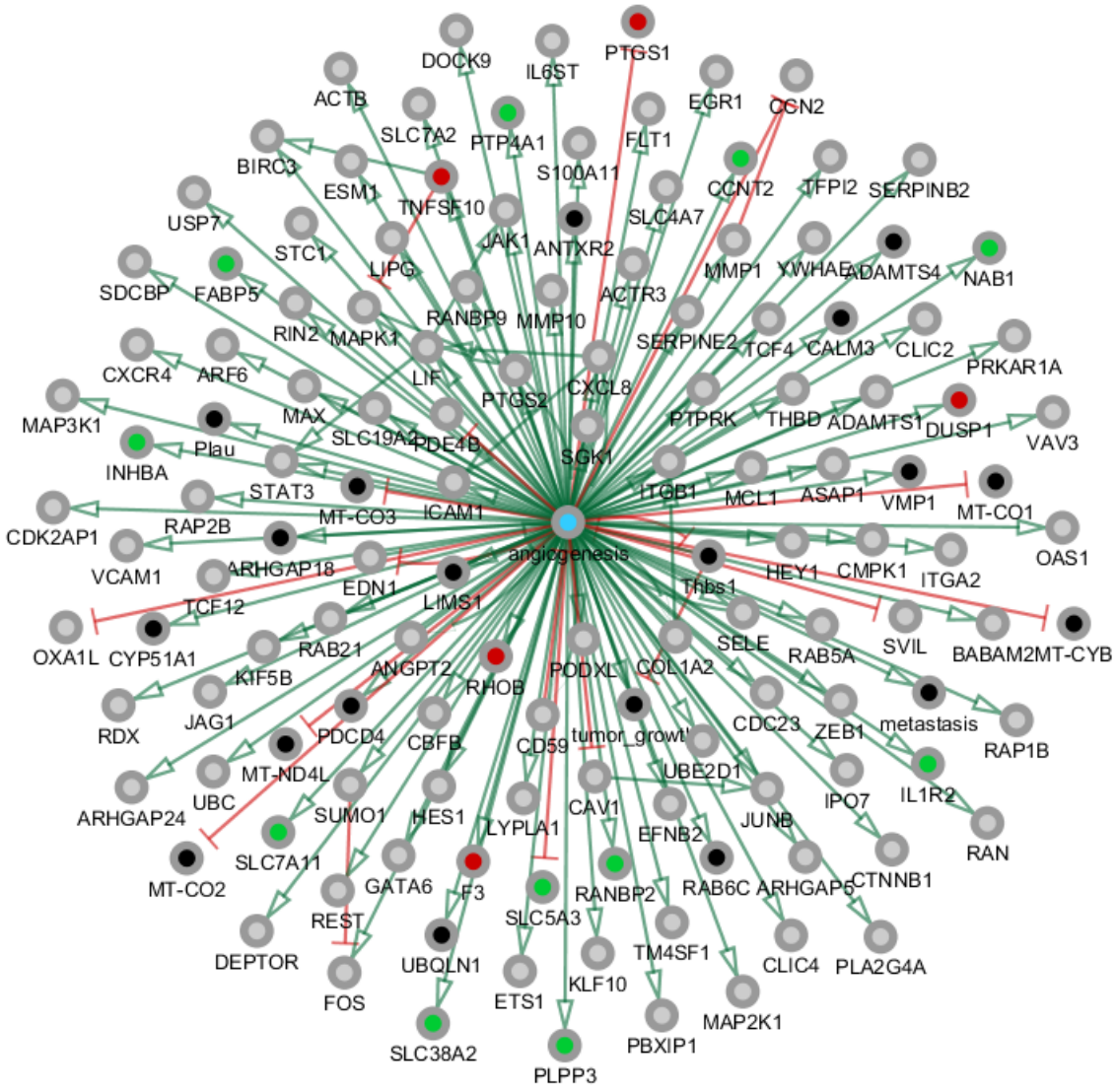

Figure 3.4: HYP corresponding to *angiogenesis* biological process.

Respecting the NPA methodology, there were 72 biological processes with a non-zero value of Strength and GPI. Table 3.7 display the top-10 mechanisms with higher absolute value of Strength.

Table 3.7: Top 10 HYPs with absolute higher values of Strength.

Name	Strength
mast cell activation	-0.955
embryo implantation	0.837
stress-induced premature senescence	0.390
response to UV-A	-0.327
actin filament polymerization	0.306
feeding behavior	0.303
response to mechanical stimulus	0.295
neutrophil activation	0.295
histone deacetylation	0.273
cell growth	0.208

Observing table 3.7, a biological mechanism with negative strength values suggest a decrease in their activity: *mast cell activation* and *response to UV-A*. On the other way, the remaining mechanisms, with positive strength values, denote an increase in their expression: *embryo implantation*, *stress-induced premature senescence*, *actin filament polymerization*, *feeding behavior*, *response to mechanical stimulus*, *neutrophil activation*, *histone deacetylation*, and *cell growth*.

