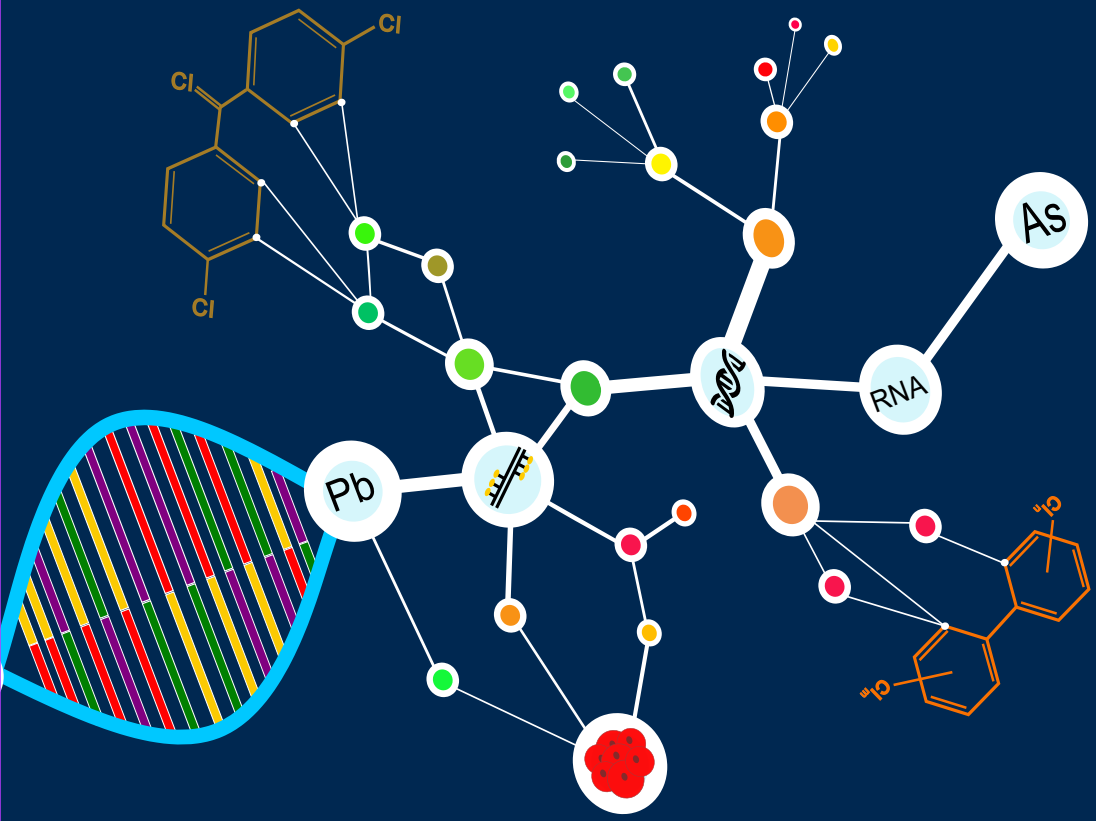
Linking the epigonome and transcriptome

Linking the epigenome and transcriptome

Integration and visualization of omics data



Tim J.M. Kuijpers

Linking the epigenome and transcriptome

Integration and visualization of omics data

Linking the epigenome and transcriptome: integration and visualization of omics data
Layout: Tim J.M. Kuijpers
Cover design: Tim J.M. Kuijpers
Printed by: Proefschriftmaken
ISBN: 978-94-6423-534-0

Linking the epigenome and transcriptome

Integration and visualization of omics data

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. dr. Rianne M. Letschert
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op woensdag 1 december 2021 om 13:00 uur

door

Tim Josephus Maria Kuijpers
Geboren op 8 februari 1988 te ’s-Hertogenbosch, Nederland

Promotores
Dr. D.G.J. Jennen
Prof. dr. J.C.S. Kleinjans

Beoordelingscommissie
Prof. dr. ir. I.C.W. Arts (Voorzitter)
Prof. dr. J.A. Aerts (Universiteit Hasselt, België)
Prof. dr. ir. N.A.W. van Riel (Technische Universiteit Eindhoven, Maastricht University)
Dr. M.A.T. Teunis (Hogeschool Utrecht)

The research described in this thesis was conducted at GROW School for Oncology and Developmental Biology of Maastricht University.

Table of Contents

Chapter 1
General introduction

The human body consists of tiny building blocks called cells. These cells provide structure for the body, convert nutrients into energy, and carry out specialized functions. To carry out all these functions, our cells make use of different “helper” structures called proteins. Proteins are complex molecules that assist the cell in maintaining its structure but also regulate and perform the cell’s function. Although proteins are very important, there is no protein storage pool but there is a well-balanced system of protein synthesis and degradation [1]. Cells have the capability to synthesis these proteins, as long as they have the correct building instruction.

This building instruction is located on the DNA, in the core of a cell. DNA does not leave this core since this will increase the likelihood of DNA damage. Therefore, it makes use of a messenger carrying the building instruction, called messenger RNA (mRNA). This mRNA molecule transfers to the cellular components responsible for protein synthesis to deliver the building instructions. Although the process of transcription (DNA to mRNA) and translation (mRNA to protein) sounds straightforward, mRNA expression is a complex process controlled at multiple levels [2]. There are various regulation mechanisms, including epigenetic, transcriptional, and post-transcriptional regulation that coordinates these processes. It is vital that there are no large perturbations to these cellular processes.

Exposure to xenobiotics – compounds foreign to our cells – such as environmental toxins, persistent organic pollutants, heavy metals, and drugs can cause these perturbations. Although not all perturbations are harmful to a cell [3], there is increasing evidence relating environmental pollutants to the development of diseases [4–6]. Current findings suggest a correlation between persistent organic pollutants and gene expression changes [7, 8], as well as a link with DNA methylation alterations [9]. However, there is still much unknown about the relations between exposure to xenobiotics and alterations in gene expression and DNA methylation.
To reduce this knowledge gap, we aim to study the effect of xenobiotics on both DNA methylation and gene expression. Combining multiple data sources can compensate for missing information in one source, and multiple sources pointing towards the same pathways or processes are more reliable [10]. Furthermore, multiple views for the same patient or sample can provide complementary information on the response to a xenobiotic.

Systems biology approaches focus on the integration of different biological entities, by studying the different molecular interactions. Unlike traditional approaches that focus on isolated components, like gene or protein expression, the new approaches focus on a more holistic view. Through the integration of different data sets, it becomes possible to study the effect of exposure to xenobiotics and the subsequent changes in DNA methylation and gene expression levels. Although recent technological advancements have made more multi-omics data sets available, integrating these multiple views is still challenging [11, 12]. There is great potential for machine learning approaches to elucidate the molecular events leading to disease development and progression. However, due to a classical problem, called

“big p, small n”, in which there are more features than samples, many statistical machine learning approaches are inappropriate. In this thesis, we will try to address these problems and design a workflow for the integration of omics data.

