Institute of Biotechnology and Division of Genetics
Faculty of Biological and Environmental Sciences
Doctoral Programme in Integrative Life Science (ILS)
University of Helsinki
Helsinki, Finland

BioMediTech Institute
Faculty of Medicine and Health Technology
Tampere University
Tampere, Finland

Strategies to Improve Standardization and Robustness of Toxicogenomics Data Analysis

Veer Singh Marwah

Academic Dissertation

To be presented, with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, for public examination in Auditorium 107, Athena building Siltavuorenpenger 3 A, Helsinki, on 6th of September 2019, at 12 noon.

Helsinki 2019

Supervisor

Associate Professor Dario Greco
Faculty of Medicine and Health Technology
Tampere University
Tampere, Finland and
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Thesis Advisory Committee Members

Research Director Petri Auvinen
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Ph.D., PI. Ari Pekka Löytynoja
Loytynoja lab
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

D.Sc. Panu Juhani Somervuo
Department of Biosciences
University of Helsinki
Helsinki, Finland

Pre-examiners

Associate Professor Paola Festa
Dipartimento di Matematica e
Applicazioni "Renato Caccioppoli",
University of Naples Federico II
Napoli, Italy

Associate Professor Merja Heinäniemi
Institute for Biomedicine
School of Medicine
University of Eastern Finland
Kuopio, Finland

Opponent

PhD, MBA, CEO Antreas Afantitis
NovaMechanics Ltd.
Nicosia, Cyprus

Custos

Professor Liisa Holm
Institute of Biotechnology and
Department of Biosciences
University of Helsinki
Helsinki, Finland

ISBN 978-951-51-5314-2 (pbk.)
ISBN 978-951-51-5315-9 (PDF)
http://ethesis.helsinki.fi

Unigrafia Oy, Helsinki University Print
Finland 2019

“Why was I chosen?'

'Such questions cannot be answered', said Gandalf. 'You may be sure that it was not for any merit that others do not possess. But you have been chosen, and you must therefore use such strength and heart and wits as you have.”

— J.R.R. Tolkien, The Fellowship of the Ring

“I think it's much more interesting to live not knowing than to have answers which might be wrong. I have approximate answers and possible beliefs and different degrees of uncertainty about different things, but I am not absolutely sure of anything and there are many things I don't know anything about, such as whether it means anything to ask why we're here. I don't have to know an answer. I don't feel frightened not knowing things, by being lost in a mysterious universe without any purpose, which is the way it really is as far as I can tell.”

— Richard P. Feynman

ABSTRACT

Toxicology is the scientific pursuit of identifying and classifying the toxic effect of a substance, as well as exploration and understanding of the adverse effects due to toxic exposure. The toxic effects on human health, biosphere, and ecosystem are essential to maintain public safety in the short and long term. The modern toxicological efforts have been driven by the human industrial exploits in the production of engineered substances with advanced interdisciplinary scientific collaborations. These engineered substances must be carefully tested to ensure public safety. This task is now more challenging than ever with the employment of new classes of chemical compounds, such as the engineered nanomaterials. Toxicological paradigms have been redefined over the decades to be more agile, versatile, and sensitive. On the other hand, the design of toxicological studies has become more complex, and the interpretation of the results is more challenging. Toxicogenomics offers a wealth of data to estimate the gene regulation by inspection of the alterations of many biomolecules (such as DNA, RNA, proteins, and metabolites). The response of functional genes can be used to infer the toxic effects on the biological system resulting in acute or chronic adverse effects. However, the dense data from toxicogenomics studies is difficult to analyze, and the results are difficult to interpret. Toxicogenomic evidence is still not completely integrated into the regulatory framework due to these drawbacks. Nanomaterial properties such as particle size, shape, and structure increase complexity and unique challenges to Nanotoxicology. Furthermore, human endeavors in engineering new nanomaterials with unique properties must be assisted with agile safety nets of toxicogenomics to reduce production costs and ultimately ensure public safety.

This thesis presents the efforts in the standardization of toxicogenomics data by showcasing the potential of omics in nanotoxicology and providing easy to use tools for the analysis, and interpretation of omics data. This work explores two main themes: i) omics experimentation in nanotoxicology and investigation of nanomaterial effect by analysis of the omics data, and ii) the development of analysis pipelines as easy to use tools that bring advanced analytical methods to general users. These tools are defined and fine-tuned by the knowledge from the investigative studies and contain the best practices to ensure reproducibility of the results. An important feature of the omics studies is the reporting of the data and related experimentation such that an independent researcher can interpret it thoroughly. For these purposes, the scientific community has defined standard formats of minimal information required to report the data (MIAME). However, there are areas of improvement in data sharing and reporting. In this work, I explored a potential solution that can ensure effective interpretability and

reproducibility. DNA microarray technology is a well-established research tool to estimate the dynamics of biological molecules with high throughput. The analysis of data from these assays presents many challenges as the study designs are quite complex and contain large cohorts of data points. I explored the challenges of omics data processing and provided bioinformatics solutions to standardize this process. With the application of omics data in toxicology and other fields, it is becoming ever more essential to ensure that the information from the high-throughput data is interpreted correctly. The responses of individual molecules to a given exposure is only partially informative and more sophisticated models, disentangling the complex networks of dynamic molecular interactions, need to be explored. However, this is a technically demanding task. An analytical solution is presented in this thesis to tackle down the challenge of producing robust interpretations of molecular dynamics in biological systems. It allows exploring the substructures in molecular networks underlying mechanisms of molecular adaptation to exposures. I also present here a multi-omics approach to defining the mechanism of action for human cell lines exposed to nanomaterials. The proposed approach can be used to infer long term functional response from relatively short-term exposures. All the methodologies developed in this project for omics data processing and network analysis are implemented as software solutions that are designed to be easily accessible also by users with no expertise in bioinformatics. Our strategies are also developed in an effort to standardize omics data processing and analysis and to promote the use of omics-based evidence in chemical risk assessment.

ACKNOWLEDGMENTS

The work presented in this dissertation was carried out at Finnish Institute of Occupational Health (2015-2016), Institute of Biotechnology, University of Helsinki (2016-2017), and BioMediTech Institute, Faculty of Medicine and Health Technology, Tampere University (2017-2019). This project would not have been possible without the funding support from Academy of Finland (Grant agreements 275151 and 292307), EU H2020 caLIBRAte Project (Grant agreement 686239), EU H2020 LIFEPATH (Grant agreement 633666), and EU FP7 NANOSOLUTIONS Project (Grant agreement FP7-309329).

I would like to thank my supervisor Associate Professor Dario Greco (Tampere University, Tampere, Finland; University of Helsinki, Helsinki, Finland) for the careful and considerate approach in defining the project. His scientific expertise and guidance were invaluable to overcome roadblocks encountered during various important stages of this project.

The thesis advisory committee comprising of Research Director Petri Auvinen (University of Helsinki, Helsinki, Finland), Ph.D., P.I. Ari Pekka Löytynoja (University of Helsinki, Helsinki, Finland), D.Sc. Panu Juhani Somervuo (University of Helsinki, Helsinki, Finland) fulfilled the obligations of their collective and individual roles with great professional courtesy. The communication with the thesis committee was clear, effortless, and productive.

I would like to thank the pre-examiners Associate Professor Paola Festa (Dipartimento di Matematica e Applicazioni "Renato Caccioppoli", University of Naples Federico II, Napoli, Italy) and Associate Professor Merja Heinäniemi (Institute for Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, Finland) for their constructive critique. Their feedback played an essential role in improving the manuscript for better clarity and for distilling the message of the dissertation.

My gratitude and thanks to Professor Juha Partanen, and Professor Pekka Heino for their considerate and kind demeanor in fulfilling their administrative roles during this dissertation.

I cannot envision the completion of this doctoral dissertation without the financial support provided by Professor Harri Alenius (University of Helsinki, Helsinki, Finland; Karolinska Institute, Stockholm, Sweden), and M.D. Ph.D. Antti Lauerma (University of Helsinki, Helsinki, Finland). I am humbled by their gracious scientific support and collaborative spirit.

This doctoral dissertation benefited from the scientific influence of Professor Kai Savolainen. His scientific experience and contribution in the field of Nanotoxicology laid the foundation of this project.

I am grateful for the mentoring provided by Assistant Professor Vittorio Fortino (University of Eastern Finland, Kuopio, Finland) in the capacity of PostDoc at GrecoLab. His Bioinformatics expertise and active leadership in explaining study design and establishing work ethics were valuable for my understanding of research guidelines. He actively collaborated in identification and implementation of statistical concepts, development of methods, and scientific discourse towards improvements of ongoing studies.

My sincere thanks to Ph.D. Pia Anneli Sofia Kinaret (University of Helsinki, Helsinki, Finland) for her expertise in the lab for biological assay experimentation, the scientific discourse of the biological concepts pertaining to study designs, and interpretation of results as well as scientific reporting.

I relished the opportunity to work with Ph.D. Giovanni Scala (Tampere University, Tampere, Finland) and I am grateful for his diligence towards statistically sound approaches for accurate processing and interpretation of biological assay data. His attention to detail and persistence to establish and calibrate benchmarks, and his impetus to identify scientific rationale as well as principles for various methodologies employed and analytical tasks performed in the studies.

This doctoral dissertation incorporates many scientific studies which could not have been completed without the specific scientific contributions of the following researchers. Ph.D. Angela Serra (Tampere University, Tampere, Finland) for the scientific critique, bioinformatics discourse, and collaborative contributions; Ph.D. Jukka Sund (European Commission, Brussels, Belgium) for is technical expertise in the laboratory and performing biological assays.

I would also like to thank the following researchers for their collaboration during various studies conducted in this project, Ph.D. Nanna Fyhrquist (University of Helsinki, Helsinki, Finland; Karolinska Institute, Stockholm, Sweden), Ph.D. Marit Ilves (University of Helsinki, Finland), and Ph.D. Lasse Ruokolainen (University of Helsinki, Helsinki, Finland).

Finally, thanks and salutations to GrecoLab members (Ph.D. Antonio Federico, and M.Sc. Laura Saarimaki) for the scientific support, encouragements and most importantly for their sincere critique.

Veer Singh Marwah

Helsinki, 2019

CONTENTS

Abstract 4
Acknowledgments 6
Contents 8
List of Original Publications 10
Abbreviations 12
1 Introduction 16
2 Traditional Toxicology 21
3 Toxicogenomics 23
4 Systems Toxicology 28
4.1 Pathways Based Toxicity Evaluation 29
5 Nanotoxicology 31
5.1 Nanoparticle interaction with the biological system 31
5.2 Nanotoxicological idiosyncrasies 32
6 Co-expression pattern of the molecular mechanism 34
7 Aims of the Thesis 37
8 Materials and Methods 38
8.1 Omics Data Preprocessing 38
8.1.1 Quality Control 38
8.1.2 Filtering 38
8.1.3 Normalization 39
8.1.4 Batch Correction 39
8.1.5 Differential Analysis 40
8.2 Network Inference 40
8.3 Centrality based Gene Ranking 41
8.4 Responsive Subnetwork 41
8.4.1 Module Detection 41
8.4.2 Characterization & Functional Assessment 41
8.4.3 Reconstruction/Merging 42
8.5 Functional Characterization 42
8.6 Integrative Analysis / Multi-omics 43
8.7 Defining Mechanism of Action 43

8.8 Bioinformatics tool implementation with graphical user interface 44
8.8.1 eUTOPIA 44
8.8.2 INfORM 45
9 Results 46
9.1 Standardization in omics reporting 46
9.2 Reproducibility and robustness of toxicogenomics data analysis 47
9.2.1 Data preprocessing 47
9.2.2 Molecular Systems Analysis 49
Network Inference 49
Robustness by consensus 49
Summarizing co-expression scores 50
Consensus by the significance of evidence 50
Modular component of the inferred network 54
Response module 55
Molecular mechanisms 58
9.2.3 Software implementation 60
Omics preprocessing 60
Molecular networks 61
9.3 Multi-omics based approach to modeling AOPs 62
10 Discussion 67
11 Conclusions 71
12 References 73

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications:

I Scala, G., **Marwah, V.**, Kinaret, P., Sund, J., Fortino, V., and Greco, D. (2018). Integration of genome-wide mRNA and miRNA expression, and DNA methylation data of three cell lines exposed to ten carbon nanomaterials. Data Brief *19*, 1046–1057.

II **Marwah, V.S.**, Scala, G., Kinaret, P.A.S., Serra, A., Alenius, H., Fortino, V., and Greco, D. (2019). eUTOPIA: solUTion for Omics data PreprocessIng and Analysis. Source Code for Biology and Medicine *14*, 1.

III **Marwah, V.S.**, Kinaret, P.A.S., Serra, A., Scala, G., Lauerma, A., Fortino, V., and Greco, D. (2018). INfORM: Inference of NetwOrk Response Modules. Bioinformatics *34*, 2136–2138.

IV Scala, G.*, Kinaret, P*., **Marwah, V.**, Sund, J., Fortino, V., and Greco, D. (2018). Multi-omics analysis of ten carbon nanomaterials effects highlights cell type specific patterns of molecular regulation and adaptation. NanoImpact *11*, 99–108.

* *Equal contributions*

The publications are referred to in the text by their Roman numerals.

Candidate contributions:

I Defined standards for data and experimental procedures reporting; drafted the manuscript.

II Defined the methodological strategies, implemented the algorithms, developed the software, conducted the case study analysis, drafted the manuscript.

III Defined the methodological strategies, implemented the algorithms, developed the software, conducted the case study analysis, drafted the manuscript.

IV Performed the data analysis, participated in the interpretation of the results, drafted the manuscript.

ABBREVIATIONS

3Rs	Refine, Reduce, and Replace
AOP	Adverse Outcome Pathway
AOP-KB	Adverse Outcome Pathway - Knowledgebase
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
AUPR20	Area Under Precision Recall with 20% high confidence
MRNET	Minimum Redundancy NETworks
MRNETb	Minimum Redundancy NETworks using Backward elimination
CHT_MW	Multiwalled Carbon Nanotube (cheaptubes)
CLR	Context Likelihood or Relatedness
CNM	Carbon Nanomaterials
BEAS-2B	adenovirus-12 SV40 hybrid virus transformed bronchial epithelial cells
BMD	Benchmark Dose
BMDL	Benchmark Dose (lower confidence limit)
CRAN	Comprehensive R Archive Network
DNA	Deoxyribonucleic Acid
DREAM	Dialogue on Reverse Engineering Assessment and Methods
ECHA	European Chemical Agency
EPA	U.S. Environmental Protection Agency
eUTOPIA	solUTion for Omics data PreprocessIng and Analysis
FDA	Food and Drug Administration
FP	False Positive
GE	Gene Expression
GEO	Gene Expression Omnibus
GIF	Graphics Interchange Format

GO	Gene Ontology
GSE	GEO Series
GSEA	Gene Set Enrichment Analysis
GUI	Graphical User Interface
INfORM	Inference of NetwOrk Response Modules
ISA-TAB	Investigation / Study / Assay Tabular
ISA-TAB-Nano	Investigation / Study / Assay Nanomaterial Tabular
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC50	Lethal Concentration, 50%
LD50	Lethal Dose, 50%
logFC	Logarithm Fold Change
LINCS	Library of Integrated Network-based Cellular Signature
MAGE-ML	Microarray Gene Expression Markup Language
MAGE-OM	Microarray Gene Expression Object Model
MAGE-TAB	Microarray Gene Expression Tabular
MANTRA	Mode of Action by NeTwoRk Analysis
MDS	Multidimensional Scaling
MIAME	Minimum Information About a Microarray Experiment
miRNA	micro Ribonucleic Acid
MIT_MW	Multiwalled Carbon Nanotube (Mitsui)
MOA	Mechanism of Action
MPM	Malignant Pleural Mesothelioma
mRNA	messenger Ribonucleic Acid
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NCI-60	National Cancer Institute 60 human cancer cell lines
NOAEL	No Observed Adverse Effect Level
OECD	Organization for Economic Cooperation and Development
PDF	Portable Document Format
POD	Points of Departure

POT	Pathways of Toxicity
PRT	Prototype Ranked list
PTGS	Predictive Toxicogenomics Space
QC	Quality Control
QSAR	Quantitative Structure-Activity Relationship
REACH	Registration, Evaluation, Authorisation, and restriction of CHemicals
RfC	Reference Concentration
RfD	Reference Dose
RNA	Ribonucleic Acid
SES-SW	Singlewalled Carbon Nanotube (SES)
SIG-MW	Multiwalled Carbon Nanotube (Sigma)
SIG-SW	Singlewalled Carbon Nanotube (Sigma)
SMITE	Significance-based Modules Integrating the Transcriptome and Epigenome
TCE	Trichloroethylene
TGP	Toxicogenomics Project
TG-GATEs	Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system
THP-1	Tamm-Horsfall Protein 1
TP	True Positive
TSS	Transcription Start Site
UI	User Interface
UML	Unified Modelling Language
XML	Extensible Markup Language

1 INTRODUCTION

Toxicology is the study of toxins/poisons and their harmful effects on human health or the environment. It is, in fact, an effort to characterize chemicals and other xenobiotic substances present in the environmental or specific exposure. However, such chemicals can be useful in some specific formulations, type of exposure, and exposure durations. Known toxins have shown promise as a drug for effective treatment of specific ailments (Cury and Picolo, 2006), while drugs in specific dosage, incorrect exposure, or due to the genetic makeup of recipient can be extremely toxic and life-threatening (Nakayama *et al.*, 2009). Furthermore, pollutants or side products of modern human activities can, in turn, produce acute and chronic toxicity (Roux *et al.*, 2002). These different dimensions of human contact with chemicals and other xenobiotics require a cautious and meticulous approach to characterize and classify toxic substances, toxicological effects, and mode of action; thus toxicity testing is an important area of research. However, the need for toxicity testing has not always been apparent. The early 20th-century boom in the chemical industry led to the production of various substances for human consumption and some serious, tragic events (Paine, 2017) resulted in the establishment of toxicity testing in animals for evidence of safety to determine whether the substance is a risk to public health. Toxicity testing standards have been refined over the last century for food additive and cosmetics, drugs, and environmental pollutants such as pesticides, industrial chemical waste, and residues from other products (Ridings, 2013). These standards have been placed into effects by various initiatives such as Food and Drug Administration (FDA), U.S. Environmental Protection Agency (EPA), Organization for Economic Cooperation and Development (OECD), and European Chemical Agency (ECHA).

Over the decades, the limitations of traditional toxicity screening methods have been realized, and improvement in experimental techniques have allowed for further reform in testing strategies. Traditional animal assays focus on apical endpoints in the whole organism, requiring the sacrifice of many animal subjects in a battery of tests. The 3Rs framework (refine, reduce, and replace) was proposed by William Russell and Rex Burch in 1959 to refine the use of animals without discouraging the scientific pursuit. Refinement is defined by the use of methods and technique to minimize the pain and suffering of test subjects, by the introduction of less intense experimental procedures. This concept led to a reduction of negative effects but also the enhancements in the welfare of the animals by improving their living conditions. Reduction is defined as improvements in the designs of scientific studies to reduce animal test subjects. Replacement is defined as the replacement of vertebrates by invertebrate subjects or absence of sentient

animal testing by employing in-vitro methods, microbiological studies, or early stage development embryos. Advanced techniques in high content and high throughput screening of in vivo and in vitro samples have allowed for testing approaches that are more cost and time effective. The modern testing strategies are capable of evaluating molecular mechanisms in a variety of different scenarios, such as neurotoxicity effects in tissue-specific fashion (Figure 1).

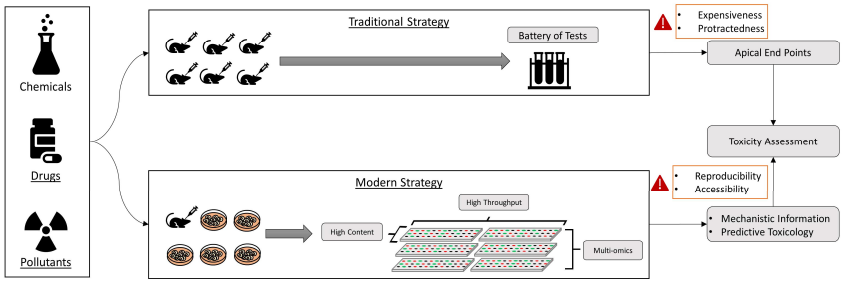

Fig. 1 - Traditional toxicology strategy of apical endpoint assessment versus the modern strategy of understanding molecular mechanisms.