Following the same label as figure 3.4, in figure 3.5 represents the HYP with the upstream node *histone deacetylation*. It contains 11 downstream nodes of which are: 1 significant up-regulated, 1 significant down-regulated, 5 with no significant change, and 5 not present in the experimental data. Once more, inter-dependency between genes is displayed and also self-loops in nodes *cyclic AMP* and *AGT* representing positive feedback. This HYP is classified as having an activity increase by NPA, with a Strength value of 0.273.

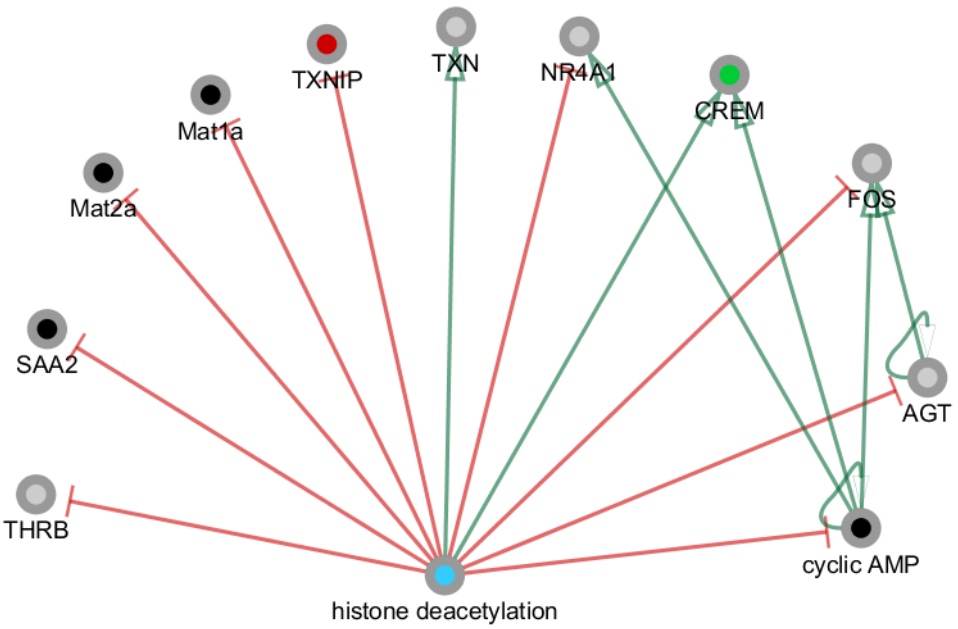

Figure 3.5: HYP corresponding to histone deacetylation aging biological process.

The table 3.8 display the top-10 mechanisms with higher absolute value of GPI.

Table 3.8: Top 10 HYPs with absolute higher values of GPI.

Name	GPI
response to UV	-0.522
angiogenesis	0.355
stress-induced premature senescence	0.325
cell fate determination	0.323
embryo implantation	0.317
histone deacetylation	0.315
inflammatory response	0.314
response to UV-B	-0.262
response to radiation	-0.261
microtubule polymerization	-0.258

Observing table 3.8, in a similar way as Strength, a negative value of GPI is associated with a reduction of activity of the respecting biological mechanism: *response to UV*, *response to UV-B*, *response to radiation* and *microtubule polymerization*. HYPs with positive values of GPI denote an increase in their activity: *angiogenesis*,

stress-induced premature senescence, *cell fate determination*, *embryo implantation*, *histone deacetylation* and *inflammatory response*.

Following the same label as figure 3.4, figure 3.6 represents the HYP with the upstream node *inflammatory response*. This sub-network is composed of 66 downstream nodes of which are 3 significantly up-regulated, 0 significantly down-regulated, 31 with no significant change, and 32 are not present in the transcriptome. This HYP is classified as having an activity increase by NPA, with a GPI value of 10.687.