Transcriptome: the expression of RNA

As the name suggests, transcriptomics is a technique to study the transcriptome of a subject consists of all the measured RNA transcripts [13]. The information for an RNA transcript is stored on the DNA. DNA stores our genetic information by using a simple but efficient four-letter code that resembles four chemical bases: Adenine (letter A), Guanine (letter G), Cytosine (letter C), and Thymine (letter T) [14]. The sequence of these bases determines the information available for building and maintaining a cell. The part of the DNA that acts as the instruction for building a protein is called a gene. Through transcription, the information on the DNA is copied into a new molecule called messenger RNA (mRNA). This mRNA molecule contains the “message” or building code for a protein and is vital for protein synthesis. Transcriptomics makes it possible to study the expression of those mRNAs.

Epigenetics: the relationship between DNA methylation and gene expression

Epigenetics refers to any heritable change in a cell that does not involve a change in the primary DNA sequence. Epigenetics is involved in many cellular processes and through genetic control can switch genes on or off, thus are important factors in regulating gene expression [15]. This phenomenon allows our bodies to contain many different types of cells (with the same DNA) such as liver cells, pancreatic cells, brain cells, and many others [16].

One of the major epigenetic mechanisms involving the modification of DNA is DNA methylation [17]. DNA methyltransferases catalyze DNA methylation by transferring a methyl group from S-adenyl methionine (SAM) to the fifth carbon of a cytosine residue to form 5mC. It is highly specific and always happens in a region where a cytosine is located next to a guanine, called a CpG site. If there is a higher concentration of CpG sites in one region of the DNA, we call this region a CpG island. Regions known as CpG islands are often associated with gene promoters and thus play a central role in gene regulation [18]. Almost all promoter associated CpG islands are usually free of DNA methylation, regardless of the transcriptional state of the gene and are hypomethylated in most tissues [19, 20].

The challenges in omics integration

From a biological point of view, the integration of different biological sources seems a step forward: it will give us more information about a biological system. From a computational point of view, it is not a straightforward process. One of the main challenges is to have high-quality data, with a high number of samples measured at the same time under the same condition. Recent improvements in technologies, such as RNA sequencing for transcriptomics, have made it possible to study a system in more detail. However, large data set with multiple omics layers and many measurements are rare, especially in the field of toxicogenomics. Furthermore,

because of their inherent difference, integrating multiple omics layers is still an ongoing challenge. To understand the complex relations between the different omics layers, one first has to understand the complex dependencies and challenges that arise when integrating different omics platforms. These challenges have a broad spectrum, going from experimental design to integration issues and leading up to the final question: how can we extract biological knowledge?

The curse of dimensionality

The current omics technologies can measure a high number of features for a given sample. The current microarray techniques can measure around 50.000 genes, whereas the current DNA methylation techniques can even measure from 200.000 up to 2.000.000 CpG sites. Although more information can be vital to understand biological systems, it brings a major challenge in analyzing the data. The high number of features introduces a new problem. In almost all data sets, there is an imbalance between the number of samples and the number of features, called “the curse of dimensionality”. Due to the high number of features, analyzing these results can be computationally intensive. Therefore, it is of great interest to reduce the number of features to a set of only important features associated with the phenotype.

The distribution of each ‘omics measurement

We measure the omics layers with different techniques and therefore the experimental data of those layers do not always follow the same distribution. For example, transcriptomic data has values that follow a distribution of positive numbers [0, inf), whereas epigenetic data can be expressed as either M-values or β-values. The β-value measures the percentage of methylation, whereas the M-value is the log2 ratio of the intensities of the methylated probe versus the unmethylated probe. We have to ask ourselves if we have to handle each layer differentially during the integration procedure, or if we can assume that both layers are just numbers. Here it is important to look at the range of values one could find in a given data set and what the meaning of these values is. We can conclude that a β-value of 0 for a CpG site (i.e. hypomethylated) has a different meaning than a value of 0 for a gene (i.e. not being expressed). When choosing an integration strategy, we should make sure that the algorithm could handle data sets with different distributions.

Omics integration strategies

There are three main categories of omics integration strategies: early integration, intermediate integration, and late integration [21]. The early integration strategy concatenates all omics layers into one big matrix. This makes it possible to apply the current single-omics algorithms to the data. Late integration is the opposite of the early integration strategy. The late integration strategy applies an algorithm to each layer individually. Finally, the outcome of the algorithm per layer is combined to obtain an omics integrated solution.

There are a couple of disadvantages to the previously mentioned integration strategies. Late integration implies that both omics layers are independent of each

other, although we know from biology that the epigenome and transcriptome are two processes tightly connected. In the early integration strategy, data concatenation can lead to imbalanced data sets when one of the omics layers is larger. As a result, the large omics layers might have more influence on the model's outcome. At the same time, the early integration strategy ignores the distribution of each omics layer.

Intermediate integration strategies try to overcome these problems by simultaneously integrating each omics layer. This strategy respects the difference in biological entities and distribution for each layer. Because the integration strategy uses complementary information in multiple layers, we believe this strategy is the best option to integrate the transcriptome and epigenome to understand toxicity and cancer-related events. Therefore, we will look further into algorithms capable of performing an intermediate integration strategy.

Machine learning techniques: unsupervised vs supervised learning
The current set of machine learning tools can be categorized into two classes: unsupervised or supervised learning [22]. In supervised learning, we use a set of input and output variables for which we want to learn the mapping function from input to output. To learn this mapping function, supervised learning makes use of a label that tells us what the output is. As an example, we can take a set of measurements from cancer-diagnosed or healthy subjects [23]. We train a model, the so-called mapping function, and learn this model how to approximate the relationship between input and output variables. During training, the labels are used as the ground truth and teach the model how to distinguish between cancer and healthy. After training the model, we can use it to classify new patients for which we do not know yet if they have cancer and classify them into either healthy or cancer. One major drawback of supervised learning is the use of labels since you have to be certain about the classification of your training data.

Unsupervised learning algorithms do not use a dependent variable or know labels. The goal of the algorithm is to infer the underlying structure of the data and to learn more about these structures in the data [24].