Access to mechanistic information from toxicity studies have opened new avenues of predictive toxicology, and scientific perspective has quickly changed to a mode of action based approaches. Toxicity pathways are enabling to build a toxicological landscape as a toxome (Bouhifd *et al.*, 2014; Bouhifd *et al.*, 2015) that can be used to classify and predict toxic substances efficiently. The combined scientific effort in omics data generation and analysis from different molecular species is revolutionizing the regulatory framework. This information is giving a possibility to build a complete picture of the biological mechanisms involved in toxicological response and is allowing for new ways to model this information for predictive toxicology.

However, there is some translational gap between the information obtained from the experimental data and its implementation in regulatory decision making. This thesis presents the scientific efforts to alleviate some of those limitations.

Data generated from omics experiments have defined formats that are specific to the raw data and do not entertain the biological relevance of the dataset. This raw data format includes only limited information concerning the experimental specifications used in generating the data, thus jeopardizing the reproducibility in follow up studies. The scientific community has proposed solutions for collecting and sharing omics assay raw data and associated metadata with sample description, study design, and experimental setups. MIAME (Minimum Information About a Microarray Experiment) standard was defined to record and report the minimum information required for interpretation of microarray data (Brazma *et al.*,

2001). MAGE-OM (Microarray Gene Expression Object Model) defined in UML (Unified modeling language), and MAGE-ML defined in XML are MIAME-compliant formats with a structured approach to facilitate the exchange of microarray data (Spellman *et al.*, 2002). MAGE-TAB (MicroArray Gene Expression Tabular) (Rayner *et al.*, 2006) is a simplified spreadsheet-based format proposed as an alternative to intensively complex MAGE-ML format. ISA-TAB (Investigation/Study/Assay TAB-delimited) was modeled on MAGE-TAB as a general purpose framework to communicate complex metadata from experiments that employ a combination of technologies such as genomics, transcriptomics, proteomics, metabolomics, (Rocca-Serra *et al.*, 2010). ISA-TAB-Nano (Investigation/Study/Assay Nanomaterial TAB-delimited) (Thomas *et al.*, 2013) extends the ISA-TAB format with the information of the material. Public repositories such as GEO (Barrett *et al.*, 2013) have been established to promote reporting and exchange of experimental data. However, there is no widely accepted repository for the ISA-TAB-Nano format reporting of data, limiting its usability. These formats are all designed to report the raw data and ensure independent inspection and interpretation to facilitate reproducibility as well as the exchange of data. One major lacking feature of these reporting formats is the absence of data analysis information. It can be argued that the complete interpretation of data is not possible without understanding intermediate analysis results. The analysis results must be reported with complete clarity of the methodology employed in analyzing the data along with the rationale for its use. The data analysis must be made independently reproducible by sharing the tools and bioinformatics scripts employed, thus ensuring that efforts from third parties can effectively produce the desired results from the very same data, or in a similarly designed independent study. This thesis presents an effort to report the data and analysis results from toxicogenomics study of nanomaterials.

One of the major concerns with omics data is the reproducibility of data and analytics. We propose here a solution for standardization of omics data analysis generated from microarray experiments by using state-of-the-art methods for data analysis in a standardized analysis workflow that is intuitive and easy to use. Such a solution can bring the technology closer to all users and enables to generate reproducible results.

Exploration of publicly available microarray data and results brings to light that raw omics data can be very noisy and may contain a poor estimation of signal for some arrays, and samples. Identification of these poor samples as outliers and removal can improve the signal-to-noise ratio (Kauffmann and Huber, 2010). Systemic data biases in large microarray datasets must be addressed by using data normalization methods before any quantitative comparison of microarray features (Bilban *et al.*, 2002). The unwanted variation can be observed in the integrated microarray datasets due to batch effects (Lazar *et al.*, 2013). These batch effects represent non-biological variation that may be biased towards particular conditions leading

to unreliable comparisons. Exploration and diagnosis of data must be performed to identify batch effects, and appropriate adjustment methods must be used for correction.

Increasing interest is devoted to the possibility to merge multiple omics data sets and re-analyze them together to increase statistical power. In such cases, the evaluation and correction of technical batch effects become crucial. Without proper processing of the raw assay values, it is impossible to investigate, identify, and remove the noise. Neglecting this step can result in type one and type two errors that could go unnoticed. Most tools for omics data analysis place a barrier of required computational expertise that either disqualifies some researchers or place them at the risk of producing erroneous results with faulty assessments. Reliability and robustness of the toxicogenomics data analysis results can be achieved in part by ensuring that noise in the assay data is estimated and avoided systematically.

Preprocessing of microarray data must check for biases and imbalances discussed above. The scientific community has actively pursued this challenge providing various packages for data correction (Lazar *et al.*, 2013). The most effective methods and guidelines have been greatly debated by comparing the performances of the proposed methods resulting in a set of best practices to ensure robustness and reliability of analysis results. Following topics discuss these best practices in brief.

The applicability of any set of tools in an efficient workflow is determined by the ease of communication between them, at the very least. R statistical programming language is one of the widely accepted and actively used platforms for development and dissemination of tools that enable bioinformatics analysis of biological assay datasets. The statistical toolset provides the appropriate platform for the development and implementation of complex algorithms. The vast library of packages in Bioconductor and CRAN public resources catered to the biological research is a testament to R's importance. The R language platform has been used to develop numerous packages for omics analysis, which have been tested and evaluated by the scientific community. Thus, I chose the R language platform for identification of gold standard tools and implementation of widely accepted best practices in omics data analysis.

Likewise, effective interpretation of omics data often represents an additional challenge. Molecular mechanisms of toxicity are not just sets of molecules dysregulated in a toxic response, but their complex patterns of interaction. A biological system is incomplete without the understanding of its molecular relationships, which can be successfully modeled by means of graph theory. It is not trivial to infer these relationships and reconstruct the system information from omics measurement data, and it is not always simple to be orientated among multiple methodological solutions. I demonstrate in this thesis the capabilities of a solution that ensures robust

inference of gene-gene co-expression network. The end-user can accomplish this task via an intuitive and easy to use graphical interface.

The multi-disciplinary nature of omics methods and analytics is often regarded as an additional obstacle for their widespread use in multiple toxicology research environments and, consequently, hampers their full implementation in regulatory assessment. The work presented in this thesis also addresses this critical issue, by the development of software solutions that can help scientists with no specific knowledge in computer science to successfully transform omics data into sensible biological knowledge while ensuring robustness and reproducibility. Finally, I showcase here a multi-omics study to model the dynamics of mechanistic information from multiple molecular species and its effectiveness in toxicity evaluation.

2 TRADITIONAL TOXICOLOGY

Traditional toxicology estimates toxic response by exposing test subjects (in vivo or in vitro) with different dosages of a substance for different time durations and observe the apical endpoint as a phenotypic change producing an adverse effect. The dosage of a substance resulting in death of 50% of the population in a defined period is known as LD50 (Lethal Dose, 50%) (LeBeau, 1983) and the concentration of the substance in mg/l capable of killing 50% of the population is known as LC50 (Lethal Concentration, 50%). These bioassays involve groups of animal replicates exposed to different concentration of the substance. The mortality rates are observed for different exposures, these data points are plotted as a graph, and LC50 is inferred from the graphical representation. No-Observed-Adverse-Effect-Level (NOAEL) (Crump, 1984) is the highest tested level of the substance that does not produce any adverse effects. It is a measure of the dose-response assessment that denotes the statistically or biologically significant maximal level of dose with no adverse effects. Reference Dose (RfD) (Barnes and Dourson, 1988) for oral exposure and Reference Concentration (RfC) for the inhalation exposure are the corrected NOAELs by a uncertainty/safety factor (Dankovic *et al.*, 2015) to balance for various aspects of experimental values, such as interspecies variability and variability in human response. However, statistical drawbacks in NOAELs led to the development of alternative approaches. A benchmark dose approach (BMD) is used to measure dose-response by modeling dose levels against the response level to identify the point of departure (POD. Dose-response model fitted on experimental data is used to measure benchmark dose (BMD) (Filipsson *et al.*, 2003) which is a significant increase in risk (10 % response) compared to background risk. BMDL is the corresponding 95% lower limit. BMD methodology can be performed at much lower sample size while NOAEL, requires many more data points to be sampled to precisely identify the corresponding dose to LOAEL (Lowest-Observed-Adverse-Effect-Level) and NOAEL. BMD is not dependent on specific dose concentrations, and it can be reliably identified from the shape of the dose-response model curve (Davis *et al.*, 2011). BMD is accepted as the preferred method by EPA.

Measurements from these bioassays have been successfully used over the decades to estimate toxicity. However, the traditional toxicology approach neglects the intermediate molecular and cellular changes leading to observable phenotypic changes occurring in the exposed individuals. Thus, we are left in the dark concerning the molecular mechanism of action of the exposure. Hence, every substance to come in human contact must be tested for its toxicity by means of bioassays and a battery of tests, this is not sustainable, as these testing methods are expensive, time-consuming, and

require many test subjects. There is a need for predictive methods for early identification of possibly toxic substances to reduce testing. Quantitative structure-activity relationship (QSAR) (Dunn, 1988; Phillips *et al.*, 1990; Dearden, 2003) methods can be used to classify substances based on their activity and their physicochemical properties. Prediction from these methods suffers from a lack of absolute correlation of the biological response with the molecular descriptors. The biological response evidence is experiment dependent and can change upon many factors. The adverse outcome of toxic exposure is the result of a substance interacting with the biological system in a specific manner by perturbing biological molecules and pathways; thus the prediction cannot be accurate in the absence of information from within the biological system. Although the categorization and grouping of similar toxicants can be employed to predict activity in the absence of test data (van Leeuwen *et al.*, 2009). The read-across technique takes advantage of the groups to suggest toxicological effects and can utilize toxicogenomics data from different apical endpoints to extrapolate predictions (*in vitro*, *in vivo*) (Schultz *et al.*, 2015). Furthermore, the concept of integrated testing can be utilized to address the drawbacks by combining complementary pieces of evidence (Ahlers *et al.*, 2008).

3 TOXICOGENOMICS

Application of genomic technologies such as genome sequence analysis, gene expression profiling, proteomics, and metabolomics in toxicological assessment is referred to as Toxicogenomics. Highly dense information generated from these techniques is combined with the toxicological effects on the phenotype of the exposed biological systems. Toxicogenomics has the potential to be more sensitive and allows for more accurate prediction of adverse effects due to toxic exposure (Krewski *et al.*, 2010).

Toxicogenomics developed as the tools of pharmacogenetics began to be applied to toxicology questions. One of the first intuition of the potential utility of toxicogenomics was the discovery during the Korean War, as soldiers of specific ethnic backgrounds developed severe hemolysis during anti-malarial treatment with primaquine (Alving *et al.*, 1960). The acute hemolytic anemia is observed in individuals with glucose-6-phosphate dehydrogenase deficiency (G6PDd) depending on the dose administered (Beutler, 1994). To date, genotyping can be done to determine this risk. Currently, in drug development, toxicogenomics is used to investigate, for instance, the mechanisms of toxicity and to predict the hazard of new drugs, while in clinical medicine toxicogenomics is used to identify patients at risk for adverse drug reactions (Rouquié *et al.*, 2015). Toxicogenomics is also used in the context of occupational exposure as a genetic variation may predispose some workers to develop disease due to specific workplace exposures (Ventura *et al.*, 2018). Genetic polymorphisms studies provide some information on the risk of disease or toxicity with specific occupational exposures and known potentially significant gene-environment interaction.

Stratification of individuals by genetic variations is vital to understand the genotype-specific health risk and predisposition to adverse effects. Genetic variants are known to be associated with opiate metabolism and toxicity (Kosarac *et al.*, 2009), genotype information can identify predisposition of risk to organophosphate toxicity (Costa *et al.*, 2013). Trichloroethylene (TCE) is an industrial solvent used as a chemical intermediate for the production of other compounds; it is known to cause several adverse effects, it is a human carcinogen; it is hepatotoxic, nephrotoxic, neurotoxic, immunotoxic; and causes fetal malformation (Chiu *et al.*, 2013). There are known genetic risk factors that are found to induce hypersensitivity dermatitis reactions (Dai *et al.*, 2015). Industrial exposure to asbestos produces adverse effects in association with genetic factors that increase susceptibility to lung cancer (Liu *et al.*, 2015b) and genetic factors may also play a role in malignant pleural mesothelioma (MPM) susceptibility (Tunesi *et al.*, 2015).

Toxicogenomics can be used in preventing medication toxicity and understanding the mechanism of toxic response to medication. Many drugs on the market are labeled with information regarding genetic polymorphisms and their association with adverse effects (Schuck and Grillo, 2016). Patient genotype data can be used to understand cases of acute toxicity after medication, toxic response due to chronic exposure to the drug in the presence of specific genetic polymorphisms. Drug metabolism is affected by the polymorphism that influences the drug metabolizing enzymes which could cause the patient to poorly metabolize the drug or be susceptible to adverse effect due to high drug dosage (Shenfield, 2004). Use of toxicogenomics in the context of medication toxicities should enable to determine the cause of adverse effects. Scientific studies have showcased the effectiveness and benefit of pharmacogenetic testing (Jorgensen *et al.*, 2019), and the implementation of such testing clinics has been evaluated (Verbelen *et al.*, 2017). Still, there is room for improvement in establishing pharmacogenetic testing as generally accepted service offered in clinics (Haga and Kantor, 2018).

Toxicological evaluation of chemical exposure to an adverse outcome is traditionally performed by observing the apical endpoints such as a phenotypic change in the organism or cell death, but this does not provide the understanding of the molecular basis of the perturbed biological system. Evaluation of the effects of toxins exposure on the transcriptome became possible with the development of DNA microarray technologies in the 1990s (Schena *et al.*, 1995); thus, the toxicogenomics field progressed rapidly. Profiling the molecular behavior during steady state and the perturbed state affords us a picture of the molecular mechanisms of action that produce the adverse outcome. Signatures of molecular response can be identified and used to predict the toxicity of a chemical or an adverse effect of a drug by using the profile of known toxic substances. Omics studies are conducted in animal models with traditional apical endpoints; alternately, the in vitro study designs use cell and tissue cultures (Collins *et al.*, 2017; LeCluyse *et al.*, 2012) to identify the adverse response that alters the biological steady state and disrupts the biological pathways by modelling the molecular response in defined durations of exposure and dosages (Adeleye *et al.*, 2015). The progressive approach of toxicology utilizes the molecular profile to identify the adverse events and significant perturbation of the defined pathway of toxicity (Brockmeier *et al.*, 2017). The in vitro screening assay data and results are assimilated in various repositories, such as ToxCast, which can be utilized for predictive toxicology by means computational methods (Knudsen *et al.*, 2015). Technological advancements have made it possible to observe the molecular changes in the whole genome and evaluation of manifold substances can be performed simultaneously / in parallel to generate a large amount of data in short time allowing to generate better response profiles from in vivo and in vitro studies. In turn, these profiles can be used to characterize chemicals and drugs against known substances, which can help

to identify toxic chemicals and drug response and efficacy at a faster pace to keep up with modern demands of the drug development cycle and chemical products. Most popularly, the genome-wide evaluation of molecular alterations is performed at the gene expression level, where the mRNA expression is observed to portray the changes in the molecular mechanisms that can be used to predict the phenotypic changes and adverse outcomes. Transcriptomics benefits from experimental techniques that are cost effective and take less time to generate the high density of data; thus it has been exploited in different biological studies like disease mechanism, preclinical studies, drug discovery, and toxicology. Genome-wide evaluation can be performed for other molecular species to get a different set of mechanistic information. Epigenomics deals with DNA structure modification, DNA protein interaction, and RNA expression (Friedman and Rando, 2015). Global patterns of methylation and chromatin modification can be used to understand the regulatory mechanisms behind changes in gene expression that preserve longer (Limonciel *et al.*, 2018). Proteomics technologies, such as mass spectrometry and protein microarrays, can be used to identify proteins and protein complexes, functional characterization and proteome-wide changes in toxic exposure and epidemiological effects (Merrick and Witzmann, 2009). Metabolomics technologies allow measuring the small molecules that are produced from the metabolic processes. It can be performed in biofluids extracted from the subject, thus providing a non-invasive method of performing repeated measurements to get metabolite dynamics in the exposure-response curve (Bouhifd *et al.*, 2013).

Transcriptomics has played an important role in the preclinical studies and drug development by giving an insight into the molecular mechanisms involved and the mode of action for drug efficacy and possible adverse effects in a dose and time-dependent manner. Gene expression profiles also allow for the identification and characterization of biomarkers in the preclinical studies (Joseph, 2017; Te *et al.*, 2016). Furthermore, these profiles can be used to identify molecular signatures or fingerprints for classification of chemicals toxicity by the specific adverse outcome in specific organs (Kim *et al.*, 2015).

Extensive resources for archival and reporting of gene expression data have been established over the years, providing expression signature information of compounds for toxicological and pathological endpoints. These resources can be used to characterize also novel compounds and drugs. The Connectivity Map (Lamb, 2007) is an extensive reference catalog of gene expression data generated from perturbation studies performed with chemicals and genetic reagents on cultured human cells. This compendium of gene expression data is used to define functional connections between disease-gene-drug, which can be used to characterize novel chemicals and identify new drug candidates. The Library of Integrated Network-based Cellular Signature (LINCS) L1000 (Liu *et al.*, 2015a) is a resource housing expression profiles induced by compounds. It allows for the discovery of

compound signatures and profiling of compounds for their drug-like qualities and toxicity. The Japanese Toxicogenomics Project consortium (TGP) has developed a large-scale toxicogenomics database of gene expression profiles and traditional toxicological data generated from in vivo and in vitro studies for exposure to 170 compounds at multiple dosages and time points. The Open TG-GATEs (Igarashi *et al.*, 2015) (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) has applications in drug safety assessment as toxicity assessment is required to be performed in test animals and cell cultured during the preclinical stage of drug development. Reference gene expression data from TG-GATEs provides the mechanistic understanding of specific toxicities to identify biomarker signatures, and it can be exploited for predictive toxicology.

Data from gene expression consortiums have been used to develop computational methods and tools with predictive capabilities to characterize novel chemicals and drugs.

Computational approaches and statistical methods have been used effectively in clustering gene expression profiles with clusters found to be significantly correlated with histopathological and clinical chemistry evidence of toxicity (Waring *et al.*, 2001) to show the applicability of transcriptomics in diagnostics. Predictive Toxicogenomics Space (PTGS) (Kohonen *et al.*, 2017) is an effort to predict unanticipated harmful effects of chemicals and drug molecules. PTGS is generated by applying a compacting modeling approach on the gene space of the Connectivity Map, and the resulting component space is fused with the cytotoxicity data from NCI-60 tumor cell line screens. MANTRA 2.0 (Carrella *et al.*, 2014) is a tool from the prediction of drug mode of action, and for drug repurposing, it uses gene expression profiles from Connectivity Map. It infers a network of drugs by obtaining a 'prototype' ranked list (PRT) of differentially expressed genes from drug treatment studies performed in multiple cell lines and at different drug dosage. Gene Set Enrichment Analysis (GSEA) is used to compute the similarity between PRT from two drugs, represented inversely as a property of the edge connecting the two drugs such that similar drugs are closer than the dissimilar ones. Exploration of various methodologies for prediction of drug sensitivity has been performed by DREAM (Dialogue on Reverse Engineering Assessment and Methods) in collaboration with NCI (National Cancer Institute) as a benchmark study which highlighted Bayesian multitask MKL (multiple kernel learning) as the best performing algorithm out of 44 drug sensitivity prediction algorithms (Costello *et al.*, 2014).

Toxicogenomics gives a tremendous advantage over traditional toxicity (Chepelev *et al.*, 2015), modern toxicological studies generate data at higher throughput, this allows for modeling of data to create profiles for characterization of novel substances. It gives better mechanistic information (Tyner, 2017) of the molecular behavior during toxic exposure allowing for

accurate toxicological assessment, and it is much more cost effective in comparison to traditional toxicology.