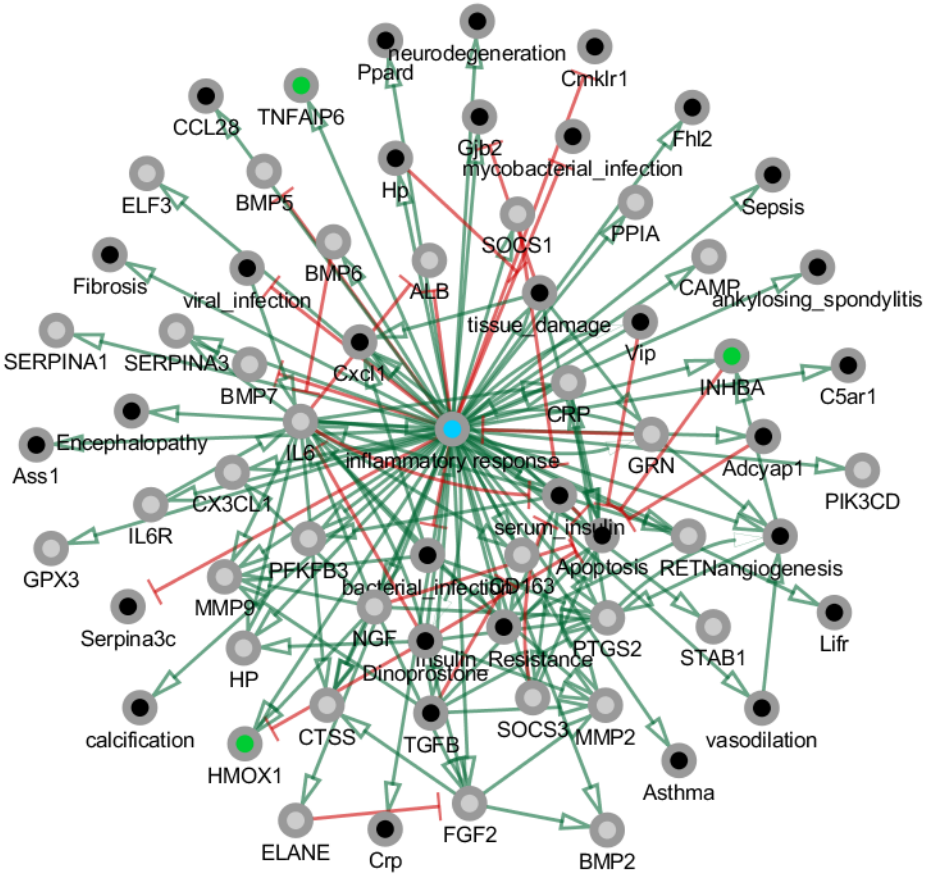

Figure 3.6: HYP corresponding to inflammatory response biological process.

4

Discussion and Conclusion

4.1 Discussion

In this project, an integrated approach that encompasses the generation of a biological causal network, differential expression gene analysis, and a combination of RCR and NPA methods has been developed. The results provide potential biological explanations and quantify the activity changes of a set of biological processes in response to occupational benzene exposure, measured by transcriptomics data.

Similarly to other techniques, RCR and NPA rely on the assumption that changes in RNA expression are equivalent to changes in the activity of the corresponding proteins, thus interpretation of results need to be done carefully.

RCR and NPA present the advantage relative to other techniques, by considering the directionality between the upstream regulator (here, a bio-process) and the downstream measurable entities (here, genes). HYPs, when compared to gene sets have the benefit of adding up the contributions of the individual downstream nodes. RCR is used to identify qualitatively the likelihood of the biological response of network perturbation. NPA offers the possibility to quantify the biological responses to a given perturbation (here, benzene exposure). The two NPA scoring methods, Strength and GPI, employ a distinct approach to measure the magnitude of perturbation between two experimental measures for a given HYP. GPI favors small sets of strongly differentially expressed genes rather than large sets of weakly differentially expressed ones, on the other hand Strength is unbiased since it does not contain weighting factors. Thus the two metrics produce different results and they should be analyzed together to prevent the extraction of conclusions that may be specific to a particular NPA score.

Applying RCR and NPA methods to all of the present biological processes in the causal network lead to a wide range of hypothesis, instead of focusing on just one or a set of processes.

One of the limitations of this study is the transcriptome not having captured all of the genes known to be altered by benzene exposure in literature (90 are missing). This can happen for multiple reasons since the procedure has many steps from the collection of the micro-array data until the differential expression analysis. The transcriptome used in this study was extracted from PBMC, a core part of the immune system (M. Wen et al., 2020). PBMC represents a broad spectrum of different cell types, that vary a lot between subjects, adding even more variety to data (L. Wong et al., 2016). Although, they are very likely to be limited in what they can tell us in gene expression studies, even if they are easy to be obtained and stored.

For a successful application of this methodology, an accurate and reliable biological knowledge describing the relationships between all the entities in the network is fundamental. These relationships are curated every day by specialists, so the more interactions are uploaded to databases the deeper knowledge we will have about the benzene role in the human body. Possible interesting biological processes, like *blood coagulation*, *DNA recombination* or *metabolic process* that, due not having enough downstream neighbors, can not be evaluated. Furthermore, mechanisms that are not present in the biological network are not assessed by RCR and NPA, thus mechanisms that may even be seriously affected by benzene are not captured. The increase in our knowledge and understanding of the relationships between the biological entities of the network and sub-networks will increase the scoring results.