The method of choice, unsupervised or supervised clustering, depends mainly on the data set and problem you want to address. Supervised clustering needs many samples to build the model, which is not a problem in large studies like the cancer genome atlas (TCGA [25]) but a problem in the much smaller toxicogenomics data sets. Unsupervised clustering could be more suitable in this case since it can work with smaller data sets. Furthermore, unsupervised clustering does not need any label for training which can be beneficial if the effect of exposure to a compound is unknown. In population studies, we often see that we measure the exposure profile of an individual, but we do not know if that exposure is low or high. Moreover, if an individual is exposed to a mixture of compounds it becomes even harder to determine the correct output label. Here, an unsupervised learning method could be used, to determine to which group each individual belongs. In the end, one could

compare each group to unravel if there is an elevated exposure to a compound and if there is a causal relationship between the exposure and the health status of that group.

Unsupervised learning techniques: matrix factorization
Matrix factorization belongs to the class of unsupervised techniques and is applied to transform a high-dimensional data set into a low-dimensional structure while preserving the most important information. The most popular matrix factorization methods are principal component analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF). Although in theory their goal is the same, the reduce the data set to a low-dimensional structure, they follow different approaches. PCA aims to identify as much variability in the data by finding principle components that maximize the variance [26]. ICA and NMF try to learn distinct patterns from the data by applying a different technique [27]. ICA learns factors that are statistically independent, whereas NMF tries to identify the patterns driving a cluster.

As previously mentioned, the techniques follow different objectives but also different constraints. NMF imposes a non-negativity constraint on the input data, whereas PCA enforces orthogonality, and ICA maximizes the separability of the data. The principal components (or basis vectors) in PCA can be ranked by the extent to which they explain the variation in the data, and not all principal components are equally important. In both ICA and NMF, the basis vectors are assumed to have equal weight [27]. One advantage of NMF over PCA is that the basis vectors can be assigned to a specific cluster. Therefore, it is possible to score the variables and extract those variables driving the clustering of samples. Another advantage of NMF is the nonnegativity constraint. The principal components (PCA) and the basis vectors (ICA) both contain positive and negative coefficients. These mixed signals indicate that modeling the original data involves complex cancellations between the positive and negative values. NMF projects the data onto a basis space only containing positive values. It has been shown that those positive values in the basis space can be scored to extract more relevant biological molecules.

It is therefore of interest to see if NMF is suitable to detect transcriptomic and epigenetic patterns driving a phenotype.

Nonnegative matrix factorization: method to overcome the challenges in omics integration?
Nonnegative matrix factorization (NMF) is a technique designed to reduce the dimensionality of the original data matrix to two smaller new matrices (W and H) with much smaller dimensions (Figure 1.1). These two smaller matrixes can be used as a reduced representation of the data samples in the data matrix [28]. NMF could be a great option to tackle the challenges we have to address for the integration of the different omics data sets.

We can consider an input matrix X whose rows contain the measured probes and the columns the samples (Figure 1.1). For omics data, the number of measured probes is higher than the number of samples. We can also expect that not all measured probes explain the observed phenotype in our samples. As mentioned earlier, the goal of NMF is to find a small number of feature probes in W that can explain the k clusters with their samples in H.

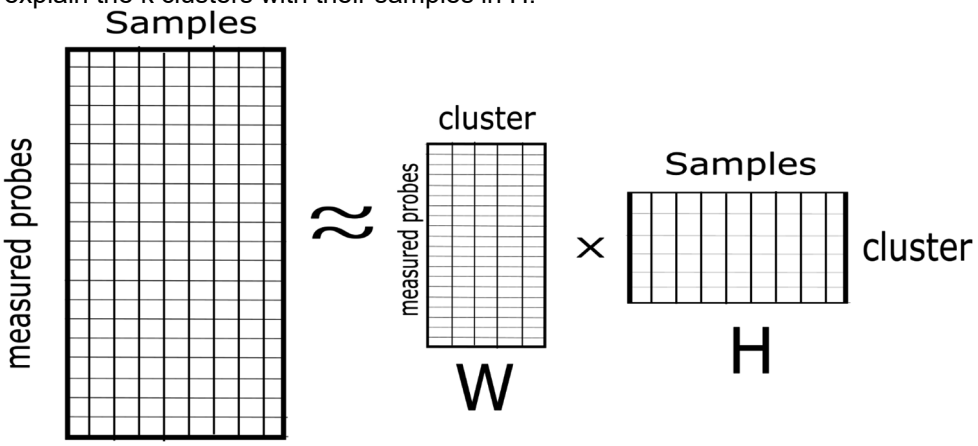


Figure 1.1 Nonnegative matrix factorization**.** The matrix containing the sample and probe values are decomposed in to two smaller matrices W (number of probes as rows, clusters as columns) and H (clusters as rows and samples as columns).

The value of k reflects the number of clusters hidden in our data and is therefore nontrivial to choose the optimal value for k. Different measures can be used to determine the optimal rank k, such as the cophenetic clusters coefficient or the silhouette coefficient. By performing 50 simulations with varying values for k, we can quantify the optimal value for k for our final simulation.

Given the input matrix X with the desired rank k, NMF computes the approximation X≈WH (Equation 1) while minimizing the objective function (Equation 2). Here, the objective function selected is called the Frobenius norm, based on the Euclidean distance. Note that this is not the only distance metric and others can be used. In the first step, both W and H are initialized with random numbers. In the second step, we iteratively update W (Equation 3) and H (Equation 4), meaning we first update W after which we update H. The second step is repeated until the object function does not longer decrease. Although the multiplicative update rules are a bit slow, it is guaranteed to monotonically reduce the objective function. Once a solution is found, it is possible to calculate the probability of a probe (in W) explain the cluster in H. The obtained values in H can be used to divide samples into their clusters.

$$X \approx WH \quad (1)$$

$$f(W,H) = \frac{1}{2}||A - WH||_F^2 \quad (2)$$

$$W_{w+1} = W * \frac{\frac{X}{WH}H^T}{\sum H} \quad (3)$$

$$H_{H+1} = H * \frac{\sum \frac{X}{WH}^{T}}{\sum W^{T}} \quad (4)$$

NMF is an unsupervised approach and therefore does not need any prior knowledge. It has been used to answer various research questions [29, 30] and was successful in separating different cancer cell lines [31, 32]. Furthermore, it can handle different data sets of different sizes as long as the number of samples are equal. In theory, this would mean that we could integrate different omics platforms, such as DNA methylation and gene expression. Earlier research has proposed NMF as a promising method for the integration of omics data [33, 34].

However, NMF alone cannot address all of the challenges in the integration of omics data. After identifying features explaining the data, it cannot tell us the relation between the proposed features. In other words, it can propose different genes and DNA methylation regions, but it cannot tell how these will affect biological processes. One major objective of multi-omics studies is to discover new biomarkers [35, 36] and therefore the proposed candidate makers should have biological relevance. Studies using multi-omics data without applying biological knowledge frequently end with the nomination of molecules in the network that are less relevant [37]. Nowadays, multiple databases exist that can give us valuable information about the molecules in our network. Functional annotation and ontology data, such as gene ontology [38, 39], pathway information [40], disease association [41], and drug-target interactions [42, 43] give us insight into the role of the different network molecules.