The advent of high content and high throughput omics techniques has made it possible to infer a snapshot of the perturbed biological system, thus opening new possibilities of better and more accurate predictive methods. One important class of substances that can benefit from the Toxicogenomics approach are nanoparticles. The unique properties of nanoparticles (Murty *et al.*, 2013) have been used to design better composite substances leading to their introduction in various products such as cosmetics, toys, electronics, sports goods, personal care products, textiles, food, and beverages. Carbon nanotubes (CNT) exhibit remarkable elasticity and tensile strength due to the smaller diameter and larger surface area. CNTs can be used to attribute strength as an additive to a composite material. Carbon black has been used as an additive to reinforce rubber in tyre manufacture. Titanium dioxide nanoparticles exhibit high ultraviolet (UV) absorption and are transparent and are employed in the formulation of sunscreens, on the contrary, bulk Titanium/titania/ titanium white is used as white pigment/dye and does not exhibit UV absorption properties. Gold nanoparticles exhibit optical and electrical properties that are employed in nanobiotechnology for cellular imaging, and it can be functionalized and employed in therapeutics to deliver drugs. Toxicogenomic assessment of nanomaterial toxicity can identify their mode of action, which can be correlated with the nanomaterial intrinsic properties (Kinaret *et al.*, 2017a). The systematic effect association with nanomaterial intrinsic properties can be used for the designing and creation of nanomaterials that avoid triggering any adverse effects while retaining their unique properties and making them safe by design (Simeonova and Erdely, 2009).

The omics data can be corrected, filtered, modeled, and transformed to highlight biologically significant events. The highly interpretive and predictive nature of toxicogenomics data makes it challenging to report the findings and describe the data in a manner that is meaningful and conclusive. Thus, bringing to light the question of data reproducibility, as it can be quite challenging to obtain the same results in repeated measurements or similarly designed studies because of the complexity and many experiment variables. Furthermore, the data is generated by different techniques which have many different instrumentations and methodologies and generate data in a variety of different raw data format, making the analytical process a further complicated step. These raw data sets can be analyzed by many different computational methods and tools, which are proposed by researchers with proof of concept evidence that speaks to the merit of these alternative choices. It becomes evident that streamlined workflows are needed to ensure ease of data processing and maintain reproducibility of results.

4 SYSTEMS TOXICOLOGY

The focus of Toxicogenomics is understanding the mechanism of action of various types of biomolecules as a constituent of resulting adverse effect due to toxic exposure. While traditional toxicology investigates the apical endpoints in relation to the intrinsic properties of the toxic substance. Systems toxicology approach assimilates the mechanistic information from the omics experimentation with the relevant intrinsic properties and utilizes data modeling for a more comprehensive understanding of the toxicological response.

Systems toxicology is the application of systems biology in evaluating the adverse biological effects of xenobiotics. Adverse response to a substance can be studied with a study design that takes into account system-level effects in different tissues to understand the tissue-specific responses as well as common effect. Furthermore, the assays performed in a different set of time points after exposure can be used to add another dimension of information that can determine acute and long-lasting effects. The information from the different conditional assays can be used to understand the effect on the biological system.

In a systems toxicology frame, for instance, the intrinsic properties of the exposure can be combined with the complex alterations taking place in the exposed biological systems to build comprehensive predictive models (Wang *et al.*, 2018).

Systems toxicology models the network of interactions between the molecules to represent the emergent response of the biological systems due to the correlated mechanisms of the molecules likely to be involved in the same biological pathways and processes, thus more accurately representing the response of the perturbed system (Wu *et al.*, 2018).

At the turn of the century, there was recognition and need for improving the toxicity testing as the traditional methods are very costly and come at the expense of animal health and welfare (Hartung, 2011) and do not have suitable predictive features. Computational toxicology in modern 21st century incorporates aspects of bioinformatics, chemo-informatics, with the growing need for biokinetics modeling such as the use of physiologically based pharmacokinetic time models (Lipscomb *et al.*, 2012). These approaches also make use of the existing databases that contain the latest information, some of which are publicly available.

Systems toxicology models the cascade of events underlying the direct action of the xenobiotic exposure and toxicity endpoints. It can further model the indirect response of the xenobiotic exposure that might lead to an

adverse outcome and can be observed as the emergent mechanisms from the systemic response and signal cascade.

Nascent Toxicogenomics evaluates the molecular features as a set of independently acting responses that represent stress of toxicity exposure at the various dosage and exposure durations. This forms a straightforward path from exposure to expression profile with respect to the adverse effect. In order to properly understand the molecular mechanisms, the biological system must be modeled to represent molecular interactions that can be combined with toxicological parameters and adverse effect (Barel and Herwig, 2018; Mulas *et al.*, 2017; Ventura *et al.*, 2018; Yamane *et al.*, 2016).

4.1 PATHWAYS BASED TOXICITY EVALUATION

Mechanistic information from high throughput and high content screening techniques allow for the representation of the toxicity systematically by using pathways. The concept of the pathway of toxicity emerged where instead of focusing on apical endpoints in organisms, changes could be observed at the cellular and molecular level that would be predictive of adverse outcomes. These evaluations could be performed without the use of animals while also allowing for much higher throughput, thus keep in pace with the thousands of chemicals that are being introduced each year.

Pathways of Toxicity (PoT) (Kleensang *et al.*, 2014) can be used to get a defined set of pathways encompassing various molecular events that can be employed in toxicity regulation decision making. A human toxome (Bouhifd *et al.*, 2015) comprising of PoTs can then be used to combine information from adverse effects to identify whether a substance is triggering a specific pathway and should thus be assessed for its potential toxicity. These pathways are different in nature from cellular pathways that have been described by the use of omics technologies such as KEGG (Kanehisa and Goto, 2000), Reactome (Fabregat *et al.*, 2018), Wiki Pathways (Slenter *et al.*, 2018), Gene Ontology (Ashburner *et al.*, 2000). PoTs are more specific to the concept of toxicity mechanism, which involves information about exposure and adverse outcome from resources, such as ToxCast (Richard *et al.*, 2016).

The adverse outcome pathway concept was introduced in 2010 (Ankley et al., 2010) as a structured format for the purpose of connecting toxicity pathways with adverse outcomes. The AOP consist of two components, the key events as nodes in the pathway, and key event relationships as edges. The key events represent a change in biological state, and two specialized key events are identified: a molecular initiating event is the initial interaction of a chemical with the biological system which is the first step in the pathway, followed by the mediating events of molecular or cellular response and ends with an adverse outcome at an organ, organism, or population level, if a

human health or an ecological risk assessment is being considered. The final adverse outcomes of this pathway are of regulatory interest and have been measured in traditional studies. Unlike the mode of action framework, AOPs are not specific to chemicals, so a particular AOP is relevant for multiple chemicals and is not specifically designed.

The organization for economic cooperation and development (OECD) started an AOP development program for harmonization of AOP definition and to support the development and use AOP framework. The AOPs intended to be part of the OECD development program are incorporated into a knowledge base AOP-KB central repository. The information is collected in a central repository and made available (Villeneuve et al., 2014).

The high density of data generated from toxicogenomic experimentation allows for a comprehensive description of the toxicity response by virtue of the mechanistic information that it provides, which informs the PoTs with cellular and molecular events that further bring this data closer to the regulatory assessment.

5 NANOTOXICOLOGY

Nanotoxicology is the regulatory process of assessing nanoparticle toxicity (Marquis *et al.*, 2009). These nanoparticles are of great interest to various industries because of their unique properties. Nanoparticles are defined generally as particles that have at least one dimension lesser than 100 nanometers. Substances that do not have any significant physical and chemical properties in bulk are capable of some intriguing properties at the nanoscale (Murty *et al.*, 2013). Nanoparticles have existed in nature and have been present in the environment. Humans have been interacting and employing nanoparticles since ancient times without a deeper understanding of their functions. A good example is colloidal gold, which is a colloidal solution of gold nanoparticles, used throughout human history to color objects. Nanoparticle nature of the colloidal gold and scientific reasoning of the red color as the optical property was described in the 1850s by Michael Faraday. Thus, establishing the phenomenon of color production due to the scattering of light by nanoparticles and the effect of nanoparticle size on color hue. Advancements in nanotechnology have not only resulted in broadening the understanding of nanoparticles already present in nature but have also led to the possibility to design and create nanomaterials. These engineered nanomaterials such as carbon nanotubes, which are hexagonal planar sheet circularized into a hollow cylindrical structure, are known to have mechanical, electrical, thermal, and optical properties of interest. They have a wide range of applications such as composite polymers, transistors, and biomedicine (Meredith *et al.*, 2013) Human interaction with these nanoparticles can be accidental or occupational, for example, exposure during the manufacturing process. The nanoparticle-human interaction can be through dermal routes, ingestion, or more likely by inhalation of particles suspended in the air leading to trachea and lung exposure (Morimoto *et al.*, 2013).

5.1 NANOPARTICLE INTERACTION WITH THE BIOLOGICAL SYSTEM

Nanoparticle exposure leading to toxic response is manifested due to its physical properties such as size, shape, surface area, as well as its chemical properties such as hydrophobicity and surface charge. (Podila and Brown, 2013). Size of nanoparticles plays a major role in the nano-bio interactions as it determines invasiveness of the particles and their final resting place (Jiang *et al.*, 2008) (Chen *et al.*, 2015). The nanoparticles can be in the size range of viruses and can interact with the same host response machinery. The shape

of the nanoparticle can also determine its invasiveness and site of deposition (Truong *et al.*, 2015). Nanoparticles are taken up by cells through the process of phagocytosis or pinocytosis (Geiser, 2010), leading to acute responses such as inflammation or chronic responses by triggering early onset of complex diseases such as asthma (Meldrum *et al.*, 2017), cancer or translocation to other organs leading to neurological and cardiovascular diseases (Simeonova and Erdely, 2009). Thus, nanotoxicology is vital for maintaining public health and safety.

5.2 NANOTOXICOLOGICAL IDIOSYNCRASIES

Regulation of nanotechnology is an area of active pursuit with increasing concern for public health and environmental effects due to the rapid influx of nanomaterial-based products. The Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH) regulation in Europe are concerned with the regulation of nanomaterial safety as it is covered under the definition of 'substance', but this is still early stages and there remains the need for refinement nanotoxicology regulation. Toxicity assessments measures from the bulk materials cannot be directly applied to the nanomaterials as they have different sets of characteristics and physicochemical properties that lead to the manifestation of nano-bio interactions. Determination of nanoparticle toxicity requires a very cautious systems toxicology approach to understand the perturbed molecular mechanisms in correlation with nanoparticle properties such as size, shape, and surface area. Omics technologies facilitate the systems toxicology approach of understanding the perturbed mechanisms by experimental measurement of different molecular species (Fröhlich, 2017). In contrast to traditional assessment methods that employ high dose exposure to elicit an observable phenotypic change, omics experiments can be performed at low doses to identify biomarkers of nanoparticle effect in the absence of the phenotypic change. The low dose assessment is more appropriate for nanoparticles because it helps to avoid agglomeration. The high density is likely to affect the dispersion of nanoparticles leading to the aggregate formation, thus affecting the cellular response. It is possible to identify the adaptive response of a system exposed to nanoparticles in low doses to replicate the precursor state of toxicity, which is a more realistic exposure scenario. The nanoparticle activity in the biological system is also defined by the formation of the corona around the nanoparticle due to absorption of proteins on the surface (Lundqvist *et al.*, 2008). The functional and interactive properties attributed to the nanoparticle due to this corona formation changes in different environments and conditions (Lundqvist *et al.*, 2011). Thus, it is imperative that true nano-bio interaction must be understood by measuring the molecular activity within the biological system.

Mechanistic information from the toxicogenomics assessment of nanoparticle can be used to characterize the nanoparticle, and this information can be used to predict the possible toxicity of new, untested materials in existence or development. It enables better designing paradigm of engineered nanomaterials, which will be safe by design. Advancements in omics techniques for gene expression can enable low cost and rapid assessment of toxic materials. Transcriptomics experiments have been performed for in vivo and in vitro studies to identify the dose-dependent response as perturbed molecular mechanisms and biomarkers that might not have been identified with traditional methods (Costa *et al.*, 2018).

6 CO-EXPRESSION PATTERN OF THE MOLECULAR MECHANISM

Research in genomics has transformed with the advent of high throughput and high content technologies such as DNA microarray and next-generation sequencing. These technologies are capable of generating high-density data, opening new possibilities of modeling complex biological events. In this frame, biological systems are explained as complex patterns of relationships between different parts of the molecular machinery, which produce a systemic effect during normal biological functions and adverse conditions, for instance, due to an external stimulus (Currie *et al.*, 2014).

Systems can be modeled as networks that can be explored with graph theory (Barabási *et al.*, 2011). In this setup, the set of vertices/nodes represent the biological entities, and the edges between the vertices represent their interactions. This information can be modeled as a simple undirected or directed network, where the network connections represent information flow channeling from one node to another, forming a signaling cascade. The connections can be given weight to represent the significance of the relationship between two specific entities and develop system-wide dynamics of connectivity and interaction.

Biological networks generally follow the power law distribution of connectivity and contain hubs with a high degree while most vertices have a low degree, making them robust to random failures but susceptible to targeted attacks (Cooper *et al.*, 2006). Random failures are not likely to produce a significant effect, but a major hub node disruption is of concern (Jeong *et al.*, 2001).

Gene expression profiles derived from transcriptomics data can be used to understand the interaction between the genes by the measure of coherence in their expression (Huang *et al.*, 2010). This modeling of expression profile as a network represents the biological pathways and processes active in the biological system in specific conditions (van Noort *et al.*, 2003). The gene co-expression networks can be inferred by partial correlation or mutual information under the assumption that co-expressed genes are most likely also co-regulated and participate in the same biological functions (Michalak, 2008). Researchers have been actively pursuing the space of gene co-expression network inference, which has led to the development of numerous inference methods over the years, and it continues to be a field of interest. Various algorithms for inference of co-expression networks have shown good performance, but, the choice of algorithm is usually determined empirically. Multiple recent studies have shown that in fact, the network inference

process benefits from combining multiple algorithms able to highlight specific types of relationships within the network (Marbach *et al.*, 2012). The "wisdom of crowds" approach to integrating the prediction from algorithms that infer gene relationship by different methods Regression, Mutual information, Correlation, Bayesian networks, and other mixed approaches resulted in robust and high-confidence networks. Mutual information and Correlation based methods are more conducive to inference of feed-forward loop defining the relationship of a gene (G) and two transcription factors (T1, and T2) where one transcription factor T1 regulates the gene G and the other transcription factor T2 which in turn regulates the gene G. The Regression and Bayesian network based methods are more accurate in the prediction of linear cascade relationships (Marbach *et al.*, 2012) where transcription factor T1 regulates transcription factor T2 which in turn regulates gene G. The inherent biases from each type of approach is complemented with the other approaches to overcome the limitations of individual predictions.

Once inferred, biological networks can be studied by looking at their intrinsic properties derived from the network topology. These properties can be used to identify the most important genes in a gene co-expression network. As stated above, few genes have high connectivity while most of the genes have low connectivity (Albert *et al.*, 2000). This means that all genes are not of equal importance in the network, there are hub genes that when disrupted will have a significant impact in the connectivity of local community of genes and distant parts of the network. Disruption of these hub genes can disturb the network in the most drastic manner and might break the network dynamics, so they are extremely important in the network structure while other genes with low connectivity can be disrupted with minimal to no change in the network behavior. This may be used to assume, for instance, that more toxic chemical insults tend to affect more central genes in the network (Chen *et al.*, 2017; Gopalacharyulu *et al.*, 2009). Centrality measures can be used to understand which nodes are important and essential in a network, on the basis of different network properties, for instance, on the number of connections that every gene has in the network. In addition to the number of connections, connecting pathways can also be used to identify important vertices by looking at their betweenness centrality. If many direct connectivity paths are flowing through a particular vertex, then it is essential for the communication channels to persist. These essential genes by their nature are likely to be more central in the network then to the periphery. This can be evaluated by measuring closeness. Importance can also be assigned not only by high connectivity but also by the quality of the connected neighbors. A vertex connected to a neighbor with a high degree itself is important because it is associated with high degree vertices. Multiple high centrality neighbors further increase the importance of the vertex, which can be measured with eigenvector centrality (Griffiths *et al.*, 2007). Centrality measures are important in identifying network bottlenecks vertices with high centrality (such as high degree or betweenness), which are

more likely to be essential molecules. This is also stated by the centrality-lethality hypothesis by which essential vertices are central and are more likely to produce lethal phenotypes upon their disruption (Yu *et al.*, 2007). A challenge in the application of network models to interpret complex biological events is translating the intrinsic network properties into biological importance. This is evidently dependent on the intrinsic network properties used to identify hub genes. Focusing on any specific centrality property is likely to lead to a biased assessment. A solution could be to integrate multiple centrality properties to increase the robustness of the assessment (Fei *et al.*, 2017). Other biological measures of gene importance, such as the differential expression in specific experimental pairwise comparisons, can be added to temper the hub nature with biological information.

Networks are gaining popularity in systems biology research, but there are different avenues of interpreting the biological significance of the network models. Gene co-expression network from biological perturbations, such as toxic exposure, has been used to suggest hub genes as possible biomarkers (Zhang *et al.*, 2013). These hub genes represent better dynamics than genes identified solely by the differential expression studies. In association, standard enrichment analysis is performed on the differentially expressed genes to find significantly enriched known manually curated pathways (Zhang *et al.*, 2017). However, a better approach could be to identify structural units that represent a closely connected set of genes co-expressing in this perturbed state. This can be accomplished by searching for subnetworks, defined as sub-communities of closely related nodes with unique signature and importance (Ideker *et al.*, 2002). These subnetworks are more likely to represent homogenous biological mechanisms and molecular events (Chen and Yuan, 2006). Gene co-expression networks are extremely dense and represent broad systems information from multiple molecular events, and thus, we cannot understand the biological significance of the whole network. Subnetworks drill down and focus on natural subunits and bring forth underlying biological significance. Categorization of subnetworks by annotations such as gene ontology can highlight significantly enriched known biological mechanisms.

Subnetworks/modules are in concept better solutions for defining molecular events occurring in a perturbed biological state of toxic exposure. They are free from any bias of known manually created pathway and thus are better representatives of the biological system inferred/reconstructed from the gene expression profiles. This present a novels opportunity to characterize subnetworks with molecular events for designing AOPs that are better representations of the molecular machinery of the perturbed system and can have an impact on toxicity assessment.

7 AIMS OF THE THESIS

Toxicogenomics studies have evolved with the implementation of experimental methods with a large amount of data. Which makes it possible to design more complex studies and use the computation methods to model the data. This requires tools and methods for data processing and analysis to obtain predictive toxicological results. There are two areas of much-required attention in current toxicogenomics approach, first is the study design phase which has various diverse approaches that incorporate different technologies and have different set of questions and expected results, these designs have no defined standards but mostly follow a general set of guidelines that have been accommodated from successfully performed and published studies. Another area is the processing, modeling, and analysis of large data from omics experiments, there are different computation methods and data processing techniques that are interchangeably used sometimes to similar effects while sometimes a different set of analytical steps produces different results, some methods turn out to be more sensitive than others while some methods are more prone to type I and type II errors than others and require more careful quality control of processed data. Different desired results raise the complexity of this process. It is essential to attain a proper standardization of toxicogenomics data processing and analysis to ensure reproducibility of quality of the results for predictive and deterministic toxicology results. Our focus is towards standardization of toxicogenomics data analysis with a set of easy to use graphically interactive tools which ensure reproducibility and quality of results. These tools are designed for all types of researchers with varying levels of expertise in toxicogenomics experimentation and data analysis.

1. Effective exchange of toxicogenomics data and reporting of analysis results from the study of nanomaterial toxicity.
2. Strive to standardize omics data processing with required quality checks and diagnosis of data to avoid type I and type II errors.
3. Research and development of a standardized systems biology approach to obtain robust gene expression networks.
4. Develop a methodology to identify important genes in gene networks by using features of graph theory. Identification of responsive subnetworks and their characterization by functional annotation.
5. Promote the usability of omics data by the integration of analysis methods into an analysis pipeline as easy to use tools.
6. Analysis and biological interpretation of toxicogenomics data from nanomaterial study by using systems biology and multi-omics approach to understand complex biological mechanisms, formulate MOA (Mechanism of Action), and AOPs (Adverse Outcome Pathways).