The results achieved by this study can give an important clue of what mechanisms are altered by benzene occupational exposure and the biological response of the human body to this exposure. Further experiment investigation need to be conducted to verify the conclusions of this work. The identified and quantified mechanisms by RCR and NPA can be used to corroborate previous literature findings or to propose novel-mechanisms for benzene toxicity.

Analyzing all the hypothesis suggested by RCR and NPA is not doable. By overlapping the top-25 results from each metric of RCR and NPA, a list of 5 bioprocesses is achieved and proceeds to a detailed discussion. This list contains the following biological processes: *stress-induced premature senescence*, *histone deacetylation*, *embryo implantation*, *inflammatory response*, and *cell fate determination*. All of these biological processes have an inferred increase in their activity. Beyond the list of the 5 biological processes, there are an inferred decrease of mechanisms related to UV rays such as *response to UV, UV-A, UV-B*, suggested across all of the results.

The *stress-induced premature senescence* process is a set of biological mechanisms

that arrests the proliferation of premalignant cells, but there is also evidence that contributes to aging (J. Campisi et al., 2007). These processes are invoked by oxidative stress, DNA damage, oncogene activity or suboptimal culture conditions (P. R. Coleman et al., 2010, A. Bielak-Zmijewska et al., 2018). Regarding the processes that elicit a senescence response, benzene induces DNA damage (J. Li et al., 2018) and promotes oxidative stress through the production of reactive oxygen species (C. Costa et al., 2016), thus the increase in the activity of this biological process is consistent with the literature.
The *histone deacetylation* is the process of removing an acetyl group from the histone structure. By deacetylating the histone tails, the DNA becomes more tightly wrapped around the histone cores, making it harder for transcription factors to bind to the DNA. This leads to decreased levels of gene expression and it is known as gene silencing (A. J. M. De Ruijter et al., 2003). Literature showing deacetylation in human exposed to benzene was not found. However, according to the study of S. Qian et al, mice exposed to low concentrations of benzene results in deacetylation, increased autophagy and haematopoietic toxicity (S. Qian et al., 2019), which is in concordance with the findings of this work.

The *embryo implantation* is the stage of pregnancy at which the embryo adheres to the wall of the uterus. No evidence is found referring the effect of benzene exposure with mechanisms associated with the implantation of the embryo, suggesting that this result is most-probably a false positive. However, further investigation needs to be done, because there are studies reporting that exposure to low-level of benzene could interrupt the function of hypothalamic-pituitary-ovarian axis (H. Chen et al., 2001). The hypothalamic-pituitary-ovarian axis, that is a tightly regulated system controlling female reproduction, responsible to select a dominant follicle for ovulation, meanwhile preparing the endometrium for implantation (S. Mikhael et al., 2019).

The *inflammatory response* is a type of response by the immune system to a variety of factors, including pathogens, damaged cells and toxic compounds. This response is characterized by redness, swelling, heat, pain, loss of tissue function and also microcirculatory events such as vascular permeability changes, leukocyte recruitment and accumulation and inflammatory mediator release (O. Takeuchi et al., 2010, L. Ferrero-Miliani et al., 2007). Multiple studies report that occupational exposure to benzene can induce alterations in the immune system biology (McHale et al., 2008, Mchale et al., 2011), corroborating the high expression of genes related to this process.

The *cell fate determination* mechanism is the process of how a particular cell develops into a final cell type. These processes include cell proliferation, differentiation, cellular movement and programmed cell death. Deregulations in processes related to *cell fate determination* can result in tumors, since oncogenic mutations can disrupt the signaling systems that govern a cell fate (F. G. Giancotti, 2014). The increase of the activity of this bioprocess is consistent with key carcinogen characteristic: "alters cell proliferation, cell death or nutrient supply". The study from F. Zolghadr et al. reports that low doses of benzene lead to an increase in the mesenchymal stem cells while, higher concentrations of benzene can induce cell death (F. Zolghadr et al., 2012)., supporting once more the results achieved.

4.2 Conclusion

In this project, a biological causal network is built for benzene, the human transcriptome is processed under two conditions (exposed vs unexposed) and RCR and NPA algorithms are implemented in a comprehensive workflow. A set of results are produced by these algorithms, some of them being consistent with literature evidence, proving the efficiency of these methods in capturing reliable mechanisms of the internal biological response to the given stimulus, benzene exposure. It is possible to conclude that the developed procedure is capable of answering the following question: what mechanisms were actually affected by benzene exposure and with which magnitude?

Further Work

In the future, the network created in this project can be enriched with new interactions allowing a deeper understanding of the benzene role in the body, extending the spectrum of interactions far beyond carcinogenic mechanisms.
A new type of transcriptome such as dose-dependent can be used to assess the maximum levels of benzene exposure to limit and regulate the exposure levels.
Another possible approach is to use other type of measurable entity instead of genes, for example, proteomics measurements could be used to compute scores for HYPs that relate the process of interest to changes in protein level. Ideally, a combination of measurements could eventually lead to better results.
Furthermore, this methodology can be applied to other experiments such as, different therapeutic agents, consumer products, exposures, or even disease progress to better understand the human biology.