Therefore, we will extend the workflow and use biological networks to study the effects on biological processes, interactions between multiple genes, and the effect of DNA methylation changes on gene expression.

The role of biological networks in omics integration

Biological networks are a powerful tool for data analysis. They are used to study the effect of perturbations on a biological system, to identify potential targets for drug discovery [43, 44] and to study disease-based mechanisms [44, 45]. Because of their design, biological networks are useful to visualize complex information in a simplified way.

We visualize biological networks via network graphs, with nodes referring to molecules and edges representing the interaction among molecules. These edges can be either directed or undirected. To define how a biological network is translated to a network graph, let us take the set of genes defined by its members: ‘gene A’, ‘gene B’, and ’gene C’ (Figure 1.2A). Gene A is a transcription factor of gene B and through its interaction can control the transcription of gene B. Therefore, we define an interaction between gene A and gene B as “Gene A interacts with gene B” by drawing a line between the two nodes (Figure 1.2B). We have created an undirected network from which we can see that genes A and B share an interaction. However,

we cannot tell from this interaction what kind of interaction gene A and gene B share. We still miss important information about the transcriptional property of gene A. To add this information to the network, we can draw a line with an arrowhead (Figure 1.2C). We call this property the directionality of the interaction. From figure 1.2C, we now can see that gene A and gene B share an interaction but more importantly, we can even see that gene A exerts an effect on gene B. We can improve our understanding of the edges in our network if we color the interactions based on their effect, green for activation and red for inhibition (Figure 1.2D). If one has a look at a network such as displayed in figure 1.2D, it becomes clear that gene A activates the transcription of gene B, gene B activates the transcription of gene C, and finally, gene C can inhibit the transcription of gene B.

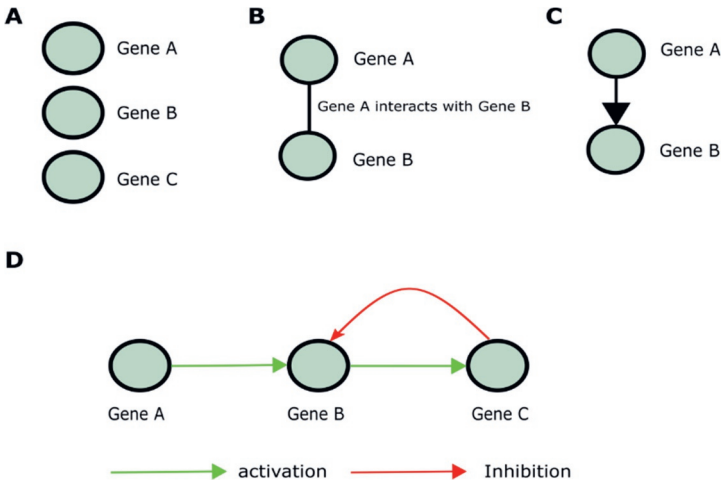


Figure 1.2 How to create a network graph. **A**: We represent a set of genes as nodes (circles). **B**: If two nodes interacted with each other, we define an edge (line) between the two nodes. **C**: To increase our understanding of the interactions, we can try to define if an interaction has a directionality. Here the interaction means Gene A interacts with gene B but not the other way around. **D**: Extra visualization options can add information about the interaction type to the network.

Network graphs and hubs

In a network, we can define the degree of a node by the number of interactions (i.e. edges). For directed graphs, we can define the inner-degree as the number of edges directed at the node and the outer-degree as the number of edges going out of the node. If we calculate the degree for all the nodes, we can define so-called HUB nodes. HUB nodes have a higher number of edges in comparison with other nodes in the network and are believed to play central roles in the network [46,47]. It is suggested that biological networks are robust against perturbations, but disruption of HUB genes causes the system to fail [48, 49] and are often associated with disease development [50, 51].

Community detection

Since research gives us more knowledge about the different interactions between genes, such as transcriptional activation or inhibition, the number of edges between

nodes can grow very quickly. This will make a network graph more complex to analyze and harder to gain new knowledge.

Here, we hypothesize that in biological systems, molecules do not work independently but in a coordinated way, where molecules involved in the same biological processes are likely to be highly connected. Therefore, we can look for communities of highly connected genes in a network and study different events in those communities. An important property of a community is that it is separable from other modules [52] and its members have more connections among themselves [53].

In this thesis, the Louvain method will be applied on the genomic interaction networks. The Louvain method is a simple and efficient method to identify communities in very large networks. The algorithm uses two phases to identify communities in a network. First, the algorithm looks for smaller communities to optimize the modularity locally. Second, it aggregates nodes that belong to the same communities to build a larger community. These two steps are repeated until the modularity stops to increase. Here, modularity is a measure of the numbers inside communities compared to links between communities. Thus, we expect to obtain high modularity when we have a low number of edges between communities and a high number of edges within a community.

Dynamic biological networks
Biological systems are not static processes but are highly dynamic where interactions between molecules or the expression of genes change over time. In toxicogenomics, it is vital to understand these dynamic processes as a response to chemical exposure in biological systems and therefore we need to study the effect of time [54, 55]. A classical approach in network biology is to visualize these dynamic processes by color-coding the nodes based on their expression value. However, visualizing only changes in expression is not enough. Multiple studies show perturbations in interactions between nodes due to chemical exposures. For example, Wolters et al [56] showed by exposing primary human hepatocytes to VPA a relation between perturbations in gene expression and gene-gene interactions. Therefore, it is not only important to visualize the changes in expression values but also the change in interactions between nodes.

The gap between epigenetics and transcriptomics in toxicogenomics
A single omics technique will detect biomolecules in one layer, and thus captures changes only for small subsets of the components in a biological system. Therefore, applying single omics analysis in toxicogenomics led to the identification of biomarkers for certain exposures but not a systemic understanding of toxicity. Epigenetics processes are significantly modulated by exposure to a compound, and its downstream effect is vital to understand the mechanisms of perturbations on a transcriptome level. The integration of multiple omics datasets will improve our understanding of the underlying biology, and therefore we will gain a better mechanistic aspect of the system.

Research and aim of this thesis

The abundance of biological data has made data integration approaches increasingly popular over the past decade. Understanding cellular processes and molecular interactions by integrating molecular networks has just been one of the challenges in data integration. In this thesis, the main objective is to gain more insight into the effect of alternations on the transcriptome and epigenome through omics integration. We hypothesize that through the integration of multiple omics data sets, we can increase our understanding of the relationship between environmental exposure, DNA methylation, and gene expression.