8 MATERIALS AND METHODS

Materials and Methods	Publication(s)
Omics Data Preprocessing	I, II, IV
Batch Effect Mitigation	I, II, IV
Network Inference	III, IV
Centrality based Gene Ranking	III
Responsive Subnetwork	III, IV
Functional Characterization	I, II, IV
Integrative Analysis / Multi-omics	I, IV

8.1 OMICS DATA PREPROCESSING

DNA microarray raw data files were transformed and imported into the R environment by using *read.maimages()*, *justRMA()*, and *read.metharray.exp()* functions from limma (Ritchie *et al.*, 2015), affy (Gautier *et al.*, 2004), and minfi (Aryee *et al.*, 2014) R packages respectively.

Preprocessing of the microarray raw data was performed by using a carefully designed analysis workflow with well-defined steps 1) Quality Control, 2) Filtering, 3) Normalization, 4) Batch Correction, and 5) Differential Analysis.

8.1.1 QUALITY CONTROL

Quality control reporting of imported raw data is performed by using yaqcaffy (Gatto, 2017), arrayQualityMetrics (Kauffmann *et al.*, 2009), and shinyMethyl (Fortin *et al.*, 2014) R packages. Quality of the samples was estimated by inspecting the QC report, and poor quality samples were marked for removal.

8.1.2 FILTERING

Filtering of imported raw data is performed by estimating the expression distribution of negative control probes and thus identifying a significant

expression score cutoff as base background noise. This cutoff score based evaluation was performed for each feature across all the samples, and the feature is expected to be consistently expressed above the background noise. However, a cushion for outliers was provided by specifying a minimum percentage of samples needed for the qualification of each microarray feature. A similar approach is used for Illumina methylation arrays where a p-value is obtained by using *detectionP()* function from minfi (Aryee *et al.*, 2014) R package, which compares the methylated + unmethylated signal versus the background signal (negative control positions). A significant p-value threshold was used to qualify positions across the majority of samples.

8.1.3 NORMALIZATION

Normalization of Agilent array data was performed by using functions from limma (Ritchie *et al.*, 2015) R package. Affymetrix array data normalization and annotation was done in step with the import of raw data to the R environment by using *justRMA()* function from affy (Gautier *et al.*, 2004) R package. Illumina methylation array data was normalized by using the functions from minfi (Aryee *et al.*, 2014) R package.

Distribution of signal in samples from different microarrays was observed by means of following diagnostic plots 1) Box plot displays signal distribution in each sample. 2) Density plot displays the density of the signal in each channel or beta values (M/M+U) for methylation arrays. 3) Mean difference plot displays the log intensity ratios (difference) versus log intensity average (mean) for two-color channel data. Distribution of signal should be consistent in different arrays and channels after normalization.

8.1.4 BATCH CORRECTION

Batch effects are technical sources of variation associated with the samples processed in batches during various preparation techniques and is also a result of pooling data from multiple experiments for meta-analysis. Technical variation in the normalized data is represented by Multidimensional Scaling (MDS) plot, which is a 2-dimensional scatterplot that displays the distance between each pair of samples computed as Euclidean distance of the top features. Sample annotation provided as the phenotype data is represented as heatmap generated by using *confounding()* function from the swamp R package, it displays the linear correlation between the annotations. Correlation between the technical variation in the data and sample annotation is represented as a heatmap generated by using *prince()* and *prince.plot()* functions from the swamp R package, it displays the principal components of variation in the data and their correlation the sample annotations. The technical variation information from prince plot was used to identify possible batches as non-biological annotations with high influence

on variation. Correlation between the possible known batches and the biological variables of interest was inferred from the confounding plot. Independent batch variables were selected for correction of noise.

Hidden sources of variation were identified as surrogate variables by using *sva()* function from sva (Leek *et al.*, 2012) R package. These surrogate variables might represent the technical batches or biological annotation (e.g., subtypes). The batch surrogate variables were identified by observing their correlation with principal components of variation by means of an updated prince plot and the correlation with the known variables was observed from the updated confounding plot with the added surrogate variables.

Batch correction was performed by using the *ComBat()* function from sva R package for correction known/unknown batches. The cross-batch correction effects were avoided by iteratively applying *ComBat()* function for one batch variable at a time, while the rest of batch variables were added to covariates of interest, a constant biological variable of interest is maintained in all iterations. Diagnosis of corrected data was performed by using the post-correction prince plot and MDS plot.

8.1.5 DIFFERENTIAL ANALYSIS

Differential analysis was performed by using functions from the limma R package. The linear model for limma was specified with the biological variable of interest and covariates. The batch variables identified from the previous steps were specified as covariates along with the biological covariates. Some dataset did not require explicit correction batch variables since the grouping of samples in the MDS plot aligned very well with the biological variables of interest. However, the corrected or uncorrected batch variables were specified in the limma linear model as covariates. The annotation from the biological variable of interest user was used to define the contrasts for differential analysis.

8.2 NETWORK INFERENCE

A robust gene co-expression network was reconstructed from the gene expression profile data as a consensus of the networks inferred by combining different inference algorithms from minet R package (Meyer *et al.*, 2008). An ensemble of network inferences was generated by using the combination of method, estimator, and disc options from *minet()* function, which resulted in 96 valid possible combinations. Different scales of statistical association measures from the various inference algorithms were accounted for by using a two-level ensemble process to combine the inferred networks. At the first level inferences from the same algorithm were merged by summarizing the co-expression measure between all pair of genes across the networks with

median, mean or max. The summarized gene-gene co-expression measures from the level one ensembles were used to obtain ranks for gene pairs or edges, the ranks from the various level one ensembles were summarized to a single rank list by using *Borda()* function from TopKLists (Schimek *et al.*, 2015) R package. The summarized rank list was used to form the final consensus network by using one of the following described approaches. The first approach iteratively adds ranked edges starting with the top-ranked edge from the ranked list until all the genes in the network have at least one edge. An alternate approach defines a cutoff as an n% of top-ranked edges from the ranked list. Thus, the final consensus is composed of most confidently inferred ranked gene-gene pair co-expression.

8.3 CENTRALITY BASED GENE RANKING

Ranking of genes in the network inferred from expression profile was performed by using the centrality property scores to define the importance of a node in the network. Node centrality scores were computed by using the *degree()*, *betweenness()*, *eigen_centrality()*, and *closeness()* functions from igraph (Csardi and Nepusz, 2006) R package. These different centrality properties were combined by using a rank based approach. A rank list was generated for each centrality measure, and these different ranks were then reduced to single unified rank by using *Borda()* function from TopKLists (Schimek *et al.*, 2015) R package. Thus, combining the influence of all the properties in the final ranked gene list. A differential expression score based rank list was also be added for condition-specific information.

8.4 RESPONSIVE SUBNETWORK

Responsive subnetwork was obtained by a three step process of 1) Module detection, 2) Characterization & Functional Assessment, and 3) Reconstruction/Merging.

8.4.1 MODULE DETECTION

Module detection was performed by using *cluster_walktrap()*, *cluster_spinglass()*, *cluster_louvain()*, and *cluster_fast_greedy()* functions from igraph (Csardi and Nepusz, 2006) R package.

8.4.2 CHARACTERIZATION & FUNCTIONAL ASSESSMENT

Characterization of detected modules was performed by inspecting the median rank of the genes in the module by centrality, differential p-value,

and differential logFC; median rank of edges in the module; and module size. The comparison of module statistics within and across different modules was performed by reporting it as a radar chart by using *chartJSRadar()* function from radarchart (Ashton and Porter, 2016) R package. Functional assessment of the detected modules was performed by enrichment of Gene Ontology (GO) annotations represented by the module genes. Semantic similarity between sets of enriched GO terms from two different modules was computed by using *mgoSim()* from GOSemSim (Yu *et al.*, 2010) R package. This semantic similarity was used to compute a GO specific Jaccard index as a measure of similarity between the modules based on their enriched functional annotations. The computed similarity matrix was plotted as a heatmap by using *heatmap.2()* from gplots (Warnes *et al.*, 2016) R package to identify functionally similar modules.

8.4.3 RECONSTRUCTION/MERGING

Reconstruction/Merging of the modules was performed to define a singular responsive subnetwork by combining two or more modules based on the module characteristics and functional similarity. The final responsive subnetwork itself was characterized by observing its module characteristic in the previously described manner and by inference of the summarized functional annotations.

8.5 FUNCTIONAL CHARACTERIZATION

Functional characterization of network modules was performed by using gene ontology (GO) annotations from organism-specific R Bioconductor annotation libraries. The participant genes from the modules were used to perform enrichment analysis using Fisher's exact test. The hierarchical graph based structure of gene ontology was used to cluster and summarize the annotation. The clustering of functional annotations from the summarized annotation was represented as a tileplot (Supek *et al.*, 2011) by using the *treemap()* function from treemap (Tennekes, 2017) R package. The functionally related annotations were summarized to a singular visual tile, and these major tiles were labeled with the most significant gene ontology term in each cluster.

The functional similarity between the GO terms was computed by using *mgoSim()* function from GOSemSim (Yu *et al.*, 2010) R package. The GO term similarity was converted to a reciprocal distance matrix, and *hclust()* function was used to perform clustering on this distance matrix. A defined set of clusters was obtained by cutting the cluster tree with function *cutree()*. This clustered annotation was further summarized by selecting the most significant GO term from each cluster as the cluster representative.

8.6 INTEGRATIVE ANALYSIS / MULTI-OMICS

SMITE (Wijetunga *et al.*, 2017) toolkit was used to perform integrated omics analysis, where the combined differential influence from all the omics layers (mRNA, miRNA, and DNA methylation) was used to rank the genes. The adjusted p-value and fold change from mRNA differential analysis were directly assigned to each gene. Epigenetic effects were represented by mapping the miRNA and DNA methylation information to the regulatory regions in the gene structure. Gene promoter was defined as a region of -1 kb and +1 kb flanks from the transcription start site. Gene body was defined as a region from TSS +1 kb to the transcript termination site. Differentially methylated CpGs were associated with gene promoter and body while the miRNA was associated with the gene body of its known target genes (top 10% scoring targets from targetScan (Agarwal *et al.*, 2015)). A summarization of miRNA and CpG values was performed to integrate multiple values from each region to obtain region-specific values. The summarized methylation differential values for each gene region were obtained by using Stouffer's method (weighted by the distance of the CpG from TSS) to integrate values from all CpGs associated with that region. The miRNA differential values were similarly summarized by using the Sidak method. These new integrated p-values from miRNA and methylation were logit transformed and rescaled to the range of minimum and maximum logit transformed mRNA expression p-values, thus, avoiding any biased effects in the downstream analysis. Gene scoring was performed by using the differential p-values from all layers. This gene score was then incorporated into the known Reactome52 protein interaction network (Croft *et al.*, 2011) where the nodes were assigned gene scores, and the edges were assigned an average of the scores from the connected nodes. This annotated network was used to identify modules by using Spinglass algorithm followed by annotation of these enriched modules with KEGG pathways.

8.7 DEFINING MECHANISM OF ACTION

Mechanism of action was defined on the basis of the biological pathways associated to functional genes obtained by combining genes from all identified functional modules. These functional genes were segregated by the direction of regulation information from multiple omics layers. A concordant set of functional genes was defined as 1) Concordantly upregulated with the upregulated gene, hypomethylation in promoter or downregulation of targeting miRNAs. 2) Concordantly downregulated with the downregulated gene, hypermethylation in promoter or upregulation of targeting miRNAs. While a discordant set of functional genes was defined as complementary to the concordant set defined before. These three set of functional genes 1) Whole set of functional genes, 2) Concordant subset of functional genes, and 3) Discordant subset of functional genes were used to perform enrichment of

KEGG pathways by using *enrichKEGG()* function from clusterProfiler (Yu *et al.*, 2012) R package with an adjusted p-value threshold of 0.05. Furthermore, the direction of regulation was assigned to each enriched KEGG pathway as the median expression fold change of the genes in the pathway. These set of upregulated and downregulated KEGG pathways was used to compute the distance between different exposures as a measure of functional congruence. The following function was used to compute the distance by separately computing Jaccard index on common upregulated jacc_up(A,B) and common downregulated jacc_down(A,B) pathways between A and B exposures.

$$dist(A,B) = 1 - [jacc\_up(A,B) + jacc\_down(A,B)]/2$$

8.8 BIOINFORMATICS TOOL IMPLEMENTATION WITH GRAPHICAL USER INTERFACE

The R programming language/infrastructure, R libraries (CRAN, Bioconductor, and GitHub), and R shiny (Chang *et al.*, 2017) web framework was used to create bioinformatics tools with graphical user interface (GUI). The steps for preprocessing and analysis of microarray data were incorporated into a guided workflow which can be executed from the bioinformatics tool eUTOPIA. The sets for network based exploration of gene expression profile and identification of the functional response module to define the biological response were defined as an analysis pipeline which can be executed form the bioinformatics tool INfORM. The graphical interface implementation has a standard R shiny architecture which is separated into a user side code UI and server-side code Server. Both UI and Server code is implemented in the R programming language and uses R shiny library along with other R libraries that extend the shiny web framework with customized UI widgets, javascript functionalities. These tools were developed in the scope of this doctoral study; they are actively maintained with continued support to future usability.

8.8.1 EUTOPIA

eUTOPIA incorporates the step for preprocessing of the microarray data from different platforms. Some components from the UI and Server are switched on/off depending on the microarray platform of choice to present an optimized and curated pipeline for each platform. The graphical interface layout is vertically split into two persistent panels. The smaller left panel is designed to execute steps in the pipeline while the larger right panel is used to display the data representations as tables and plots. The right panel contains a nested tabbed view where the tabs in the main panel correspond to the workflow steps in the left panel and displays the corresponding results

from the execution of each step as plots and tables. The processing of user input is handled by the R function in Server while functions for more complicated processes with custom implementation such as filtering, batch correction, and differential analysis are defined in a separate R source file. Some additional files are the GIF file displayed during the execution of steps in the GUI, R markup file used to compile and export the PDF analysis report with the plots from the GUI, and the list of unreliable cross-hybridizing methylation (Chen *et al.*, 2013).

8.8.2 INFORM

The analysis pipeline for the network based exploration of gene expression profile and identification of the functional response module is implemented as a set of functions that can be executed individually with required parameters to perform computations or generate plots for the data. These set of functions are defined in an R source file and can be used to perform the complete analysis from the command line. However, the UI and Server implementation of INfORM in R shiny web framework streamlines this process, and the user can perform the complete analysis by a single click of the button after uploading the necessary input files. The simple visual layer is separated into two distinctly marked panels “Upload” to upload the data and setup the analysis and “Display Area” to display the data and results by means of the tables and plots. Further customization of the analysis setup can be performed by adjusting the advanced parameters from the hidden panel in “Upload”. The Server code effectively executes the different components of the pipeline sequentially from the processing of the uploaded data to the detection of modules and functional annotation by using the functions from the R source file. The user defines the final response module, and thus it is not an automated task executed by the pipeline implemented in the GUI. An additional R markup file is used as a template to generate the PDF for the tileplot representation of the summarized functional annotation.

9 RESULTS

9.1 STANDARDIZATION IN OMICS REPORTING

The MIAME compliant format is the prescribed standard for sharing omics data with the research community. It is an established prerequisite for any data to be published in peer-reviewed scientific journals. These reporting formats are supported by many public resources, which are great platforms for archiving the omics data files and preserving them for future use by other researchers. However, these reporting formats present very minimal information about the experiment, study design, data analysis methodology, tools/programming scripts employed in the analysis, and interpretation of the results. The research articles report the biological study with limited focus on data analysis and main emphasis in the interpretation of the final results, to describe the biological findings or discoveries of biological phenomena. The guidelines for research articles have been improved with more journals requesting analysis code/programming scripts for reproducibility of analysis performed. In our efforts to ensure clarity and completeness of reporting the omics data, we encountered the concept of data articles. We employed this new type of focused publication (article I) for reporting the study design, omics assay strategy, omics data, analysis methodology, and to a great extent the intermediate analysis result (Figure 2).

The data from the multi-omics study of nanomaterial exposure effect on human cell lines (article IV) was described along with the experimental procedures, analysis methodologies, and the results. This information is collated and provided in a comprehensible format of a scientific article which is expressed by the author towards independent researchers (article I). There are three sets of microarrays for 1) mRNA, 2) miRNA, and 3) DNA methylation each.

mRNA

mRNA expression assay was performed for 96 samples by using the Agilent SurePrint G3Human GE 8×60K array.

miRNA

miRNA expression assay was performed for 91 samples by using the Agilent SurePrint G3 Unrestricted Human miRNA_V21 8×60K array.

DNA methylation

DNA methylation assay was performed for 99 samples by using the Illumina HumanMethylation450 BeadChip array.

The analysis source code and the programming scripts implemented for data analysis were shared with great detail to ensure reproducibility of analysis results. The data article has a unique focus towards more comprehensive data reporting and thus provides an ideal platform for sharing the technical information that has been neglected in the research articles of the same datasets. The higher comprehension of this reported data is ensured by this reporting format of a scientific journal that is agnostic to any specific data format and can be understood by a generic independent researcher.

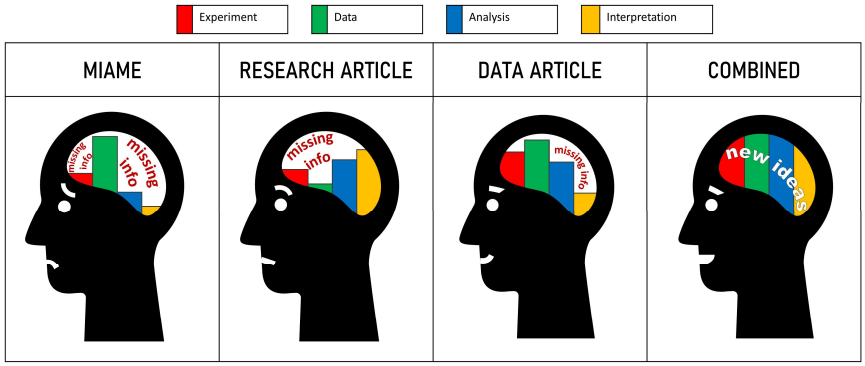

Fig. 2 - Omics data reporting strategies MIAME, research articles, and data articles have contrasting information content capabilities. Data articles are capable of providing holistic information.

9.2 REPRODUCIBILITY AND ROBUSTNESS OF TOXICOGENOMICS DATA ANALYSIS

9.2.1 DATA PREPROCESSING

Standardization of toxicogenomics data analysis is much needed to ensure the reproducibility and quality of result for implementation in toxicological evaluations.

In this thesis, I engaged in the effort of implementing the best practices for DNA microarray omics data analysis into a robust and user-friendly workflow to avoid serious pitfalls in the omics data processing. I was able to crystallize the following analysis steps as essential in the successful processing of DNA microarray omics data; **i)** I recognized the need for the quality assessment of samples from the raw assay data. I identified the tools for **Quality Control** of samples for identification and reporting of poor quality samples. **ii)** In step with quality control of samples, I also recognized the need for assessment and identification of poor quality probes in the raw data. False estimations in the analysis must be avoided by **Probe**

Prefiltering to drop poor quality probes. **iii)** I identified the **Normalization** of raw data as the minimum requirement before the application of statistical methods. The data must be normalized before performing quantitative comparisons to avoid false estimates of positive effects and negative effects. **iv)** I identified **Batch Effect Mitigation** as a critical step in microarray data processing during the phase of exploring and establishing the best practices required to avoid pitfalls. Batch-wise processing of samples during sample preparation and omics assay introduces non-biological noise that affects the quantitative comparison of sample groups due to batch effects. The latent noise can be estimated by identifying the sources of variation in the data. The variation in the data is determined as principal components obtained by orthogonal transformation of assay data. The batch effect can be estimated by observing the association of the technical variables with the principal components representing high variation. The identified batches must be adjusted to reduce their effect before any quantitative comparison is performed. The batch parameters are estimated by using the empirical Bayesian method of determining priors from the standardized data followed by adjustment of estimated batch effect. The adjustment of multiple batches must be performed iteratively to ensure that specific batch effects are estimated while the effect of other biological and batch variables is preserved. **v) Differential analysis** of the biological molecules is performed between groups of samples to observe the positive or negative effect in a perturbed biological state. This comparison is facilitated by advanced linear model methodology implemented in limma; it is a very robust method of modeling assay data of biomolecules locally and globally to account for variance. The set of biomolecules with significant positive or negative effect are qualified by specifying the scale of comparison as log transformed and significance as a probability value. However, it is advised to perform multiple testing correction and obtain adjusted probability values to account for false positive results due to the chance of testing a large number of biomolecules.