Bibliography

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J.P. MesirovMolecular signatures database (MSigDB) 3.0 Bioinformatics, 27 (2011), pp. 1739-1740

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000, 25: 25-29. 10.1038/75556

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. In Nature Reviews Genetics. https://doi.org/10.1038/nrg2918

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., & Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Research. https://doi.org/10.1093/nar/gki890

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bielak-Zmijewska, A., Mosieniak, G., & Sikora, E. (2018). Is DNA damage indispensable for stress-induced senescence? In Mechanisms of Ageing and Development. https://doi.org/10.1016/j.mad.2017.08.004

Bradley, G., & Barrett, S. J. (2017). CausalR: Extracting mechanistic sense from genome scale data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx425

Breitling, R., Amtmann, A., & Herzyk, P. (2004). Iterative Group Analysis (iGA): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-5-34

Campisi, J., & D'Adda Di Fagagna, F. (2007). Cellular senescence: When bad things happen to good cells. In Nature Reviews Molecular Cell Biology. https://doi.org/10.1038/nrm2233

Carlson, M. (2016). hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.2.3.

Catlett, N. L., Bargnesi, A. J., Ungerer, S., Seagaran, T., Ladd, W., Elliston, K. O., & Pratt, D. (2013). Reverse causal reasoning: Applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-14-340

Chen, H., Wang, X., & Xu, L. (2001). Effects of exposure to low-level benzene and its analogues on reproductive hormone secretion in female workers. Zhonghua Yu Fang Yi Xue Za Zhi [Chinese Journal of Preventive Medicine]

Chindelevitch, L., Ziemek, D., Enayetallah, A., Randhawa, R., Sidders, B., Brockel, C., & Huang, E. S. (2012). Causal reasoning on biological networks: Interpreting transcriptional changes. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts090

Coleman, P. R., Hahn, C. N., Grimshaw, M., Lu, Y., Li, X., Brautigan, P. J., Beck, K., Stocker, R., Vadas, M. A., & Gamble, J. R. (2010). Stress-induced premature senescence mediated by a novel gene, SENEX, results in an anti-inflammatory phenotype in endothelial cells. Blood. https://doi.org/10.1182/blood-2009-11-252700

Costa, C., Ozcagli, E., Gangemi, S., Schembri, F., Giambò, F., Androutsopoulos, V., Tsatsakis, A., & Fenga, C. (2016). Molecular biomarkers of oxidative stress and role of dietary factors in gasoline station attendants. Food and Chemical Toxicology. https://doi.org/10.1016/j.fct.2016.01.017

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., Mc-

Morran, R., Wiegers, J., Wiegers, T. C., & Mattingly, C. J. (2017). The Comparative Toxicogenomics Database: Update 2017. Nucleic Acids Research. https://doi.org/10.1093/nar/gkw838

De Ruijter, A. J. M., Van Gennip, A. H., Caron, H. N., Kemp, S., & Van Kuilenburg, A. B. P. (2003). Histone deacetylases (HDACs): Characterization of the classical HDAC family. In Biochemical Journal. https://doi.org/10.1042/BJ20021321

Esteller, M. (2008). Epigenetics in Cancer. New England Journal of Medicine. https://doi.org/10.1056/nejmra072067

Ferrero-Miliani, L., Nielsen, O. H., Andersen, P. S., & Girardin, S. E. (2007). Chronic inflammation: Importance of NOD2 and NALP3 in interleukin-1β generation. In Clinical and Experimental Immunology. https://doi.org/10.1111/j.1365-2249.2006.03261.x

Forni, A. (1979). Chromosome changes and benzene exposure. A review. In Reviews on Environmental Health.

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). "affy—analysis of Affymetrix GeneChip data at the probe level." Bioinformatics, 20(3), 307–315. ISSN 1367-4803, doi: 10.1093/bioinformatics/btg405.

Gebharter A., Kaiser M.I. (2014) Causal Graphs and Biological Mechanisms. In: Kaiser M., Scholz O., Plenge D., Hüttemann A. (eds) Explanation in the Special Sciences. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), vol 367. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7563-3_3

Geistlinger, L., Csaba, G., Küffner, R., Mulder, N., & Zimmer, R. (2011). From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr228

Giancotti, F. G. (2014). Deregulation of cell signaling in cancer. In FEBS Letters. https://doi.org/10.1016/j.febslet.2014.02.005

Gillardon, F., Moll, I., Meyer, M., & Michaelidis, T. M. (1999). Alterations in cell

death and cell cycle progression in the UV-irradiated epidermis of bcl-2-deficient mice. Cell Death and Differentiation. https://doi.org/10.1038/sj.cdd.4400455

Guil, S., & Esteller, M. (2009). DNA methylomes, histone codes and miRNAs: Tying it all together. In International Journal of Biochemistry and Cell Biology. https://doi.org/10.1016/j.biocel.2008.09.005

Gyori B.M., Bachman J.A., Subramanian K., Muhlich J.L., Galescu L., Sorger P.K. From word models to executable models of signaling networks using automated assembly (2017), Molecular Systems Biology, 13, 954.