In **chapter 2.1**, we created a network visualization tool to visualize a dynamic biological network and at the same time show biological knowledge from different data sources. In **chapter 2.2**, we have created a python work frame to construct multi-omics networks from curated information about gene-gene interactions, drug-gene interactions, disease-gene interactions, and DNA methylation – gene interactions.

In **chapter 3**, we investigated the use of nonnegative matrix factorization to integrate DNA methylation and gene expression from two cancer data sets. In **chapter 3.1**, we tested the workflow on the NCI60 data set, a small data set of cancer cell lines with gene expression (microarray platform), and DNA methylation (Illumina 450K methylation array). **In chapter 3.2**, we performed a simulation on the 2019 Cancer Cell Line Encyclopedia to investigate the use of the proposed workflow on a large data set with high heterogeneity for which we have gene expression (RNA-sequencing platform) and DNA methylation (RBBS platform).
In **chapter 4**, we investigated the integrative molecular effects of chronic human exposure to environmental pollutants, by applying a cross-omics computational approach to investigate the relation between exposure and alteration in gene expression and/or DNA methylation in European cohort studies. These cohort studies contain subjects exposed to persistent environmental pollutants and heavy metals during their lifetime.

In **chapter 5**, we further deepened our understanding between the epigenome and the transcriptome by using single nucleotide polymorphisms (SNPs). SNPs are forms of DNA variation among individuals and may influence promoter activity, messenger RNA, conformation, and subcellular localization of mRNA. Therefore, alterations on the DNA due to SNP can be a missing piece to understand the link between the epigenome and transcriptome and is of interest to study. To investigate this hypothesis, we made use of gene expression, DNA methylation, and SNP information from a Pakistani cohort study exposed to arsenic-polluted drinking water. We hypothesized, taking into account polymorphism, DNA methylation, and gene expression with arsenic exposure, we can identify exposure-related profiles for subgroups with different susceptibility to arsenic exposure, independently of the dose of arsenic exposure.

References

1. Alberts B, Johnson A, Lewis J et al. Molecular biology of the cell. In: Molecular biology of the cell. 2002.
2. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012;28:2458–66.
3. COMAN G, DRAGHICI C, CHIRILA E, SICA M. POLLUTANTS EFFECTS ON HUMAN BODY – TOXICOLOGICAL APPROACH. In: Chemicals as Intentional and Accidental Global Environmental Threats. Dordrecht: Springer Netherlands. p. 255–66. doi:10.1007/978-1-4020-5098-5_19.
4. Kim HS, Kim YJ, Seo YR. An Overview of Carcinogenic Heavy Metal: Molecular Toxicity Mechanism and Prevention. J Cancer Prev. 2015;20:232–40. doi:10.15430/JCP.2015.20.4.232.
5. Huat TJ, Camats-Perna J, Newcombe EA, Valmas N, Kitazawa M, Medeiros R. Metal Toxicity Links to Alzheimer's Disease and Neuroinflammation. J Mol Biol. 2019;431:1843–68. doi:10.1016/j.jmb.2019.01.018.
6. Freeman MD, Kohles SS. Plasma Levels of Polychlorinated Biphenyls, Non-Hodgkin Lymphoma, and Causation. J Environ Public Health. 2012;2012:1–15. doi:10.1155/2012/258981.
7. Mitra PS, Ghosh S, Zang S, Sonneborn D, Hertz-Picciotto I, Trnovec T, et al. Analysis of the toxicogenomic effects of exposure to persistent organic pollutants (POPs) in Slovakian girls: Correlations between gene expression and disease risk. Environ Int. 2012;39:188–99. doi:10.1016/j.envint.2011.09.003.
8. Espín-Pérez A, Hebels DGAJ, Kiviranta H, Rantakokko P, Georgiadis P, Botsivali M, et al. Identification of Sex-Specific Transcriptome Responses to Polychlorinated Biphenyls (PCBs). Sci Rep. 2019;9:746. doi:10.1038/s41598-018-37449-y.
9. van den Dungen MW, Murk AJ, Kampman E, Steegenga WT, Kok DE. Association between DNA methylation profiles in leukocytes and serum levels of persistent organic pollutants in Dutch men. Environ Epigenetics. 2017;3. doi:10.1093/eep/dvx001.
10. Kim D. Methods of integrating data to uncover genotype – phenotype interactions. Nat Publ Gr. 2015;16:85–97. doi:10.1038/nrg3868.
11. Canzler S, Schor J, Busch W, Schubert K, Rolle-Kampczyk UE, Seitz H, et al. Prospects and challenges of multi-omics data integration in toxicology. Arch Toxicol. 2020;94:371–88. doi:10.1007/s00204-020-02656-y.
12. López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, et al. Challenges in the Integration of Omics and Non-Omics Data. Genes (Basel). 2019;10:238. doi:10.3390/genes10030238.
13. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLOS Comput Biol. 2017;13:e1005457. doi:10.1371/journal.pcbi.1005457.
14. WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171:737–8. doi:10.1038/171737a0.
15. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33:245–54. doi:10.1038/ng1089.
16. Skinner MK. Role of epigenetics in developmental biology and transgenerational inheritance. Birth Defects Res Part C Embryo Today Rev. 2011;93:51–5. doi:10.1002/bdrc.20199.
17. Holliday R. DNA methylation and epigenetic mechanisms. Cell Biophys. 1989;15:15–20. doi:10.1007/BF02991575.
18. Long HK, King HW, Patient RK, Odom DT, Klose RJ. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. Nucleic Acids Res. 2016;44:6693–706. doi:10.1093/nar/gkw258.
19. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. 2007;39:457–66. doi:10.1038/ng1990.
20. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, et al. Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol. 2009;16:564–71. doi:10.1038/nsmb.1594.
21. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion. 2019;50:71–91. doi:10.1016/j.inffus.2018.09.012.
22. Dey A. Machine Learning Algorithms: A Review. Int J Comput Sci Inf Technol. 2016;7:1174–9. www.ijcsit.com.
23. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17. doi:10.1016/j.csbj.2014.11.005.
24. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019;20:76. doi:10.1186/s13059-019-1689-0.

25. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature Genetics. 2013;45:1113–20.
26. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans R Soc A Math Phys Eng Sci. 2016;374:20150202. doi:10.1098/rsta.2015.0202.
27. Stein-O'Brien GL, Arora R, Culhane AC, Favorov A V., Garmire LX, Greene CS, et al. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. 2018;34:790–805. doi:10.1016/j.tig.2018.07.003.
28. Lee DD, Seung H.S. Algorithms for non-negative matrix factorization. Adv Neural Inf Proccessing Syst. 2000;:556–62.
29. Devarajan K, Ebrahimi N. Class Discovery via Nonnegative Matrix Factorization. Am J Math Manag Sci. 2008;28:457–67. doi:10.1080/01966324.2008.10737738.
30. Shao C, Höfer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. Bioinformatics. 2017;33:235–42. doi:10.1093/bioinformatics/btw607.
31. Frigyesi A, Höglund M. Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes. Cancer Inform. 2008;6:CIN.S606. doi:10.4137/CIN.S606.
32. Ma X, Gu J, Wang K, Zhang X, Bai J, Zhang J, et al. Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection. Sci Rep. 2019;9:7073. doi:10.1038/s41598-019-43548-1.
33. Fujita N, Mizuarai S, Murakami K, Nakai K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Sci Rep. 2018;8:9743. doi:10.1038/s41598-018-28066-w.
34. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS One. 2017;12.
35. Olivier M, Asmis R, Hawkins GA, Howard TD, Cox LA. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. Int J Mol Sci. 2019;20:4781. doi:10.3390/ijms20194781.
36. Hu Z-Z, Huang H, Wu CH, Jung M, Dritschilo A, Riegel AT, et al. Omics-Based Molecular Target and Biomarker Identification. 2011. p. 547–71. doi:10.1007/978-1-61779-027-0_26.
37. Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62:R21–45. doi:10.1530/JME-18-0055.
38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9. doi:10.1038/75556.
39. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011;39 suppl_1:D712–7. doi:10.1093/nar/gkq1156.
40. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39 Database:D691–7. doi:10.1093/nar/gkq1018.
41. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40:D940–6. doi:10.1093/nar/gkr972.
42. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36 suppl_1:D901–6. doi:10.1093/nar/gkm958.
43. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 2019;47:D948–54. doi:10.1093/nar/gky868.
44. Hao T, Wang Q, Zhao L, Wu D, Wang E, Sun J. Analyzing of Molecular Networks for Human Diseases and Drug Discovery. Curr Top Med Chem. 2018;18:1007–14. doi:10.2174/1568026618666180813143408.
45. Mulder NJ, Akinola RO, Mazandu GK, Rapanoel H. Using biological networks to improve our understanding of infectious diseases. Comput Struct Biotechnol J. 2014;11:1–10. doi:10.1016/j.csbj.2014.08.006.
46. Goymer P. Why do we need hubs? Nat Rev Genet. 2008;9:651–651. doi:10.1038/nrg2450.
47. Zotenko E, Mestre J, O'Leary DP, Przytycka TM. Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Comput Biol. 2008;4:e1000140. doi:10.1371/journal.pcbi.1000140.
48. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008;6:e264. doi:10.1371/journal.pbio.0060264.
49. Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, et al. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci. 2018;14:124–36. doi:10.7150/ijbs.22619.

50. Fu Y, Zhou Q-Z, Zhang X-L, Wang Z-Z, Wang P. Identification of Hub Genes Using Co-Expression Network Analysis in Breast Cancer as a Tool to Predict Different Stages. Med Sci Monit. 2019;25:8873–90. doi:10.12659/MSM.919046.
51. Wei S, Chen J, Huang Y, Sun Q, Wang H, Liang X, et al. Identification of hub genes and construction of transcriptional regulatory network for the progression of colon adenocarcinoma hub genes and TF regulatory network of colon adenocarcinoma. J Cell Physiol. 2020;235:2037–48. doi:10.1002/jcp.29067.
52. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47–52. doi:10.1038/35011540.
53. Tornow S. Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res. 2003;31:6283–9. doi:10.1093/nar/gkg838.
54. Cavill R, Kleinjans J, Briedé J-J. DTW4Omics: Comparing Patterns in Biological Time Series. PLoS One. 2013;8:e71823. doi:10.1371/journal.pone.0071823.
55. Bar-joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nat Publ Gr. 2012;13:552–64. doi:10.1038/nrg3244.
56. Wolters JEJ, van Breda SGJ, Grossmann J, Fortes C, Caiment F, Kleinjans JCS. Integrated 'omics analysis reveals new drug-induced mitochondrial perturbations in human hepatocytes. Toxicol Lett. 2018;289:1–13. doi:10.1016/j.toxlet.2018.02.026.

Chapter 2
Network visualization and knowledge integration

Chapter 2.1: DYNOVIS: a web tool to study dynamic perturbations for capturing dose-over-time effects in biological networks

T.J.M. Kuijpers[1],*, J.E.J. Wolters[1, 2], J.C.S. Kleinjans[1] and D.G.J. Jennen[1]
1 Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

Abstract
Background: The development of high-throughput sequencing techniques provides us with the possibility to obtain large data sets, which capture the effect of dynamic perturbations on cellular processes. However, because of the dynamic nature of these processes, the analysis of the results is challenging. Therefore, there is a great need for bioinformatics tools that address this problem.

Results: Here we present DynOVis, a network visualization tool that can capture dynamic dose-over-time effects in biological networks. DynOVis is an integrated work frame of R packages and JavaScript libraries and offers a force-directed graph network style, involving multiple network analysis methods such as degree threshold, but more importantly, it allows for node expression animations as well as a frame-by-frame view of the dynamic exposure. Valuable biological information can be highlighted on the nodes in the network, by the integration of various databases within DynOVis. This information includes pathway-to-gene associations from ConsensusPathDB, disease-to-gene associations from the Comparative Toxicogenomics databases, as well as Entrez gene ID, gene symbol, gene synonyms, and gene type from the NCBI database.

Conclusions: DynOVis could be a useful tool to analyze biological networks that have a dynamic nature. It can visualize the dynamic perturbations in biological networks and allows the user to investigate the changes over time. The integrated data, from various online databases, makes it easy to identify the biological relevance of nodes in the network. With DynOVis we offer a service that is easy to use and does not require any bioinformatics skills to visualize a network.

Published in : BMC Bioinformatics, 2019

Background

The development of high-throughput sequencing techniques allows us to obtain complex data sets, which are capable of revealing molecular responses of cellular processes [1]. For instance, changes in gene expression play a role in signal transduction mechanisms, metabolic pathways, and responses to harmful events in the cell [2]. For enabling a deeper mechanistic understanding these data sets have necessitated the development of mathematical models. For example, transcriptomic data have been used to construct gene regulatory networks, which provide valuable insights into the regulatory mechanisms of differential gene expression [3–5].

However, it is well known that these changes in biological processes are not static but dynamic. Therefore, currently, high-throughput sequencing techniques are combined with time-series experiments. Although this approach will increase the overall knowledge of dynamic cellular responses, the temporal analysis adds another layer of complexity Multiple tools have been designed to translate experimental results into network graphs for the purpose of studying time series data [6, 7] or dynamic perturbations, such as dose-over-time effects [8] . Currently, there are a number of different tools available that focus on visualizing dynamic networks [9–11], but these do not focus on the integration of dynamic visualization with biological knowledge.