I ensured reproducibility of the results by implementing these best practices as pipeline/workflow (Figure 3) that requires minimal setup, presents the results of the data processing steps with meaningful graphical representation to understand and appreciate the processing effect, and is easy to execute. I created an effortlessly efficient solution for generic end users by implementing this robust workflow with an easy to use graphical interface developed in R shiny web development framework.

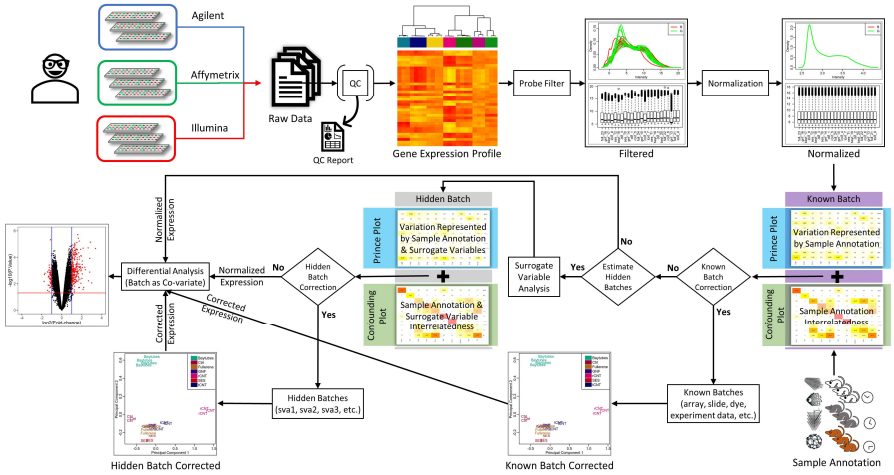

Fig. 3 - Microarray data preprocessing and batch effect mitigation decision-making approach in eUTOPIA.

9.2.2 MOLECULAR SYSTEMS ANALYSIS

The data from omics assays is usually employed to identify the significantly perturbed biomolecules. However, the perturbed biological state can rarely be defined by the action of the independent individual biomolecules. The complex biological system can be inferred by modeling the patterns of biomolecular activity in response to external stimuli. The biological system level activity can be used to identify the altered molecular mechanisms that define the effects on the biological system. Biological systems can be inferred as molecular networks that represent their coherent behavior in specific molecular mechanisms and possible molecular interactions.

Network Inference

Reconstruction of a molecular system model from gene expression profiles and exploration of system dynamics by graph theory is an area of great interest. There are many methods that have been proposed for inference of network from gene expression data. However, there is lack of consensus regarding the best choice of inference method since they have comparable performances and can be complementary. The power of these different methods can be used in combination to ensure the robustness of network inference (Marbach *et al.*, 2012).

Robustness by consensus

I employed the multiple mutual information based network inference algorithms from minet R package to infer a robust gene co-expression

network from perturbed gene expression profiles to ensure that the inferred network is not biased by choice of inference algorithm. Evidence from multiple inferences increases the confidence of inferred gene-gene connections. The robustness attributed to the consensus network is in terms of its accuracy to represent co-expression between molecules.

Different network inference scenarios were obtained by combining various options of network inference algorithms ARACNE (Margolin *et al.*, 2006), MRNET (Meyer *et al.*, 2007), MRNETb (Meyer *et al.*, 2010), CLR (Faith *et al.*, 2007); entropy estimators (pearson, spearman, kendall, mi.empirical, mi.mm, mi.shrink, mi.sg); and discretization methods (equalfreq, equalwidth, globalequalwidth).

Summarizing co-expression scores

I performed the algorithm specific consensus by combining the inferences from all the networks computed by a specific algorithm. I further evaluated the accuracy of the consensus networks by using the NetBenchmark (Bellot *et al.*, 2015) R package. The AUPR20 scores generated by NetBenchmark highlight that network accuracy was better for higher consensus networks, and the accuracy progressively improved with the assimilation of more network inferences. Variability in the performance by algorithms was observed as, CLR, MRNET, and MRNETb inference algorithms had better accuracy scores than ARACNE.

I tested different measures mean, median, and max for summarization of the co-expression scores. The most robust summarization across all inference algorithms was obtained by mean, while the median was quite similar to mean. Summarization by max performed best for ARACNE, but it performed poorly for the rest.

Consensus by the significance of evidence

Algorithm specific consensus network inferences contain consolidated evidence of robust molecular relationships inferred by each algorithm. However, the scale and distribution of score defining these interactions vary by the algorithms, which makes it difficult for further consolidation (Figure 4).

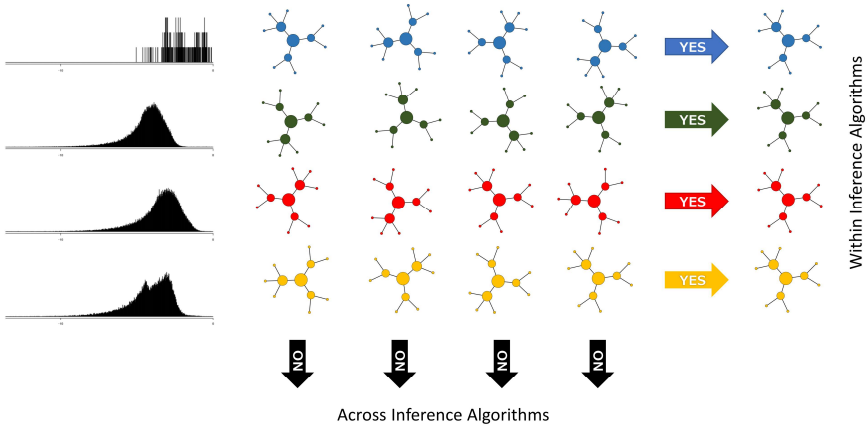

Fig. 4 - Different scale and distribution of co-expression score inferred by different inference algorithms hinder the summarization of co-expression scores across inference algorithms.

I defined an alternative approach to represent the significance of molecular relationships inferred by each algorithm consensus in isolation to the molecular co-expression scores by transforming the scores to ranks. Thus, we were able to isolate the significance of the connections from the raw measure of co-expression, allowing to avoid the imbalances of co-expression score inferred by different algorithms. The rank of the connections represents their prediction accuracy, which is used to incorporate the connections with accuracy. The rank wise significance of molecular connections from various algorithm consensus' is consolidated into a single consensus rank list by using the widely accepted Borda method.

Finally, the reconstruction of the perturbation modulated molecular system is accomplished by rank wise selection of the top molecular connections from the consolidated rank list until all the molecules are represented in the inferred network by at least one connection (Figure 5).

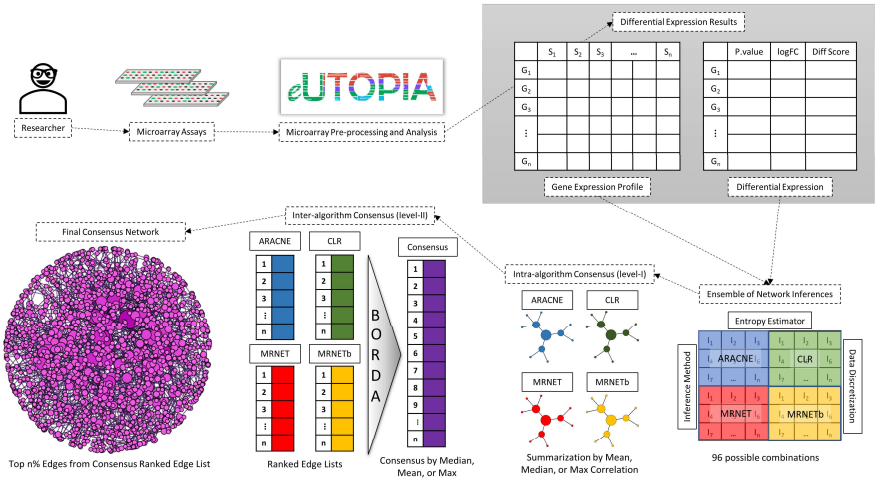

Fig. 5 - Robust network inference by consensus from different inference algorithms.

I have evaluated the efficiency of the consensus network creation process by the exploration of simulated expression data from a biological source network of E. *coli* (Ma *et al.*, 2004). A simulated network dataset Syntren300 was obtained from R package grndata (Bellot *et al.*, 2014), this dataset is created by using the SynTReN simulator (Van den Bulcke *et al.*, 2006). This dataset contains a binary true network of 300 genes (subgraph from the biological source network) and the corresponding simulated expression dataset. Evaluations of different levels (set size) of consensus networks can be used to observe the robustness of the consensus network creation process (Figure 6).

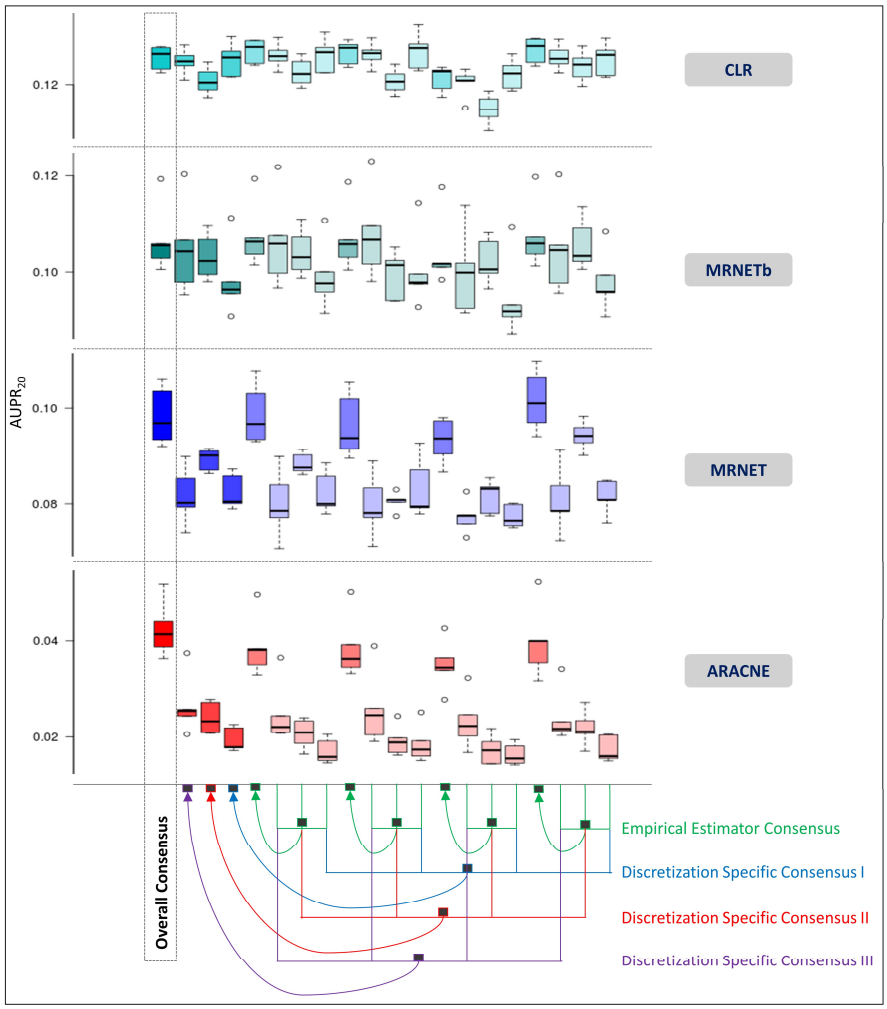

Fig. 6 - The consensus creation strategy of INfORM was evaluated on a simulated expression dataset syntren300. The accuracy measure AUPR20 is on the y-axis, and different levels of consensus are represented by the darker shade of the box. Different compositions of consensus are reported in the x-axis margin. The overall consensus highlighted by the dotted outline consolidates all individual inferences. The accuracy of consensus inferences is better than the individual inferences, and the overall consensus effectively maintains high accuracy.

I evaluated the performance of the rank based inter-network consensus by inspecting the number of true positive (TP) edges in final ensemble network (inter algorithm), level I ensemble networks (intra algorithm), and individual network inferences. The intersection of TP edges was inspected across the spectrum of networks inferred by each algorithm along with its level I ensemble and the final ensemble. In our benchmark experiments, the

intersection of TP was more consistent within MRNETb and CLR inferences. MRNET level I consensus consisted of more TP edges than the individual inferences, thus highlighting the consensus benefits. ARACNE level I ensemble compiled more TP edges than all the individual inferences and even the final ensemble. However, the f-measure for ARACNE level I ensemble is very low which means that it has an even higher number of false positive (FP) edges and thus by itself will not be the accurate reconstruction of the biological system. Overall comparison of TP edges in all level I ensembles and the final ensemble highlights the robust nature of the ensemble which is able to correct for low accuracy of some algorithms (ARACNE) (Figure 7).

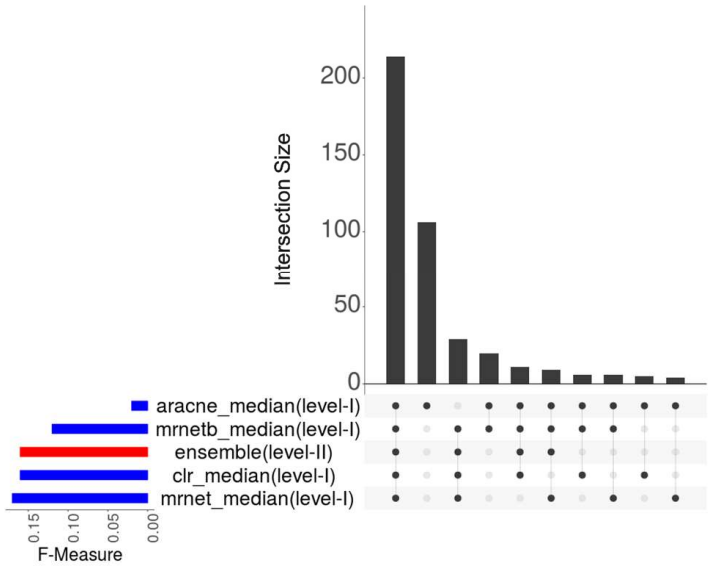

Fig. 7 - UpSet plot for the intersection of true positive edges inferred by different level-I ensembles and the final level-II ensemble

Modular component of the inferred network

Biological networks of interacting molecules contain hub nodes of high importance that influence the communication of many molecules with the smaller and larger chain of interactions. Moreover, modular subunits exist within the networks, which consist of closely connected molecules that are more interactive with their neighbor nodes compared to the rest of the network. These modules define a specific expression of the system in a given biological condition. The module detection can be performed by various algorithms, thus posing a question of the best choice of method to ensure more accurate results. I investigated the performance of module detection

algorithms by identifying modules from a test ensemble network (syntren300) by different methods walktrap, spinglass, greedy, and louvain (Csardi and Nepusz, 2006). The similarity is computed by means of the jaccard index between the set of genes in pairs of modules. Clustering was performed on the similarity matrix of modules detected from all methods. The heatmap representation highlights the similarity cluster formed by individual modules from different methods, thus suggesting a high level of similarity in the module detection methods (Figure 8).

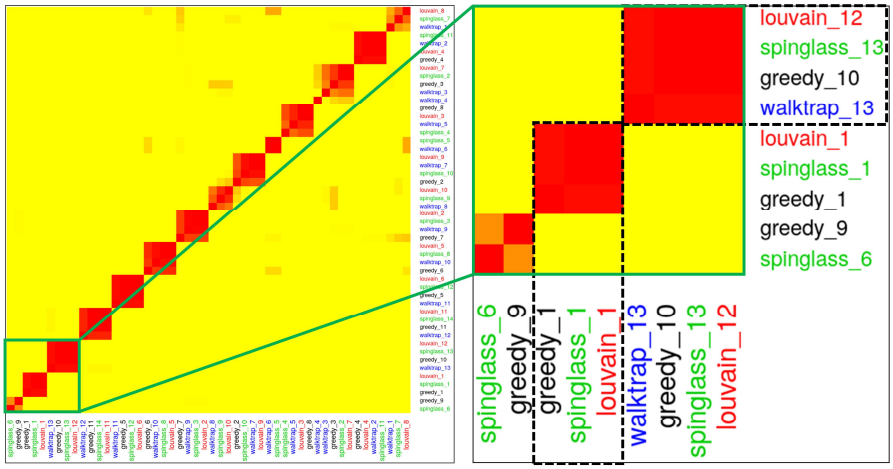

Fig. 8 - Benchmarking the possible divergence in the performance of Walktrap, Spinglass, Louvain, and Greedy methods. Their performance was found to be fairly complementary.

Response module

Gene networks are complex structures that are difficult to navigate through. Their biological significance is concealed behind the high density of nodes and edges. It is imperative to identify not only the genes but also the network dynamics that best describe/explain the biological query at hand. In the scope of this thesis work, I explored the idea of identifying the core responsive subnetwork that highlights the network dynamics of biological significance. The targeted assessments bring forth the mechanisms that are hidden in the overall network due to dilution by non-relevant genes and network dynamics. I defined an approach for response module detection by using the top ranked genes in the network. The genes in the overall network are ranked by different centrality measures (betweenness, degree, eigenvector, clustering coefficient) and differential expression score. These individual ranks of significance are aggregated by the Borda method, and the median aggregate rank list is used to identify the top ranked genes in the network (Figure 9).

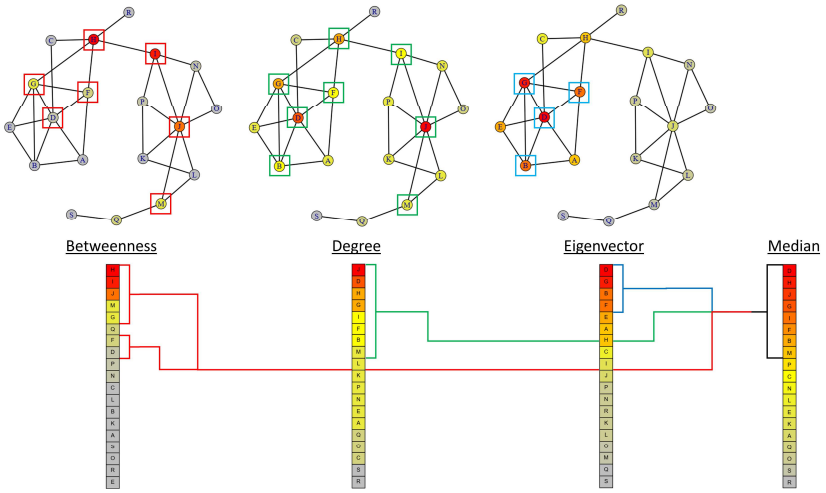

Fig. 9 - Different network centrality properties highlight different genes of importance depending on the network dynamics of interest.

I identified the most important genes by using a selection cutoff of top *n* ranked genes from the aggregated ranked list. The nodes representing the selected genes are used to draw a constellation by finding connections between all pairs of selected gene nodes. All the shortest between each pair of nodes are identified, and finally, the accumulated information of paths is used to infer the responsive subnetwork (Figure 10).

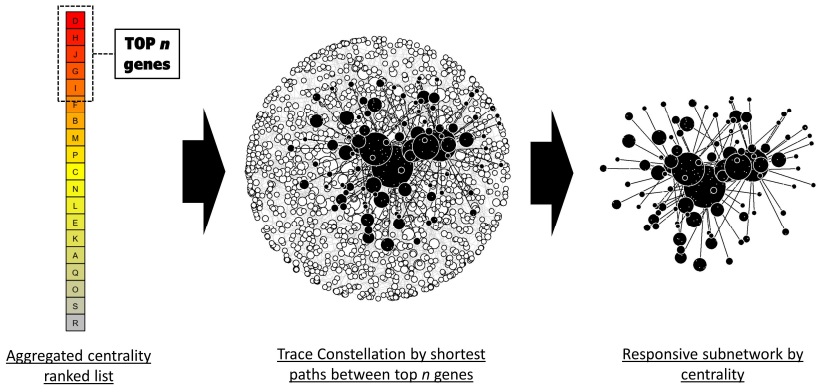

Fig. 10 - Responsive subnetwork definition by tracing constellations between most important genes by network centrality and biological scores.