Hosseinzadeh, R., & Moosavi-Movahedi, A. A. (2016). Human hemoglobin structural and functional alterations and heme degradation upon interaction with benzene: A spectroscopic study. Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy. https://doi.org/10.1016/j.saa.2015.12.014

Hoyt, C. T., Konotopez, A., & Ebeling, C. (2018). PyBEL: A computational framework for Biological Expression Language. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx660

Huang, X. Y. (1991). Influence on benzene and toluene to reproductive function of female workers in leathershoe-making industry. Zhonghua Yu Fang Yi Xue Za Zhi [Chinese Journal of Preventive Medicine]

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, No. 100F. Chemical agents and related occupations (benzene). Lyon, France: International Agency for Research on Cancer; 2012:249–294. https://monographs.iarc.fr/wp-content/uploads/2018/06/mono100F-24.pdf

IARC monographs on the evaluation of carcinogenic risk of chemicals to man. Volume 7. (1974). WHO.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England). https://doi.org/10.1093/biostatistics/4.2.249

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010, 38: D355-D360. 10.1093/nar/gkp896.

Katukam, V., Kulakarni, M., Syed, R., Alharbi, K., & Naik, J. (2012). Effect of benzene exposure on fertility of male workers employed in bulk drug industries. Genetic Testing and Molecular Biomarkers. https://doi.org/10.1089/gtmb.2011.0241

Kašuba, V., Rozgaj, R., & Šentija, K. (2000). Cytogenetic changes in subjects occupationally exposed to benzene. Chemosphere. https://doi.org/10.1016/S0045-6535(99)00265-9

Kolachana, P., Subrahmanyam, V. V., Meyer, K. B., Zhang, L., & Smith, M. T. (1993). Benzene and Its Phenolic Metabolites Produce Oxidative DNA Damage in HL60 Cells in Vitro and in the Bone Marrow in Vivo. Cancer Research.

Lan, Q., Zhang, L., Li, G., Vermeulen, R., Weinberg, R. S., Dosemeci, M., Rappaport, S. M., Shen, M., Alter, B. P., Wu, Y., Kopp, W., Waidyanatha, S., Rabkin, C., Guo, W., Chanock, S., Hayes, R. B., Linet, M., Kim, S., Yin, S., ... Smith, M. T. (2004). Hematotoxicity in workers exposed to low levels of benzene. Science. https://doi.org/10.1126/science.1102443

Li, B., Li, Y. Q., Yang, L. J., Chen, S. H., Yu, W., Chen, J. Y., & Liu, W. W. (2009). Decreased T-cell receptor excision DNA circles in peripheral blood mononuclear cells among benzene-exposed workers. International Journal of Immunogenetics. https://doi.org/10.1111/j.1744-313X.2009.00832.x

Li, J., Xing, X., Zhang, X., Liang, B., He, Z., Gao, C., Wang, S., Wang, F., Zhang, H., Zeng, S., Fan, J., Chen, L., Zhang, Z., Zhang, B., Liu, C., Wang, Q., Lin, W., Dong, G., Tang, H., ... Li, D. (2018). Enhanced H3K4me3 modifications are involved in the transactivation of DNA damage responsive genes in workers exposed to low-level benzene. Environmental Pollution. https://doi.org/10.1016/j.envpol.2017.11.042

Loomis D, Guyton KZ, Grosse Y et al. . Carcinogenicity of benzene. Lancet Oncol. 2017;18(12):1574–1575.

Marchetti, F., Eskenazi, B., Weldon, R. H., Li, G., Zhang, L., Rappaport, S. M., Schmid, T. E., Xing, C., Kurtovich, E., & Wyrobek, A. J. (2012). Occupational exposure to Benzene and chromosomal structural aberrations in the sperm of Chinese men. Environmental Health Perspectives. https://doi.org/10.1289/ehp.1103921

Martin, F., Thomson, T. M., Sewer, A., Drubin, D. A., Mathis, C., Weisensee, D., Pratt, D., Hoeng, J., & Peitsch, M. C. (2012). Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. BMC Systems Biology. https://doi.org/10.1186/1752-0509-6-54

Mathur, R., Rotroff, D., Ma, J., Shojaie, A., & Motsinger-Reif, A. (2018). Gene set analysis methods: A systematic comparison. BioData Mining. https://doi.org/10.1186/s13040-018-0166-8