To improve upon this, we developed a tool that integrates dynamic network visualization with functional biological information. To visualize and understand the influence of time and dose on cellular responses, integration of high-throughput sequencing time series data has been implemented. To enable a functional interpretation of different nodes and interactions in the established networks, information extraction methods have been developed to pull relevant information from various biological databases.

Implementation

DynOVis offers an easy web-based tool for visualizing dynamic gene expression data on a biological network and is made freely available at https://bitbucket.org/mutgx/dynovis/src for downloading and running locally. DynOVis is an integrated work frame of R packages and JavaScript libraries and is made freely available using the R Shiny package. R shiny is used to control the webpage and the web server, whereas the D3js JavaScript library is used to visualize the network. DynOVis is designed to guide the user through the different steps that translate a network structure into a network image. The user can decide to upload a static or dynamic network (directed or undirected) but also if they want to map biological knowledge onto the network. After the network has been constructed, the user can start analyzing the network (Figure 2.1).

Initialization and creation of the network

DynOVis allows for handling different network structures, that can either be an edge list (format: A interacts with B), adjacency matrix (format: matrix of 0 and 1's that define an interaction between nodes), or a Cytoscape file (in case previous analyses of the relevant data set has been performed using Cytoscape). We have provided a workflow that handles the processing of the network file into a network structure. Therefore, DynOVis does not require the user to specify the extension of the network file but it will automatically detect the file type and process it. The network structure is then converted into a network graph by applying the D3js force-directed graph algorithm (Figure 2.1A). This algorithm calculates the position of every node by applying an attractive force between each pair of connected nodes, as well as a repulsive force between the nodes. This will create a network with the least amount of overlapping nodes, which is important for studying any network. After the network has been created, the user is also free to drag and place nodes at different places and change the node positions.

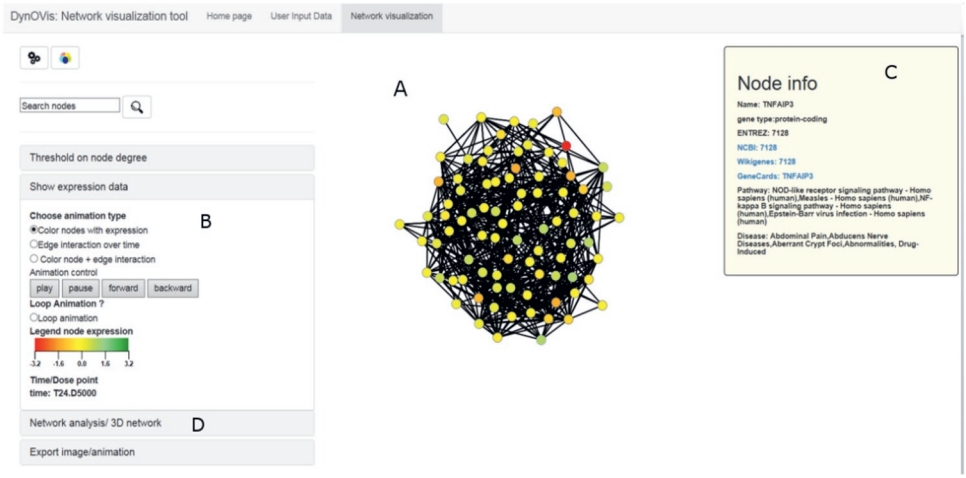


Figure 2.1 DynOVis: network visualization frame. In the center, the network is displayed (**A**). The control panel with the threshold degree and animation controls are placed at the left (**B**) and the biological information panel for each node is shown at (**C**) after a node has been clicked. Different graph theory features have been integrated to identify important HUB nodes (**D**).

Analyzing experimental data in your network

In its most general appearance, a network is a collection of nodes and the interaction between the nodes is defined by edges. These nodes may represent different entities, for instance, genes if one has built a gene regulatory network or CpG islands if one has built an epigenomic network. To investigate the relationship between changes in expression of nodes, the user may also upload a file containing expression data generated to visualize dose, time, or a combination of dose-over-time series experiments (Figure 2.1B). DynOVis does not request for a specific expression format, such as intensity or log fold change and the user is free to use their own desired format.

DynOVis applies a mapping function to assign the experimental data to the corresponding node and translates the value into a color by a red/green or red/blue linear color scale function. To determine the domain of the color scale, a built-in DynOVis function calculates the maximum absolute value of the experimental data is calculated and used as the left and right boundary of the scale. By using the absolute maximum value, we ensure that a dark red color (meaning very low expression) is equal to a bright green (or blue) color (meaning high expression).

Analyzing dynamic events on the network
DynOVis offers different options for analyzing dynamic perturbations in a biological network to study the changes in expression or interactions between nodes in the network (Figure 2.1, B). First, the user can upload a dynamic node expression file, from which DynOVis builds an animation that shows the changes in expression over time (see first example case study). Second, it is also possible to upload a dynamic edge interaction file that can be used to study changes in interactions between nodes (see second example case study). Third, both the dynamic edge and node expression file may be used to create an animation that shows how differences in node expression are associated with differences in edge interactions. This animation is to be viewed as a video, but also frame-by-frame by using the forward and backward animation control buttons. DynOVis has a built-in function that creates a 2D line graph, showing the changes in node expression if the user selects a local neighborhood of nodes. The degree of each node may change over time, which is visualized by either changing the degree of the nodes during the animation or by looking at the node degree table. This will highlight the importance of certain nodes at a given time point. Furthermore, it is possible to save the animation and use it in a presentation or on a website.

Integrated database to the network
To increase the understanding of the different nodes in a network, it is possible to add biological knowledge to the network by making use of the internal database of DynOVis. This database is built from various online sources such as ConsensusPathDB [12], Comparative Toxicogenomics database [13], NCBI database [14] and allows the user to access this information without having to visit any of those websites. To get the biological information for each node, the node identifier is connected with the corresponding key identifier in the DynOVis database. The user can select during the initial stage of building the network to incorporate gene information for Homo sapiens, Rattus Norvegicus, or Mus Musculus. This information includes pathway-to-gene associations from ConsensusPathDB [12], disease-to-gene associations from the Comparative Toxicogenomics databases [13], as well as Entrez gene ID, gene symbol, gene synonyms, and gene type from the NCBI database [14]. Here, we have developed a function that shows a pop-up window with the biological information that only becomes visible when the user clicks on a node (Figure 2.1C).