This analysis approach proved to be beneficial in the study of Carbon Nanomaterial (CNM) toxicity response (Kinaret *et al.*, 2017a). In that context, we studied the effects of 6 different CNMs on THP-1 macrophage cell line as well as the lung tissue of mice (Kinaret *et al.*, 2017a). This system biology study observes the effect of nanomaterial exposure in relation to their intrinsic properties (length, diameter, surface area, and aspect ratio) and

compares the molecular response from in vivo experimental assay in mouse lung tissue versus the in vitro assay in cell lines derived from Human THP-1 macrophages.

Genes expression information from Mouse in vivo and Human in vitro assay was integrated for the selected set of 3868 orthologous genes (Ensemble database). The genes that exhibit a strong response to geometric properties of the CNMs are selected for further analysis. The gene co-expression networks were inferred for individual Human and Mouse datasets as well as the integrated orthologous datasets. The responsive subnetwork is identified by drawing a constellation of shortest paths between the top ranked genes in the network that form a subnetwork, representing the biological system that is composed of and is under the influence of the top ranked genes in the network. Functional characterization of this subnetwork explains the molecular mechanisms, events, components, and pathways with active involvement in producing the adverse outcome.

We observed that the initial set of differentially expressed gene sets were widely divergent and had minimal overlap while the congruence of functional annotation after network analysis was much higher. Our results suggest that, even though the significantly enriched genes from the in vivo and in vitro assays are different, they are still involved in the same biological functions. Thus, we hypothesize that the in vitro study at the functional level is able to reproduce the results from the in vivo study (Figure 11).

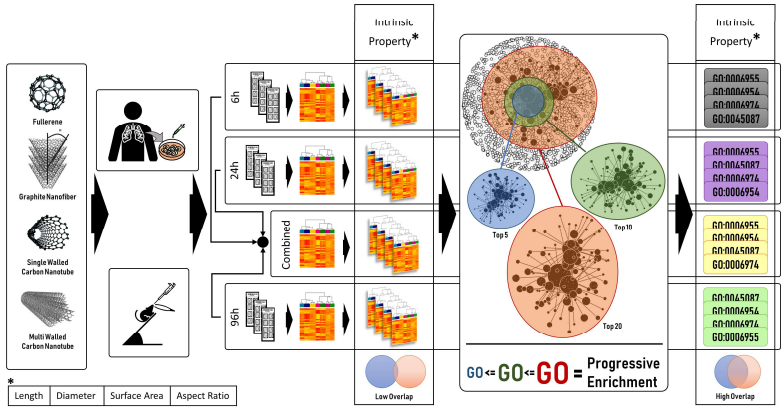

Fig. 11 - Congruency of carbon nanomaterial exposure response in Human cell line and Mouse lung tissue.

I further refined the responsive subnetwork detection methodology to incorporate the network topology based modular structure. Modules can be detected in the network as a segregated cluster of nodes that are more closely connected locally than the rest of the network. These modules derived from the gene co-expression relationships are highly likely to represent coherent

biological molecular mechanisms (Gene Ontology). However, the discovered modules represent the generally representative molecular functions and need some measure of distinction to identify the biologically significant modules. In this refined methodology, I characterized the modules by the median gene ranks in each module by different properties i) centrality, ii) differential p.value, iii) differential logFC, and also by the iv) median edge rank obtained during consensus network creation. Furthermore, these modules can be functionally congruent (GO semantic similarity) and can be represented as a singular functional response. Thus, the module characterization scores and the functional congruence is used to define the final response module that best represents the biological query of interest (Figure 12).

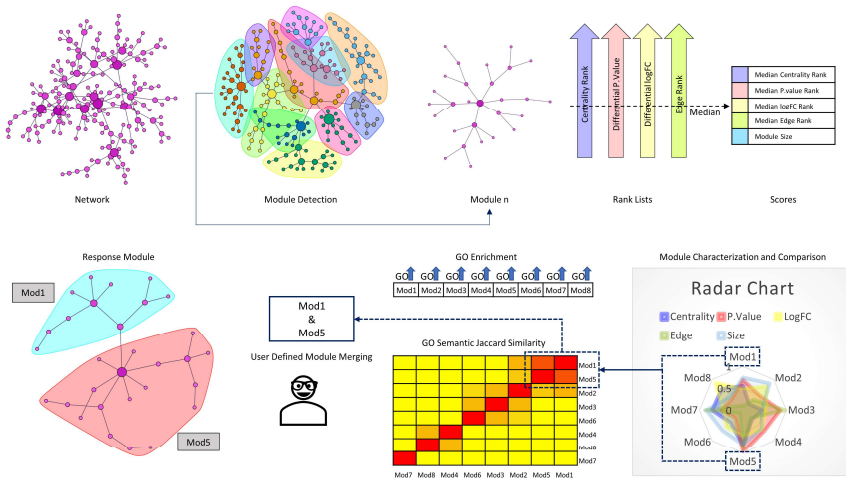

Fig. 12 - Responsive subnetwork definition by characterization of modules by node and edge importance, as well as the functional similarity between detected modules.

Molecular mechanisms

The modular nature of a network of molecular interaction can be associated with known biological mechanisms based on the functional annotation of molecules. Enrichment of gene ontology associated with the molecules can be used to ascertain the biological mechanisms. Modules from larger molecular networks inferred from expression assays represent the active biological pathways and functions in the perturbed biological condition. The enriched biological pathways characterize these natural modules present in the reconstructed biological system. Thus, I postulate that, by characterizing these modules, it is possible to interpret the mechanism of action of exposure (Figure 13). Response to toxic substances can be characterized by this approach, making it possible to define adverse outcome pathways.

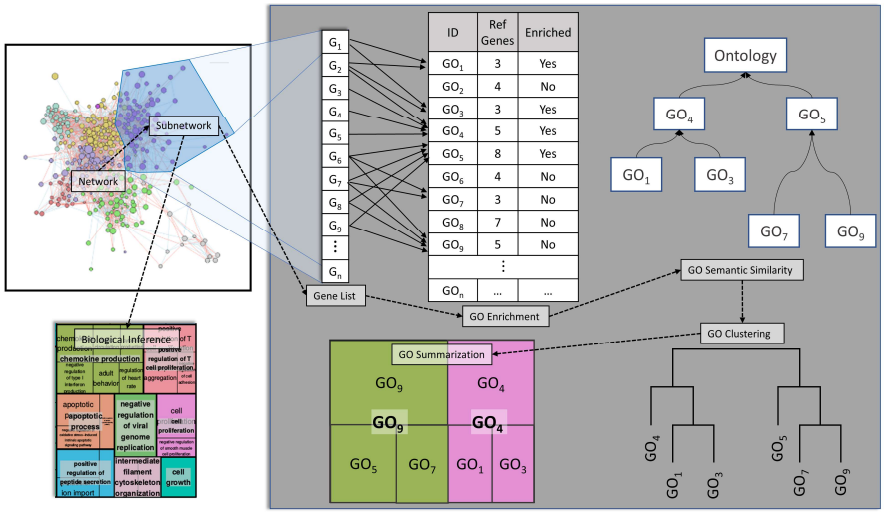

Fig. 13 - Functional characterization of subnetwork by GO annotation.

INfORM was used to analyze the data from a publicly available gene expression dataset obtained from NCBI Gene Expression Omnibus (GSE13355) (Nair *et al.*, 2009; Swindell *et al.*, 2011). It is a study on psoriasis patient with samples from lesional skin areas, normal skin areas, and healthy donors.

The summarized GO annotation tileplot (Figure 14) generated from the responsive subnetwork highlighted following biological mechanism 'immune response' (Liang *et al.*, 2017), 'keratinization' (Iizuka *et al.*, 2004), and 'proteolysis' (Dubertret *et al.*, 1984). These GO terms are of relevance to psoriasis pathogenesis.

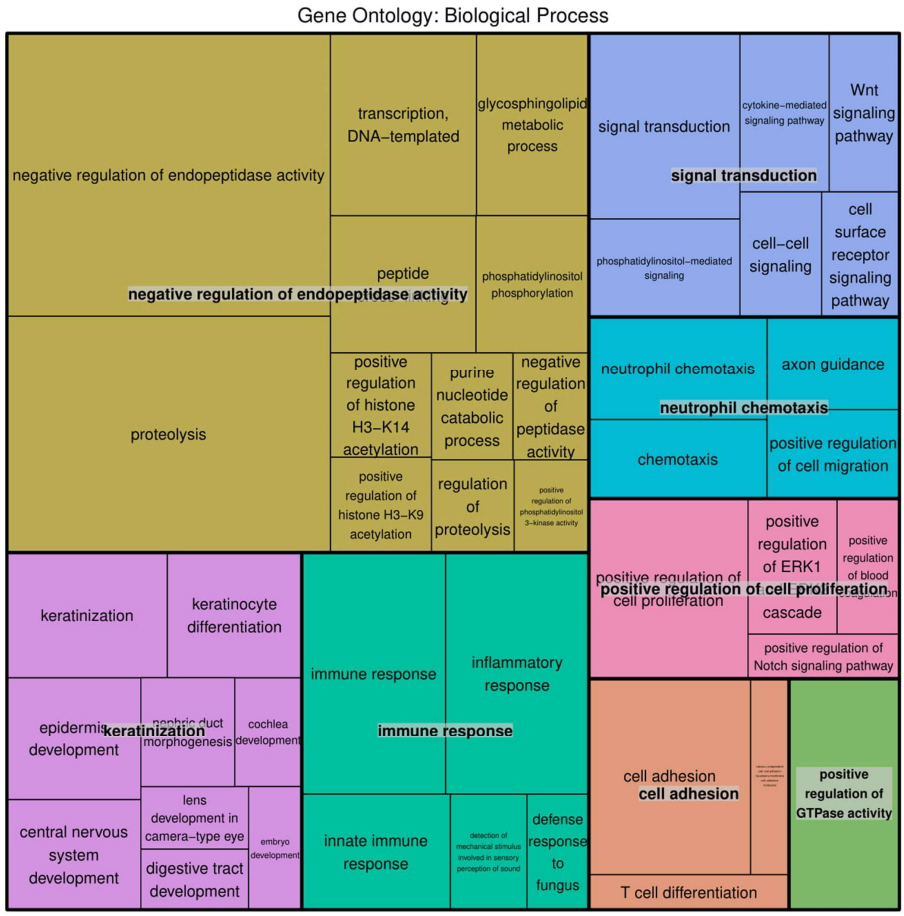

Fig. 14 - GO (Gene Ontology) terms summarized by clustering on the basis of their similarity to minimize redundant annotation information and highlight the important clusters of terms.

9.2.3 SOFTWARE IMPLEMENTATION

Omics preprocessing

This thesis presents eUTOPIA, a software solution for analysis of microarray omics data. eUTOPIA makes the microarray data analysis accessible to the experimental researchers or to an end user who does not have expertise in statistical and computational tools to perform analysis. The tools employed in microarray analysis require a certain measure of expertise in statistics and computational methods, it discourages users from approaching the analysis or more worryingly leaves the novice users susceptible to pitfalls of failure in making necessary checks and corrections. eUTOPIA bridges this knowledge

gap by allowing the user to engage with each analytical step through a graphical interface that guides them forward and gives the feedback via a meaningful graphical representation of data. Ultimately, enabling a novice or an expert user to perform microarray data analysis with reproducible results and allow for more focus on the biological interpretation of results. The user can familiarize with eUTOPIA's functionality by following the user guide, which is provided as a supplementary to the published article II. The user guide highlights different components and features of eUTOPIA by showcasing the analysis of Agilent 2-color sample data.

Molecular networks

INfORM is an easy to use graphical tool that can be used to execute multi-step workflow for network based analysis of gene expression profile with a single click of a button (Figure 15). Its workflow is configured with a set of default parameters that were determined with a careful evaluation strategy (Figure 6-9). The user can confidently proceed with the analysis with minimal analytical setup and configuration. This allows the user to spend more time interpreting the results and less time struggling with the analytical steps. While the whole analytical setup is customizable and a more experienced user with specific requirement and ideas can alter the analytical setup to their needs.

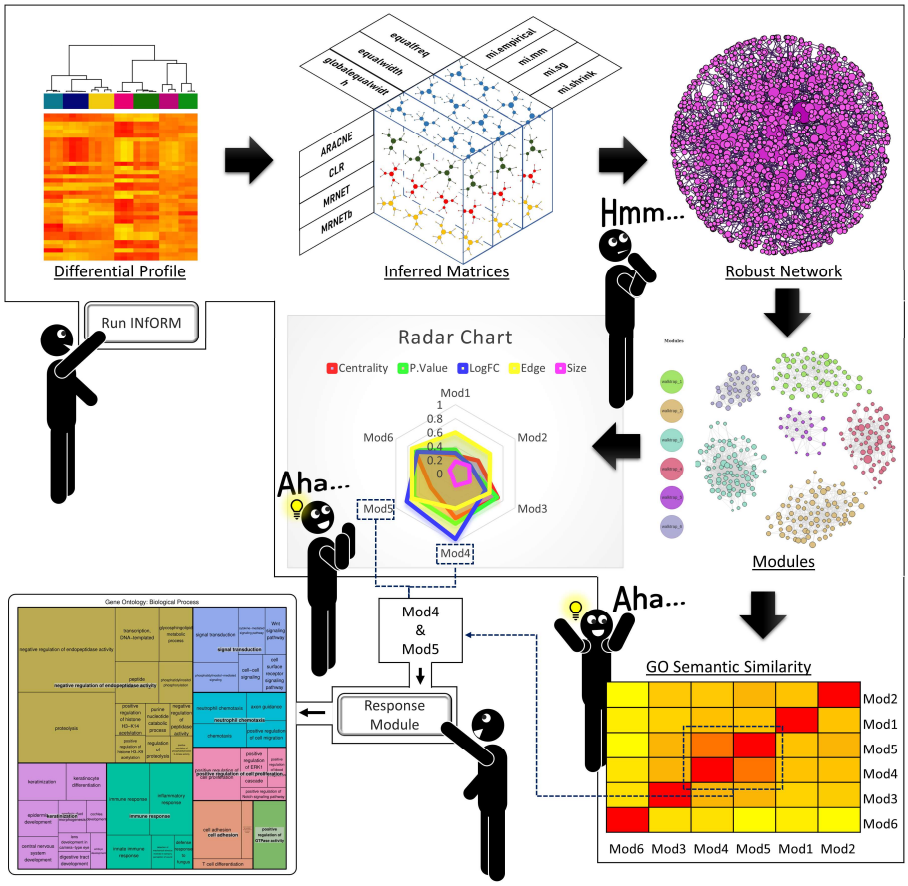

Fig. 15 – INfORM workflow execution is a simplified process that enables the user to build a network, identify modules of interest, define the response module, and interpret biological significance with minimal technical input.

9.3 MULTI-OMICS BASED APPROACH TO MODELING AOPS

I present here a study that we performed to observe the effect of 10 different carbon nanomaterials (4 different types). Three different cell lines representative of human lung resident cells, A549 (Human alveolar epithelium), BEAS-2B (bronchial epithelium), and THP-1 (differentiated macrophages), were exposed to the carbon nanomaterials and the response was observed by assay of mRNA expression, miRNA expression, and DNA methylation (Figure 16).

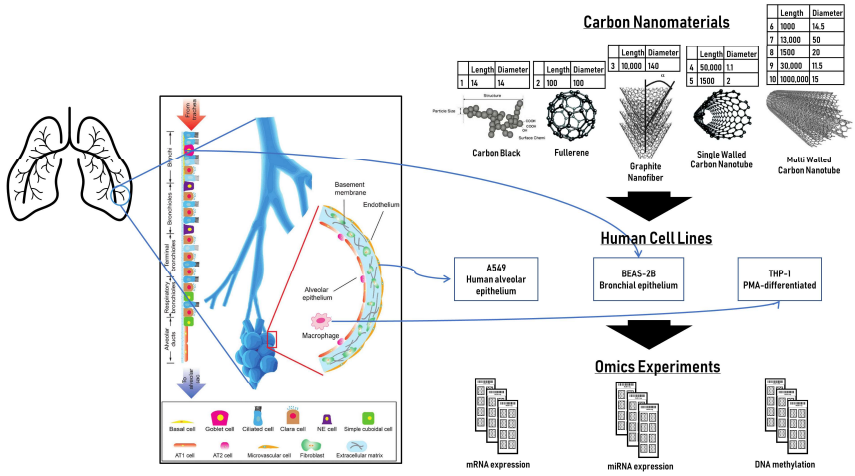

Fig. 16 - Multi-omics study design for observation of 10 carbon nanomaterial exposure on three cell lines derived from human lungs by assay of mRNA expression, miRNA expression, and DNA methylation.

These three different molecular profiles of carbon nanomaterial exposure response were chosen to define the gene regulatory model that can help in the interpretation of acute response as well as in inferring long term response. The relationship between different molecular layers was defined as the inverse correlation of mRNA expression information with the DNA methylation in the gene promoter, in addition to the inverse correlation with miRNA expression. Thus, we were able to define a model where the gene is concordantly upregulated, its promoter is hypomethylated, and a possible targeting miRNA is downregulated, vice-versa the concordant downregulation is defined by gene underexpression, hypermethylation of promoter, and overexpression of miRNA (Figure 17).

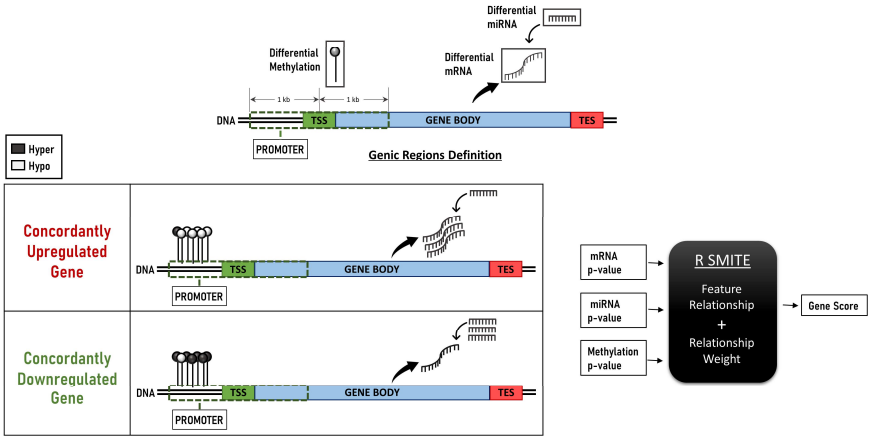

Fig. 17 - Multi-omics model of concordant gene response by using coherent information from mRNA expression, miRNA expression, and DNA methylation to obtain a singular gene score.

The concordance model describes the molecular feature relationships between mRNA, miRNA, and methylation. This relationship information, along with the weights of influence was specified in R SMITE analysis pipeline to summarize the p.values of significance from individual molecular assay to obtain a single gene score. We used the summarized gene score to obtain a weighted Reactome protein interaction network. Modules are detected in this weighted network by using the spinglass algorithm and were tested against random permutations to obtain significantly enriched modules. The total set of genes in the enriched modules were tested evaluated against the defined concordance model and segregated into concordant genes and discordant genes. Functional annotation enrichment was performed with KEGG pathways for both gene sets. Finally, we were able to obtain characteristic mechanism of action of for each CNM in each cell line (Figure 18) that can be used to hypothesize long term effect by concordant pathway maps and possible acute response by discordant pathway maps. By using this multi-omics model, we were able to observe and interpret the long term effect from a short exposure of 48 hours. This is a significant result for the in vitro toxicity exposure studies.