Mchale, C., Zhang, L., Lan, Q., Li, Q., Hubbard, A., Porter, K., Vermeulen, R., Shen, M., Rappaport, S., Yin, S., Smith, M. T., & Rothman, N. (2008). Low-Dose, Occupational Exposure to the Leukemogen Benzene Induces Robust Changes in the Blood Transcriptome Associated with Altered Immune System Biology. Blood. https://doi.org/10.1182/blood.v112.11.1207.1207

McHale, C. M., Zhang, L., Lan, Q., Li, G., Hubbard, A. E., Forrest, M. S., Rothman, N. (2009). Changes in the peripheral blood transcriptome associated with occupational benzene exposure identified by cross-comparison on two microarray platforms. Genomics. https://doi.org/10.1016/j.ygeno.2008.12.006

McHale, C. M., Zhang, L., Lan, Q., Vermeulen, R., Li, G., Hubbard, A. E., Porter, K. E., Thomas, R., Portier, C. J., Shen, M., Rappaport, S. M., Yin, S., Smith, M. T., & Rothman, N. (2011). Global gene expression profiling of a population exposed to a range of benzene levels. Environmental Health Perspectives. https://doi.org/10.1289/ehp.1002546

Mchale, C. M., Zhang, L., & Smith, M. T. (2012). Current understanding of the mechanism of benzene-induced leukemia in humans: Implications for risk assessment. Carcinogenesis. https://doi.org/10.1093/carcin/bgr297

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

Mikhael, S., Punjala-Patel, A., & Gavrilova-Jordan, L. (2019). Hypothalamic-pituitary-ovarian axis disorders impacting female fertility. In Biomedicines. https://doi.org/10.3390/biomedicines7010005

Perez-Riverol, Y., Zorin, A., Dass, G., Vu, M. T., Xu, P., Glont, M., Vizcaíno, J. A., Jarnuczak, A. F., Petryszak, R., Ping, P., & Hermjakob, H. (2019). Quantifying the impact of public omics data. Nature Communications. https://doi.org/10.1038/s41467-019-11461-w

Phillipson, R. P., Tobi, S. E., Morris, J. A., & McMillan, T. J. (2002). UV-A induces persistent genomic instability in human keratinocytes through an oxidative stress mechanism. Free Radical Biology and Medicine. https://doi.org/10.1016/S0891-5849(01)00829-2

Reutman, S. R., LeMasters, G. K., Knecht, E. A., Shukla, R., Lockey, J. E., Burroughs, G. E., & Kesner, J. S. (2002). Evidence of reproductive endocrine effects in women with occupational fuel and solvent exposures. Environmental Health Perspectives. https://doi.org/10.1289/ehp.02110805

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma. Nucleic Acids Research. https://doi.org/10.1093/nar/gkv007

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Research. https://doi.org/10.1101/gr.1239303

Sheets, P. L., & Carlson, G. P. (2004). Kinetic factors involved in the metabolism of benzene in mouse lung and liver. Journal of Toxicology and Environmental Health - Part A. https://doi.org/10.1080/15287390490273488

Smith, M. T. (2010). Advances in understanding benzene health effects and susceptibility. In Annual Review of Public Health. https://doi.org/10.1146/annurev.publhealth.012809.103646

Smith, M. T., Guyton, K. Z., Gibbons, C. F., Fritz, J. M., Portier, C. J., Rusyn, I., DeMarini, D. M., Caldwell, J. C., Kavlock, R. J., Lambert, P. F.,

Hecht, S. S., Bucher, J. R., Stewart, B. W., Baan, R. A., Cogliano, V. J., & Straif, K. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. In Environmental Health Perspectives. https://doi.org/10.1289/ehp.1509912

Snyder, R. (2012). Leukemia and benzene. In International journal of environmental research and public health. https://doi.org/10.7326/0003-4819-99-6-885

Snyder, R., & Hedli, C. C. (1996). An overview of benzene metabolism. Environmental Health Perspectives. https://doi.org/10.2307/3433158

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.0506580102

Takeuchi, O., & Akira, S. (2010). Pattern Recognition Receptors and Inflammation. In Cell. https://doi.org/10.1016/j.cell.2010.01.022

Thomson, T. M., Sewer, A., Martin, F., Belcastro, V., Frushour, B. P., Gebel, S., Park, J., Schlage, W. K., Talikka, M., Vasilyev, D. M., Westra, J. W., Hoeng, J., & Peitsch, M. C. (2013). Quantitative assessment of biological impact using transcriptomic data and mechanistic network models. Toxicology and Applied Pharmacology. https://doi.org/10.1016/j.taap.2013.07.007

Uzma, N., Kumar, S. S., & Hazari, M. A. H. (2010). Exposure to benzene induces oxidative stress, alters the immune response and expression of p53 in gasoline filling workers. American Journal of Industrial Medicine. https://doi.org/10.1002/ajim.20901

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. Computing in Science and Engineering. https://doi.org/10.1109/MCSE.2011.37

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. https://doi.org/10.1038/nrg2484

Wen, M., Cai, G., Ye, J., Liu, X., Ding, H., & Zeng, H. (2020). Single-cell transcriptomics reveals the alteration of peripheral blood mononuclear cells driven by sepsis. Annals of Translational Medicine. https://doi.org/10.21037/atm.2020.02.35

Wickham H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wong, L., Jiang, K., Chen, Y., Hennon, T., Holmes, L., Wallace, C. A., & Jarvis, J. N. (2016). Limits of Peripheral Blood Mononuclear Cells for Gene Expression-Based Biomarkers in Juvenile Idiopathic Arthritis. Scientific Reports. https://doi.org/10.1038/srep29477

Zolghadr, F., Sadeghizadeh, M., Amirizadeh, N., Hosseinkhani, S., & Nazem, S. (2012). How benzene and its metabolites affect human marrow derived mesenchymal stem cells. Toxicology Letters. https://doi.org/10.1016/j.toxlet.2012.08.015

Appendices

A

Appendix A

Table A.1: Benzene queries performed in INDRA.

Queries	Search terms	PMIDs
query1	benzene[Title] AND ("pharmacokinetics"[MeSH Terms] OR "pharmacokinetics"[Subheading] OR "absorption"[MeSH Terms] OR "distribution"[Title] OR "excretion"[All Fields]) AND (1900/01/01[PDat] : 2020/03/01[PDat])	501
query2	benzene[Title] AND ("Mutation"[MeSH] OR "Cytogenetic Analysis"[MeSH] OR "Mutagens"[MeSH] OR "Oncogenes"[MeSH] OR "Genetic Processes"[MeSH] OR "genomic instability"[MeSH] OR chromosom* OR clastogen* OR "genetic toxicology" OR "strand break" OR "unscheduled DNA synthesis" OR "DNA damage" OR "DNA adducts" OR "SCE" OR "chromatid" OR micronucle* OR mutagen* OR "DNA repair" OR "UDS" OR "DNA fragmentation" OR "DNA cleavage") AND (1900/01/01[PDat] : 2020/03/01[PDat])	740
query3	benzene[Title] AND ("rna"[MeSH] OR "epigenesis, genetic"[MesH] OR rna OR "rna, messenger"[MeSH] OR "rna" OR "messenger rna" OR mrna OR "histones"[MeSH] OR histones OR epigenetic OR miRNA OR methylation) AND (1900/01/01[PDat] : 2020/03/01[PDat])	255
query4	benzene[Title] AND ("reactive oxygen species"[MeSH Terms] OR "reactive oxygen species"[All Fields] OR "oxygen radicals"[All Fields] OR "oxidative stress"[MeSH Terms] OR "oxidative"[All Fields] OR "oxidative stress"[All Fields] OR "free radicals"[All Fields]) AND (1900/01/01[PDat] : 2020/03/01[PDat])	332

Continued on next page

Table A.1 – *Continued from previous page*

Queries	**Search terms**	**PMIDs**
query5	benzene[Title] AND (inflamm* OR immun* OR chemokine OR cytokine OR leukocyte OR white blood cell) AND (1900/01/01[PDat] : 2020/03/01[PDat])	577
query6	benzene[Title] AND ("Hormones, Hormone Substitutes, and Hormone Antagonists"[MeSH] OR "Endocrine Disruptors"[MeSH] OR "Thyroid Hormones"[MeSH] OR "Estrogens"[MeSH] OR "Progesterone"[MeSH] OR "Receptors, Estrogen"[MeSH] OR "Receptors, Androgen"[MeSH] OR "Receptors, Progesterone"[MeSH] OR "Receptors, Thyroid Hormone"[MeSH] OR "Receptors, Aryl Hydrocarbon"[MeSH] OR "Peroxisome Proliferator-Activated Receptors"[MeSH] OR "constitutive androstane receptor"[Supplementary Concept] OR "farnesoid X-activated receptor"[Supplementary Concept] OR "liver X receptor"[Supplementary Concept] OR "Retinoid X Receptors"[MeSH]) AND (1900/01/01[PDat] : 2020/03/01[PDat])	74
query7	benzene[Title] AND ("Cell Transformation, Neoplastic"[MeSH] OR "Cell Proliferation"[MeSH] OR apoptosis OR "necrosis"[MeSH] OR "DNA Replication"[MeSH] OR "Cell Cycle"[MeSH] OR brdu OR thymidine OR angiogenesis) AND (1900/01/01[PDat] : 2020/03/01[PDat])	240