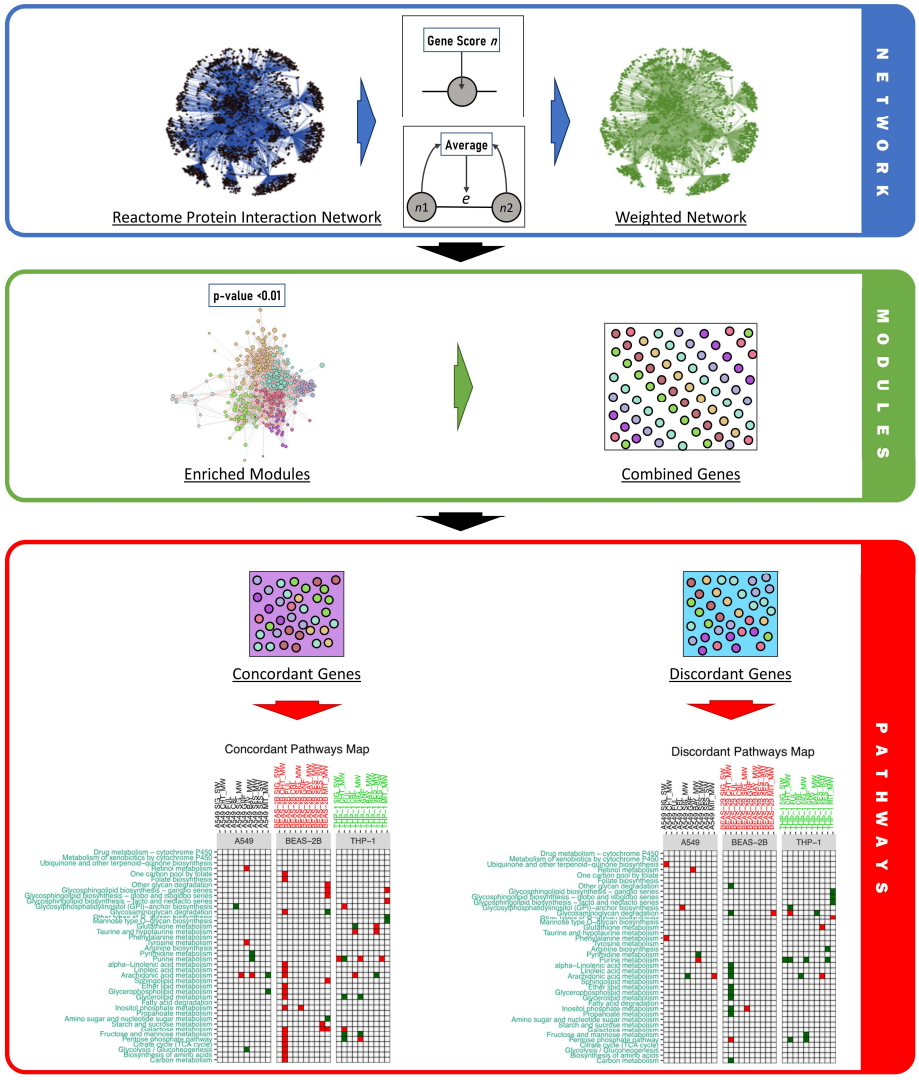

Fig. 18 - Identification of the functional mechanism of action by concordant and discordant genes from significantly enriched modules identified in weighted Reactome protein network.

The MOA derived from the KEGG based pathway analysis of functional genes along with concordance information from the integration of multiple omics layers is used to understand the finer differences between the responses of the cell lines A549 (alveolar epithelium), BEAS-2B (bronchial epithelium), and THP-1 (macrophages). Overall exposure response was strongest in BEAS-2B cells, THP-1 cells also exhibited a strong response, while A549 cells were least responsive. The overall MOA in BEAS-2B is found to be strongly biased toward upregulation while it was more balanced in A549 and THP-1 cells. We observed distinct cell line specific MOA in

response to the same CNM exposures, which can be explained by the different steady-state transcriptional patterns of the individual cell lines. The different type of cells have different mechanisms for recognizing foreign substances.

Evaluation of altered pathways suggests higher alteration potential of nanotubes with the exception of SES_SW which was observed to produce no effect on A549 cells, while the spherical CNMs with small aspect ratio was found to have a low potential to alter pathways.

10 DISCUSSION

Toxicology is an area of science that has a direct and immediate effect on human safety and thus has been under constant pressure for advancement and improvement. It is a discipline that is strictly regulated to ensure the application of best practices in evaluating the toxicity of new substances introduced into the environment and specifically designed for human consumption. The reform of toxicity assay practices under the proposed ideology of 3R is one of the most important steps forward towards modern science. The continued pursuit of reducing the cost of toxicological assay and the efforts to enrich the understanding of toxic exposure response from a biological system has motivated the incorporation of high-throughput omics technologies. Toxicogenomics evaluation in nanotoxicology is a recent development and not yet well established as a robust source for regulatory evidence. There have been scientific studies that showcase the potential of the omics in nanotoxicology (Fadeel *et al.*, 2013; Rydman *et al.*, 2014; Kinaret *et al.*, 2017b; Serra *et al.*, 2019). However, there is a shroud of uncertainty regarding its capability to generate conclusive and reproducible results. This is a major hurdle in the field of toxicogenomics and nanotoxicology. This challenge must be addressed by efforts towards standardization of omics data processing and reporting (Sauer *et al.*, 2017). Microarray omics data has rapidly become one of the most standardized and streamlined omics assay technologies that can be generated with swiftness and is cost-effective (Marzancola *et al.*, 2016). A number of well established computational methods and guidelines are available for processing omics data. These methods have evolved in parallel with the experimental techniques and experimental designs for better omics evaluation of the biological systems. The analysis methodology and data processing guidelines have been streamlined with defined steps. I identified a set of 'state-of-the-art' tools and computational methods widely accepted and employed by the scientific community that facilitate the standard analytics of microarray omics data. Microarrays allow for complex study designs with large cohorts of samples generated over multiple sets of experiments. Thus, exposure of a substance can be estimated in different aspects such as tissues, doses, and time. However, integration of data from multiple experiments also adds to potential noise due to varying levels of expressions in different experiments, and this noise can contribute to the variability of the genes and could inflate or deflate the expression thus giving rise to type I and type II errors. Large-scale integrated analysis of omics data must avoid the bias and noise by identifying the batches of data formed due to technical effects or biological effects over a subset of the samples (Chen *et al.*, 2011). On the contrary, there is a chance of overcorrection the data leading to positive bias, and caution must be observed while adjusting the data. It is not trivial to identify and

remove true batches while avoiding general pitfalls, diagnosis of data must be performed to observe the variation associated with the sample annotations and the effect of variation adjustment must be observed to determine the most appropriate strategy. Microarray data processing and analysis is a technically challenging task that requires computation expertise to execute analytical tools, build a seamless workflow for productivity, and generate meaningful representations of the data. There is a need for a guided user experience that lifts the burden of obligatory computational expertise from the user and enables them to process data with confidence. I designed an easy to use tool eUTOPIA for analysis of omics data. It incorporates the state-of-the-art tools and data analysis guidelines as a stepwise guided workflow that is accessible via an intuitive graphical interface.

The toxicological assessments have relied on the identification of significant biomolecules that can be attributed with the responsibility of producing the observed response and can be used as molecular markers (Timbrell, 1998). However, any biological response can seldom be explained by the activity of individual molecules. Over time Toxicogenomics has evolved into systems toxicology by the implementation of methods that infer system level activity due to combined molecular activity in the observed biological response (Waters and Fostel, 2004). Nanotoxicology and toxicogenomics studies, in general, must exploit the power of the dense high throughput and high content data generated from the modern omics techniques and technologies. It is trivial to estimate the behavior of a single molecule, rather the interaction of molecules in a functionally activate biological system must be estimated to understand the molecular mechanisms. Gene networks reconstructed from the transcriptomic data represent the statistical association in the molecular dynamics. Specific functional networks can be created by selecting the significantly altered molecules in the perturbed state compared to the steady state (Barel and Herwig, 2018). Graph theory can be used to identify hub genes which are bottlenecks that upon disruption can break network dynamics resulting in loss of communication between different parts of the network. In terms of systems toxicology, this could result in the disruption of biological pathways, thus producing an adverse effect. I created a tool INfORM for analysis of gene expression profiles by inferring a robust gene expression network to ensure reliability. This approach allows to confidently define gene-gene connections by observing evidence from multiple inferences. INfORM enabled the user in identifying the responsive subnetwork, and it also characterizes the network by identifying the most significant biological functions in the responsive subnetwork.

Toxicology has adopted the concept of networks in describing the biological events from the toxic response. The regulatory bodies for assessment of toxic substances such as OECD have defined programmes for describing the toxic response as AOPs. Gene co-expression networks reconstructed from the transcriptomic data represent the statistical

association in the molecular dynamics. Specific functional networks can be created by selecting significantly altered molecules in response to toxic exposure. The specific coherent molecular mechanisms can be observed as a modular structure in networks which represent the cluster of molecules in close correlation in contrast to the whole (Zhuang *et al.*, 2015). The nature of these molecular mechanisms can be understood by using known functional annotations. The biological interpretation of functional annotation is essential for understanding the molecular mechanisms, events, components, and pathways with active involvement in producing the adverse outcome. Nanotoxicology can adopt the module structure information from the biologically active gene networks to assist in the definition and identification of adverse outcome pathways (Hardt *et al.*, 2018).

The next avenue to enhance the biological characteristic of the toxicological response is the assessment of different biomolecules involved in the regulatory machinery (Jayapal, 2012). The information from individual molecular layers can be complemented and corroborated with the other layers to obtain a better interpretation of the biological response (Zhu *et al.*, 2012). The generally accepted model of gene regulation defines gene upregulation or over-expression in combination with hypomethylation in gene promoter and under-expression of gene targeting miRNA. While gene downregulation or under-expression is defined as a combination of hypermethylation in gene promoter and over-expression of gene targeting miRNA, this regulatory model is well accepted and confirmed by many omics studies explaining the molecular interactions and the biochemical process involved. We exploited this regulatory model to identify the significantly altered genes from gene expression data along with the methylation (Reamon-Buettner *et al.*, 2008) and miRNA information. Thus, enabling us to identify the biological pathways related to ENM exposure as the mechanism of action produced by exposure of human cell line to ENM. Exposure of different cell types to the same nanomaterial resulted in the different mechanism of action map, and this observation can be explained by the different steady state functional activity of the cell types and the specific roles of the cells. The acute and long-term effects can be inferred by the accordance of the regulatory model. The mRNA dynamics represent the immediate response of the perturbed biological system, while the DNA methylation dynamics represent long term response.

Another important aspect of the biological system is the study of temporal changes. Biological studies employ the omics technologies to obtain a snapshot of the biological system in a particular time point, this information can be extended to cover multiple time points, and the effect on the biological system can be observed over time. Over time the effort and interest to observe the temporal have resulted in a steady increase of timecourse data in public repositories. However, there is still a need for more efforts towards timecourse experiments and data modeling in Toxicogenomics. A cursory glance at GEO brings this to light. There are 3064 GSE (GEO Series) records

filtered by keyword toxic, out of which 455 (14.85%) GSE have at least two timepoints, while only 368 (12%) GSE records have at least three time points (Chen *et al.*, 2019). The timecourse data can be modeled as dynamic networks (Kim *et al.*, 2014) to understand the temporal change in nodes and edges to enrich the understanding of biological response in systems toxicology.

This thesis project contributed to improve the analysis of the omics data in the context of nanotoxicology studies by standardizing the data processing along with the implementation of robust and easy to use tools.

11 CONCLUSIONS

The toxicological evaluation of possible xenobiotic substances is outlined by different regulatory bodies such as Food and Drug Administration (FDA), U.S. Environmental Protection Agency (EPA), Organization for Economic Cooperation and Development (OECD), and European Chemical Agency (ECHA). The regulatory guidelines do not yet specify the analytical results from omics experiments as a piece of essential information for xenobiotic evaluation. My work presented in this thesis addresses the apparent areas of concern that have a hindered the usability of omics data in toxicology decision making. I target the core concern of reproducibility and reliability in omics analysis with my effort to standardize the data processing of microarray omics data. High throughput omics data is quite dense and must be processed appropriately in order to obtain accurate results. However, this cannot always be achieved to similar levels of satisfaction because there is lack of standardization. I incorporated the best practices and the state-of-the-art tools for omics data processing in a stepwise guided workflow. It enables the end user to perform consistent and reproducible analysis of the different datasets and repeat analyses of the same datasets.

I further addressed the challenge of estimating and interpreting the system level molecular mechanisms that explain the biological activities and events that are involved in the toxicological exposures. A straightforward approach to toxicogenomic assay data analysis is to estimate the set of significantly perturbed biomolecules and their associated biological functions. However, the biological system is in a constant flux of molecular interactions where the dynamics of a single molecule is not the sole contributor to the functional response. It is essential to interpret the dynamics of the biological system resulting from the correlated effect of individual dynamics. However, the inference of molecular relationships is subject to the choice of the algorithm, and it is a technically challenging approach which requires an understanding of graph theory. I implemented a methodology of combining the pieces of evidence from multiple network inference algorithms to obtain a robust gene network. Thus, ensuring the effective and accurate interpretation of the system level toxicological response. The interpretation of molecular mechanisms can be further enhanced by modeling the dynamics of different molecular species in the accepted regulatory model. This thesis showcases a study to determine the mechanism of action of nanomaterials in human cell lines derived from lung tissue. The information from different molecular signals mRNA, miRNA, and DNA methylation was used to create concordant dynamic model to explain the mechanisms of actions are concordantly expressed and could be used to model a consistent response in the short term and long term, alternately it

also allows to interpret the discordant response that might explain the short term acute response with mRNA or the long term chronic response with DNA methylation. The mechanistic information from the toxicological response is represented by a chain of biological events known as AOPs. I defined an approach to obtain the mechanistic information from the biological systems by that can be used to defined better AOPs by identifying the modules that represent specific biological functions.

The bioinformatics tools described in this thesis are designed to enable the general researchers in the processing and analysis of omics data by using state-of-the-art methods and best practices. These tools are accessible to the user via an intuitive and easy to use graphical interface that promotes the usability of these omics data processing and network analysis tools.

12 REFERENCES

Abts,H.F. *et al.* (2001) Sequence, Organization, Chromosomal Localization, and Alternative Splicing of the Human Serine Protease Inhibitor Gene Hurpin (PI13) Which Is Upregulated in Psoriasis. *DNA and Cell Biology*, **20**, 123–131.

Adeleye,Y. *et al.* (2015) Implementing Toxicity Testing in the 21st Century (TT21C): Making safety decisions using toxicity pathways, and progress in a prototype risk assessment. *Toxicology*, **332**, 102–111.

Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**.

Ahlers,J. *et al.* (2008) Integrated testing and intelligent assessment-new challenges under REACH. *Environ Sci Pollut Res Int*, **15**, 565–572.

Albert,R *et al.* (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.

Alving,A.S. *et al.* (1960) Mitigation of the haemolytic effect of primaquine and enhancement of its action against exoerythrocytic forms of the Chesson strain of Plasmodium vivax by intermittent regimens of drug administration: a preliminary report. *Bull. World Health Organ.*, **22**, 621–631.

Ankley,G.T. *et al.* (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.*, **29**, 730–741.

Aryee,M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25–29.

Ashton,D. and Porter,S. (2016) radarchart: Radar Chart from 'Chart.js'.

Barabasi, null and Albert, null (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barabási,A.-L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Barel,G. and Herwig,R. (2018) Network and Pathway Analysis of Toxicogenomics Data. *Front Genet*, **9**, 484.

Barel,G. and Herwig,R. (2018) Network and Pathway Analysis of Toxicogenomics Data. *Front Genet*, **9**.

Barnes,D.G. and Dourson,M. (1988) Reference dose (RfD): description and use in health risk assessments. *Regul. Toxicol. Pharmacol.*, **8**, 471–486.

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*, **41**, D991–D995.

Bellot,P. *et al.* (2015) NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference. *BMC Bioinformatics*, **16**.

Bellot,P. *et al.* (2014) grndata: Synthetic Expression Data for Gene Regulatory Network Inference.

Beutler,E. (1994) G6PD deficiency. *Blood*, **84**, 3613–3636.

Bilban,M. *et al.* (2002) Normalizing DNA microarray data. *Curr Issues Mol Biol*, **4**, 57–64.

Brandt-Rauf,P. *et al.* (2015) Genetic Screening in the Workplace: *Journal of Occupational and Environmental Medicine*, **57**, e17–e18.

Brockmeier,E.K. *et al.* (2017) The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment. *Toxicol Sci*, **158**, 252–262.

Bouhifd,M. *et al.* (2013) Review: Toxicometabolomics. *J Appl Toxicol*, **33**.

Bouhifd,M. *et al.* (2014) Mapping the human toxome by systems toxicology. *Basic Clin. Pharmacol. Toxicol.*, **115**, 24–31.

Bouhifd,M. *et al.* (2015) The Human Toxome Project. *ALTEX*, **32**, 112–124.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Carrella,D. *et al.* (2014) Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. *Bioinformatics*, **30**, 1787–1788.

Chang,W. *et al.* (2017) shiny: Web Application Framework for R.

Chen,C. *et al.* (2011) Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS One*, **6**.

Chen,G. *et al.* (2019) Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. *Database (Oxford)*, **2019**.

Chen,J. and Yuan,B. (2006) Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, **22**, 2283–2290.

Chen,K.-H. *et al.* (2015) Nanoparticle distribution during systemic inflammation is size-dependent and organ-specific. *Nanoscale*, **7**, 15863–15872.

Chen,P. *et al.* (2017) Co-expression network analysis identified six hub genes in association with metastasis risk and prognosis in hepatocellular carcinoma. *Oncotarget*, **8**, 48948–48958.

Chen,Y. *et al.* (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, **8**, 203–209.

Chepelev,N.L. *et al.* (2015) Integrating toxicogenomics into human health risk assessment: lessons learned from the benzo[a]pyrene case study. *Crit. Rev. Toxicol.*, **45**, 44–52.

Chiu,W.A. *et al.* (2013) Human Health Effects of Trichloroethylene: Key Findings and Scientific Issues. *Environ Health Perspect*, **121**, 303–311.

Collins,A.R. *et al.* (2017) High throughput toxicity screening and intracellular detection of nanomaterials. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*, **9**.

Cooper,T.F. *et al.* (2006) Effect of random and hub gene disruptions on environmental and mutational robustness in Escherichia coli. *BMC Genomics*, **7**, 237.

Costa,P.M. *et al.* (2018) Transcriptional profiling reveals gene expression changes associated with inflammation and cell proliferation following short-term inhalation exposure to copper oxide nanoparticles. *J Appl Toxicol*, **38**, 385–397.

Costello,J.C. *et al.* (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **32**, 1202–1212.

Croft,D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691-697.

Crump,K.S. (1984) A new method for determining allowable daily intakes. *Fundam Appl Toxicol*, **4**, 854–871.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Currie,H.N. *et al.* (2014) An approach to investigate intracellular protein network responses. *Chem. Res. Toxicol.*, **27**, 17–26.

Cury,Y. and Picolo,G. (2006) Animal toxins as analgesics--an overview. *Drug News Perspect.*, **19**, 381–392.

ORGANOPHOSPHATE TOXICITY. *Toxicology*, **307**, 115–122.

Dai,Y. *et al.* (2015) Performance of genetic risk factors in prediction of trichloroethylene induced hypersensitivity syndrome. *Sci Rep*, **5**.

Dankovic,D.A. *et al.* (2015) The Scientific Basis of Uncertainty Factors Used in Setting Occupational Exposure Limits. *J Occup Environ Hyg*, **12**, S55–S68.

Davis,J.A. *et al.* (2011) Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1. *Toxicol. Appl. Pharmacol.*, **254**, 181–191.

Dearden,J.C. (2003) In silico prediction of drug toxicity. *J. Comput. Aided Mol. Des.*, **17**, 119–127.

Dubertret,L. *et al.* (1984) Psoriasis: a defect in the regulation of epidermal proteases, as shown by serial biopsies after cantharidin application. *Br. J. Dermatol.*, **110**, 405–410.

Dunn,W.J. (1988) QSAR approaches to predicting toxicity. *Toxicol. Lett.*, **43**, 277–283.

Fabregat,A. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

Fadeel,B. *et al.* (2013) Nanotoxicology. *Toxicology*, **313**, 1–2.

Faith,J.J. *et al.* (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLOS Biology*, **5**, e8.

Fei,L. *et al.* (2017) A new method to identify influential nodes based on combining of existing centrality measures. *Mod. Phys. Lett. B*, **31**, 1750243.

Filipsson,A.F. *et al.* (2003) The benchmark dose method--review of available models, and recommendations for application in health risk assessment. *Crit. Rev. Toxicol.*, **33**, 505–542.

Fortin,J.-P. *et al.* (2014) shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Res*, **3**.

Friedman,N. and Rando,O.J. (2015) Epigenomics and the structure of the living genome. *Genome Res*, **25**, 1482–1490.

Fröhlich,E. (2017) Role of omics techniques in the toxicity testing of nanoparticles. *J Nanobiotechnology*, **15**, 84.

Gatto,L. (2017) yaqcaffy: Affymetrix expression data quality control and reproducibility analysis.

Gautier,L. *et al.* (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Geiser,M. (2010) Update on macrophage clearance of inhaled micro- and nanoparticles. *J Aerosol Med Pulm Drug Deliv*, **23**, 207–217.

Gopalacharyulu,P.V. *et al.* (2009) Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast. *Mol Biosyst*, **5**, 276–287.

Griffiths,T.L. *et al.* (2007) Google and the mind: predicting fluency with PageRank. *Psychol Sci*, **18**, 1069–1076.

Guengerich,F.P. (1998) The Environmental Genome Project: functional analysis of polymorphisms. *Environ Health Perspect*, **106**, 365–368.

Haga,S.B. and Kantor,A. (2018) Horizon Scan Of Clinical Laboratories Offering Pharmacogenetic Testing. *Health Aff (Millwood)*, **37**, 717–723.

Harden,J.L. *et al.* (2016) The Tryptophan Metabolism Enzyme, L-Kynureninase, is a Novel Inflammatory Factor in Psoriasis and other Inflammatory Diseases. *J Allergy Clin Immunol*, **137**, 1830–1840.

Hardt,C. *et al.* (2018) Computational Network Analysis for Drug Toxicity Prediction. *Methods Mol. Biol.*, **1819**, 335–355.

Hartung,T. (2011) From alternative methods to a new toxicology. *Eur J Pharm Biopharm*, **77**, 338–349.

Hartung,T. *et al.* (2017) Systems Toxicology: Real World Applications and Opportunities. *Chem Res Toxicol*, **30**, 870–882.

Huang,G.S. *et al.* (2010) Co-expression of GPR30 and ERbeta and their association with disease progression in uterine carcinosarcoma. *Am. J. Obstet. Gynecol.*, **203**, 242.e1–5.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18 Suppl 1**, S233-240.

Igarashi,Y. *et al.* (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res*, **43**, D921–D927.

Iizuka,H. *et al.* (2004) Unique keratinization process in psoriasis: late differentiation markers are abolished because of the premature cell death. *J. Dermatol.*, **31**, 271–276.

Jayapal,M. (2012) Integration of Next-Generation Sequencing Based Multi-Omics Approaches in Toxicogenomics. *Front Genet*, **3**.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jorgensen,A.L. *et al.* (2019) Implementation of genotype-guided dosing of warfarin with point-of-care genetic testing in three UK clinics: a matched cohort study. *BMC Med*, **17**, 76.

Joseph,P. (2017) Transcriptomics in toxicology. *Food Chem Toxicol*, **109**, 650–662.

Jiang,W. *et al.* (2008) Nanoparticle-mediated cellular response is size-dependent. *Nat Nanotechnol*, **3**, 145–150.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kauffmann,A. *et al.* (2009) arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.

Kauffmann,A. and Huber,W. (2010) Microarray data quality control improves the detection of differentially expressed genes. *Genomics*, **95**, 138–142.

Keermann,M. *et al.* (2015) Transcriptional landscape of psoriasis identifies the involvement of IL36 and IL36RN. *BMC Genomics*, **16**.

Kim,H. *et al.* (2015) Meta-Analysis of Large-Scale Toxicogenomic Data Finds Neuronal Regeneration Related Protein and Cathepsin D to Be Novel Biomarkers of Drug-Induced Toxicity. *PLoS One*, **10**.

Kim,Y. *et al.* (2014) Inference of dynamic networks using time-course data. *Brief. Bioinformatics*, **15**, 212–228.

Kinaret,P. *et al.* (2017a) Network Analysis Reveals Similar Transcriptomic Responses to Intrinsic Properties of Carbon Nanomaterials in Vitro and in Vivo. *ACS Nano*, **11**, 3786–3796.

Kinaret,P. *et al.* (2017b) Inhalation and Oropharyngeal Aspiration Exposure to Rod-Like Carbon Nanotubes Induce Similar Airway Inflammation and Biological Responses in Mouse Lungs. *ACS Nano*, **11**, 291–303.

Kirby,T.J. *et al.* (2016) Integrative mRNA-microRNA analyses reveal novel interactions related to insulin sensitivity in human adipose tissue. *Physiol Genomics*, **48**, 145–153.

Kleensang,A. *et al.* (2014) Pathways of Toxicity. *ALTEX*, **31**, 53–61.

Knudsen,T.B. *et al.* (2015) FutureTox II: In vitro Data and In Silico Models for Predictive Toxicology. *Toxicol Sci*, **143**, 256–267.

Kohonen,P. *et al.* (2017) A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nat Commun*, **8**, 15932.

Komatsu,N. *et al.* (2007) Aberrant human tissue kallikrein levels in the stratum corneum and serum of patients with psoriasis: dependence on phenotype, severity and therapy. *British Journal of Dermatology*, **156**, 875–883.

Kosarac,B. *et al.* (2009) Effect of genetic factors on opioid action. *Curr Opin Anaesthesiol*, **22**, 476–482.

Krewski,D. *et al.* (2010) TOXICITY TESTING IN THE 21ST CENTURY: A VISION AND A STRATEGY. *J Toxicol Environ Health B Crit Rev*, **13**, 51–138.

Kulski,J.K. *et al.* (2005) Gene expression profiling of Japanese psoriatic skin reveals an increased activity in molecular stress and immune response signals. *J Mol Med*, **83**, 964–975.

Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, **7**, 54–60.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lauss,M. (2017) swamp: Visualization, Analysis and Adjustment of High-Dimensional Data in Respect to Sample Annotations.

Lazar,C. *et al.* (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform*, **14**, 469–490.

LeCluyse,E.L. *et al.* (2012) Organotypic liver culture models: Meeting current challenges in toxicity testing. *Crit Rev Toxicol*, **42**, 501–548.

LeBeau,J.E. (1983) The role of the LD50 determination in drug safety evaluation. *Regul. Toxicol. Pharmacol.*, **3**, 71–74.

Leek,J.T. *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Liang,Y. *et al.* (2017) Psoriasis: a mixed autoimmune and autoinflammatory disease. *Curr. Opin. Immunol.*, **49**, 1–8.

Lima-Mendez,G. and van Helden,J. (2009) The powerful law of the power law and other myths in network biology. *Mol Biosyst*, **5**, 1482–1493.

Limonciel,A. *et al.* (2018) Persistence of Epigenomic Effects After Recovery From Repeated Treatment With Two Nephrocarcinogens. *Front Genet*, **9**.

Liu,C. *et al.* (2015a) Compound signature detection on LINCS L1000 big data. *Mol Biosyst*, **11**, 714–722.

Liu,C. *et al.* (2015b) Genome-wide gene-asbestos exposure interaction association study identifies a common susceptibility variant on 22q13.31 associated with lung cancer risk. *Cancer Epidemiol Biomarkers Prev*, **24**, 1564–1573.

Liu,Q.-P. *et al.* (2012) The association between GJB2 gene polymorphism and psoriasis: a verification study. *Arch Dermatol Res*, **304**, 769–772.

Lundqvist,M. *et al.* (2011) The evolution of the protein corona around nanoparticles: a test study. *ACS Nano*, **5**, 7503–7509.

Lundqvist,M. *et al.* (2008) Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 14265–14270.

Ma,H.-W. *et al.* (2004) An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res.*, **32**, 6643–6649.

Mahadevan,B. *et al.* (2011) Genetic Toxicology in the 21st Century: Reflections and Future Directions. *Environ Mol Mutagen*, **52**, 339–354.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat Methods*, **9**, 796–804.

Margolin,A.A. *et al.* (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, **7**, S7.

Marquis,B.J. *et al.* (2009) Analytical methods to assess nanoparticle toxicity. *Analyst*, **134**, 425–439.

Marzancola,M.G. *et al.* (2016) DNA Microarray-Based Diagnostics. *Methods Mol. Biol.*, **1368**, 161–178.

Meenu,S. *et al.* (2016) Modulation of host ubiquitin system genes in human endometrial cell line infected with Mycobacterium tuberculosis. *Med Microbiol Immunol*, **205**, 163–171.

Meldrum,K. *et al.* (2017) Mechanistic insight into the impact of nanomaterials on asthma and allergic airway disease. *Part Fibre Toxicol*, **14**, 45.

Meredith,J.R. *et al.* (2013) Biomedical applications of carbon-nanotube composites. *Front Biosci (Elite Ed)*, **5**, 610–621.

Merrick,B.A. and Witzmann,F.A. (2009) The role of toxicoproteomics in assessing organ specific toxicity. *EXS*, **99**, 367–400.

Meyer,P.E. *et al.* (2010) Information-theoretic inference of gene networks using backward elimination. *BIOCOMP International Conference Bioinformatics Computational Biology CSREA Press*, **2010**, 700-705.

Meyer,P.E. *et al.* (2008) minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Meyer,P.E. *et al.* (2007) Information-Theoretic Inference of Large Transcriptional Regulatory Networks. *EURASIP J Bioinform Syst Biol*, **2007**, 79879.

Michalak,P. (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, **91**, 243–248.

Moffat,I. *et al.* (2015) Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo[a]pyrene in drinking water. *Crit Rev Toxicol*, **45**, 1–43.

Morimoto,Y. *et al.* (2013) Inhalation toxicity assessment of carbon-based nanoparticles. *Acc. Chem. Res.*, **46**, 770–781.

Moussali,H. *et al.* (2005) Expression of hurpin, a serine proteinase inhibitor, in normal and pathological skin: overexpression and redistribution in psoriasis and cutaneous carcinomas. *Experimental Dermatology*, **14**, 420–428.

Muñoz-Galván,S. *et al.* (2015) MAP17 (PDZKIP1) Expression Determines Sensitivity to the Proteasomal Inhibitor Bortezomib by Preventing Cytoprotective Autophagy and NFκB Activation in Breast Cancer. *Mol Cancer Ther*, **14**, 1454–1465.

Mulas,F. *et al.* (2017) Network-based analysis of transcriptional profiles from chemical perturbations experiments. *BMC Bioinformatics*, **18**, 130.

Murty,B.S. *et al.* (2013) Unique Properties of Nanomaterials. In, Murty,B.S. *et al.* (eds), *Textbook of Nanoscience and Nanotechnology*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 29–65.

Nair,R.P. *et al.* (2009) Genomewide Scan Reveals Association of Psoriasis with IL-23 and NF-κB Pathways. *Nat Genet*, **41**, 199–204.

Nakajima,K. *et al.* (2013) Barrier Abnormality Due to Ceramide Deficiency Leads to Psoriasiform Inflammation in a Mouse Model. *Journal of Investigative Dermatology*, **133**, 2555–2565.

Nakayama,S. *et al.* (2009) A zone classification system for risk assessment of idiosyncratic drug toxicity using daily dose and covalent binding. *Drug Metab. Dispos.*, **37**, 1970–1977.

Okamoto,K. *et al.* (2012) Inhibition of Glucose-Stimulated Insulin Secretion by KCNJ15, a Newly Identified Susceptibility Gene for Type 2 Diabetes. *Diabetes*, **61**, 1734–1741.

Ota,T. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet*, **36**, 40–45.

Paine,M.F. (2017) Therapeutic disasters that hastened safety testing of new drugs. *Clin. Pharmacol. Ther.*, **101**, 430–434.

Parfett,C.L. and Desaulniers,D. (2017) A Tox21 Approach to Altered Epigenetic Landscapes: Assessing Epigenetic Toxicity Pathways Leading to

Altered Gene Expression and Oncogenic Transformation In Vitro. *Int J Mol Sci*, **18**.

Pavlopoulos,G.A. *et al.* (2011) Using graph theory to analyze biological networks. *BioData Min*, **4**, 10.

Pieroni,E. *et al.* (2008) Protein networking: insights into global functional organization of proteomes. *Proteomics*, **8**, 799–816.

Phillips,J.C. *et al.* (1990) Survey of the QSAR and in vitro approaches for developing non-animal methods to supersede the in vivo LD50 test. *Food Chem. Toxicol.*, **28**, 375–394.

Podila,R. and Brown,J.M. (2013) Toxicity of Engineered Nanomaterials: A Physicochemical Perspective. *J Biochem Mol Toxicol*, **27**, 50–55.

Ramena,G. *et al.* (2016) CLCA2 Interactor EVA1 Is Required for Mammary Epithelial Cell Differentiation. *PLOS ONE*, **11**, e0147489.

Rayner,T.F. *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.

Reamon-Buettner,S.M. *et al.* (2008) The next innovation cycle in toxicogenomics: environmental epigenetics. *Mutat. Res.*, **659**, 158–165.

Ridings,J.E. (2013) The thalidomide disaster, lessons from the past. *Methods Mol. Biol.*, **947**, 575–586.

Richard,A.M. *et al.* (2016) ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.*, **29**, 1225–1251.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**, e47.

Rocca-Serra,P. *et al.* (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, **26**, 2354–2356.

Rouquié,D. *et al.* (2015) Contribution of new technologies to characterization and prediction of adverse effects. *Crit. Rev. Toxicol.*, **45**, 172–183.

Roux,E. *et al.* (2002) In vitro effect of air pollutants on human bronchi. *Cell Biol. Toxicol.*, **18**, 289–299.

Ryan,N. *et al.* (2016) Moving Toward Integrating Gene Expression Profiling Into High-Throughput Testing: A Gene Expression Biomarker Accurately Predicts Estrogen Receptor α Modulation in a Microarray Compendium. *Toxicol Sci*, **151**, 88–103.

Rydman,E.M. *et al.* (2014) Inhalation of rod-like carbon nanotubes causes unconventional allergic airway inflammation. *Part Fibre Toxicol*, **11**, 48.

Schena,M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schimek,M.G. *et al.* (2015) TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat Appl Genet Mol Biol*, 311–6.

Schoenhard,J.A. *et al.* (2003) Regulation of the PAI-1 promoter by circadian clock components: differential activation by BMAL1 and BMAL2. *Journal of Molecular and Cellular Cardiology*, **35**, 473–481.

Schuck,R.N. and Grillo,J.A. (2016) Pharmacogenomic Biomarkers: an FDA Perspective on Utilization in Biological Product Labeling. *AAPS J*, **18**, 573–577.

Schultz,T.W. *et al.* (2015) A strategy for structuring and reporting a read-across prediction of toxicity. *Regul. Toxicol. Pharmacol.*, **72**, 586–601.

Serra,A. *et al.* (2019) INSIdE NANO: a systems biology framework to contextualize the mechanism-of-action of engineered nanomaterials. *Sci Rep*, **9**.

Shenfield,G.M. (2004) Genetic Polymorphisms, Drug Metabolism and Drug Concentrations. *Clin Biochem Rev*, **25**, 203–206.

Simeonova,P.P. and Erdely,A. (2009) Engineered nanoparticle respiratory exposure and potential risks for cardiovascular toxicity: predictive tests and biomarkers. *Inhal Toxicol*, **21 Suppl 1**, 68–73.

Slenter,D.N. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.

Spellman,P.T. *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **3**, RESEARCH0046.

Sporns,O. and Zwi,J.D. (2004) The small world of the cerebral cortex. *Neuroinformatics*, **2**, 145–162.

Sauer,U.G. *et al.* (2017) The challenge of the application of 'omics technologies in chemicals risk assessment: Background and outlook. *Regulatory Toxicology and Pharmacology*, **91**, S14–S26.

Suárez-Fariñas,M. *et al.* (2015) RNA sequencing atopic dermatitis transcriptome profiling provides insights into novel disease mechanisms with potential therapeutic implications. *Journal of Allergy and Clinical Immunology*, **135**, 1218–1227.

Sumantran,V.N. *et al.* (2016) Microarray Analysis of Differentially-Expressed Genes Encoding CYP450 and Phase II Drug Metabolizing Enzymes in Psoriasis and Melanoma. *Pharmaceutics*, **8**, 4.

Sun,K. *et al.* (2014a) Predicting disease associations via biological network analysis. *BMC Bioinformatics*, **15**, 304.

Sun,K. *et al.* (2014) The integrated disease network. *Integr Biol (Camb)*, **6**, 1069–1079.

Supek,F. *et al.* (2011) REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, **6**, e21800.

Swindell,W.R. *et al.* (2011) Genome-Wide Expression Profiling of Five Mouse Models Identifies Similarities and Differences with Human Psoriasis. *PLOS ONE*, **6**, e18266.

Te,J.A. *et al.* (2016) Systems toxicology of chemically induced liver and kidney injuries: histopathology-associated gene co-expression modules. *J Appl Toxicol*, **36**, 1137–1149.

Tennekes,M. (2017) treemap: Treemap Visualization.

Thomas,D.G. *et al.* (2013) ISA-TAB-Nano: a specification for sharing nanomaterial research data in spreadsheet-based format. *BMC Biotechnol.*, **13**, 2.

Timbrell,J.A. (1998) Biomarkers in toxicology. *Toxicology*, **129**, 1–12.

Truong,N.P. *et al.* (2015) The importance of nanoparticle shape in cancer drug delivery. *Expert Opin Drug Deliv*, **12**, 129–142.

Tsunekawa,N. *et al.* (2016) Heparanase augments inflammatory chemokine production from colorectal carcinoma cell lines. *Biochemical and Biophysical Research Communications*, **469**, 878–883.

Tunesi,S. *et al.* (2015) Gene-asbestos interaction in malignant pleural mesothelioma susceptibility. *Carcinogenesis*, **36**, 1129–1135.

Tyner,J.W. (2017) Integrating functional genomics to accelerate mechanistic personalized medicine. *Cold Spring Harb Mol Case Stud*, **3**.

Van Damme,K. *et al.* (1997) Ethical issues in genetic screening and genetic monitoring of employees. *Ann. N. Y. Acad. Sci.*, **837**, 554–565.

Van den Bulcke,T. *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7**, 43.

van Leeuwen,K. *et al.* (2009) Using chemical categories to fill data gaps in hazard assessment. *SAR QSAR Environ Res*, **20**, 207–220.

van Noort,V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Varshney,P. *et al.* (2016) Transcriptome profiling unveils the role of cholesterol in IL-17A signaling in psoriasis. *Sci Rep*, **6**.

Ventura,C. *et al.* (2018) Conventional and novel 'omics'-based approaches to the study of carbon nanotubes pulmonary toxicity. *Environ. Mol. Mutagen.*, **59**, 334–362.

Verbelen,M. *et al.* (2017) Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet? *Pharmacogenomics J.*, **17**, 395–402.

Villeneuve,D.L. *et al.* (2014) Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicol. Sci.*, **142**, 312–320.

Wang,X. *et al.* (2018) Toxicological Profiling of Highly Purified Single-Walled Carbon Nanotubes with Different Lengths in the Rodent Lung and Escherichia Coli. *Small*, **14**, e1703915.

Warnes,G.R. *et al.* (2016) gplots: Various R Programming Tools for Plotting Data.

Waters,M.D. and Fostel,J.M. (2004) Toxicogenomics and systems toxicology: aims and prospects. *Nature Reviews Genetics*, **5**, 936–948.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Wijetunga,N.A. *et al.* (2017) SMITE: an R/Bioconductor package that identifies network modules by integrating genomic and epigenomic information. *BMC Bioinformatics*, **18**, 41.

Wilsmann-Theis,D. *et al.* (2016) Among the S100 proteins, S100A12 is the most significant marker for psoriasis disease activity. *J Eur Acad Dermatol Venereol*, **30**, 1165–1170.

Wu,Z. *et al.* (2018) Network-Based Methods for Prediction of Drug-Target Interactions. *Front Pharmacol*, **9**, 1134.

Yamane,J. *et al.* (2016) Prediction of developmental chemical toxicity based on gene networks of human embryonic stem cells. *Nucleic Acids Res.*, **44**, 5515–5528.

Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.

Yu,G. *et al.* (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.

Yu,H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.

Zhang,D.-Q. *et al.* (2017) Identification of hub genes and pathways associated with bladder cancer based on co-expression network analysis. *Oncol Lett*, **14**, 1115–1122.

Zhang,L. *et al.* (2013) Identification of biomarkers for hepatocellular carcinoma using network-based bioinformatics methods. *Eur J Med Res*, **18**, 35.

Zhao,F. and Klimecki,W.T. (2015) Culture conditions profoundly impact phenotype in BEAS-2B, a human pulmonary epithelial model. *J Appl Toxicol*, **35**, 945–951.

Zhuang,D.-Y. *et al.* (2015) Identification of hub subnetwork based on topological features of genes in breast cancer. *Int. J. Mol. Med.*, **35**, 664–674.

Zhu,J. *et al.* (2012) Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. *PLoS Biol*, **10**.