Change the node size based on their degree
Further important features in DynOVis refer to the basic network analysis options (Figure 2.1D). These options are derived from graph theory properties and give an overview of the most important nodes, based on degree centrality. Nodes with many connected neighbors have a high degree and are therefore called HUB nodes. It is important to identify these HUB nodes since biological networks are robust against perturbations, but disruption of pivotal nodes in general causes the system to [15, 16]. In DynOVis, we have split degree centrality into three components: i) total degree, ii) inner degree and iii) outer degree. This separation helps to identify a node that is regulated by many other neighbors (inner degree) or a node that regulates many nodes (outer degree). The visualization style of the nodes (i.e. the size, a property defined by D3js) can also be changed according to the different degrees, or nodes in the network can be hidden if a chosen degree is below a certain threshold.

Study the network in three dimensions
The two-dimensional network (Figure 2.1, A) can also be converted into a three-dimensional network. This allows the user to navigate through the network in a first-person view while at the same time playing the animation showing the dynamic perturbation. DynOVis uses a custom-developed 3D force-directed layout by adding an extra formula to calculate the z-position with respect to the x and y values of the 2D D3js force-directed layout by defining the following formula:

$$z = Y * 0.8 * sin * (y_{value} * \pi) + X * 0.8 * \sin{(x_{value} * \pi)}$$

Finding pathways in your network
For each gene in the network, the associated pathways are searched for in the database that is attached to DynOVis. Since one gene may be associated with multiple pathways, the number of different pathways may be very high, depending on the number of genes in the network. Therefore, DynOVis ranks the pathways based on the number of genes associated and returns the top 10 pathways as a table. DynOVis provides a function to download the complete list of genes and associated pathways.

Results and Discussion
As has been discussed in the introduction, there are multiple applications that aim to visualize biological networks. Here, DynOVis will be compared with the most popular tool at this moment: Cytoscape. Cytoscape is a stand-alone network visualization tool that offers multiple network analysis methods.

The workflow of DynOVis is designed in such a way from start to end, thus uploading a network file to visualizing a network, is straightforward to the user. Here, a comparison will be made between the workflow of DynOVis and Cytoscape. Both Cytoscape and DynOVis can handle different input formats, including .txt, .csv or .sif files. Whereas DynOVis has an automated feature to translate the two

columns of input data (parent node and child node) to a network, Cytoscape first asks the user to specify the parent and child node column. This allows the user more freedom with respect to the number of columns in their input file, whereas for DynOVis the first and second column must imply the network structure. Although the automated process restricts the user in their choices, it makes the visualization process more straightforward.

One of the most important features is the dynamic network visualization. DynOVis directly offers this feature, whereas for Cytoscape different apps need to be installed from their app store. The Cytoscape app store contains multiple applications that allow for dynamic network visualization, such as DyNet [17], CyAnimator [11], DyNetViewer [18], and ANIMO [19]. ANIMO has been excluded from the comparison since the focus is more on signaling networks. In most cases, the user would like to study the perturbations of nodes over time or dose points in one biological network. Both DynOVis and CyAnimator have been built to show dynamic expression changes in nodes. DynOVis takes the dynamic input for every node and assigns these values as a color to each node, while directly building the animation. For CyAnimator, the user has built each frame manually before the frame can be added to the animation. Each time point requires a step in which the node expression values have to be set as fill color, which will be stored as a key frame. Consequently, if there are ten different time points the user has to build ten key frames for each data point before the animation can be created. DynOVis saves the users some time because they only need to push the play button to start the animation.

Animations serve very well to identify important expression changes in the network, for instance, a gene that shows a strong downregulation between two time points. If these events are observed, the user would like to analyze the local network around this node. DynOVis makes this possible by highlighting a local neighbor network around a selected node and the user may play the animation or analyze the changes frame-by-frame. It is also possible to drag and drop the nodes to new places and start the animation. However, these two features are not possible while using CyAnimator. The node and edge positions are fixed after creating the animation and the animation is only visible in one orientation. If a local neighbor network around a node has to be analyzed, the user has to create a new animation by hand.

Dynamic node interactions may be studied by using DynOVis, DyNetViewer, or DyNet. DyNet is a Cytoscape app that visualizes differences among multiple networks, such as edges between nodes. This is of interest if one wants to study protein-protein-interaction networks in different tissues. The different node degree calculations can be performed to get the most important nodes between each network. DyNetViewer can generate a dynamic network by adding time course data to a static network. This will result in an animation with the different edges between

nodes for the different time points. DynOVis can also generate an animation of the different changes in node interactions after the user uploads a file containing the interaction information per time point. The advantage of DynOVis over DyNet is that the latter requires at least two networks to compare, whereas DynOVis shows dynamic edge interactions in one network. DyNetViewer also shows dynamic edge interaction, but it calculates the interactions from the node expression values in a static network. This means that if one obtains a dynamic network with already known interaction changes, it cannot be uploaded in DyNetViewer. Here DynOVis has the advantage, because it can map dynamic edge interactions onto a network. DyNetViewer has an advantage if the dynamic edge interactions are yet unknown because it calculates them from node expression data alone, something DynOVis cannot do from scratch. However, DyNetViewer cannot show the expression data on the nodes while playing the animation. DynOVis is capable of showing this information and therefore the user will see whether changes in node expression are actually altering node interactions.

With DynOVis we offer the implementation of dynamic network, functional, and graph theory analysis without limitations with respect to the type of network. It is possible to study dynamic effects without the need for multiple networks while the tool immediately provides the user with information about gene function, associated pathways, and diseases. Although Cytoscape is more powerful in analyzing a static network, DynOVis has some advantages over the current Cytoscape applications regarding dynamic visualization (see Table 2.1 for an overview). The combination of dynamic visualization and functional biological annotation is one of the advantages of DynOVis. Some of these functionalities are demonstrated in the following case studies.

Case studies on time series drug treatment experiments

In the first example, the network describing the NF-kB pathway has been investigated, a pathway related to apoptosis and inflammation, key events for both drug-induced liver fibrosis and cholestasis [20]. Here, a dose-over-time network of the NK-kB pathway was constructed with DTNI [8] from human in vitro samples exposed to acetaminophen from TG-GATEs [21]. Data were measured for two biological replicates at three doses (low, middle, high) at three time points (2, 8, and 24 hours). By investigating the acetaminophen dose-response of the NF-kB pathway over time, it becomes apparent that both time and dose, play an important role in the changes in gene expression. From the animation it becomes clear that the induced effect of the acetaminophen challenge is dose- and time-specific because the alternations in gene expression are only observed at a high dose after a period of 8 hours (Figure 2.2A). TNF superfamily member 11 (*TNFSF11*) and toll-like receptor 4 (*TLR4*) show a strong downregulation at 8 hours (Figure 2.2B), whereas C-X-C motif chemokine ligand 2 (*CXCL2*) shows a strong downregulation at 24 hours (Figure 2.2C).

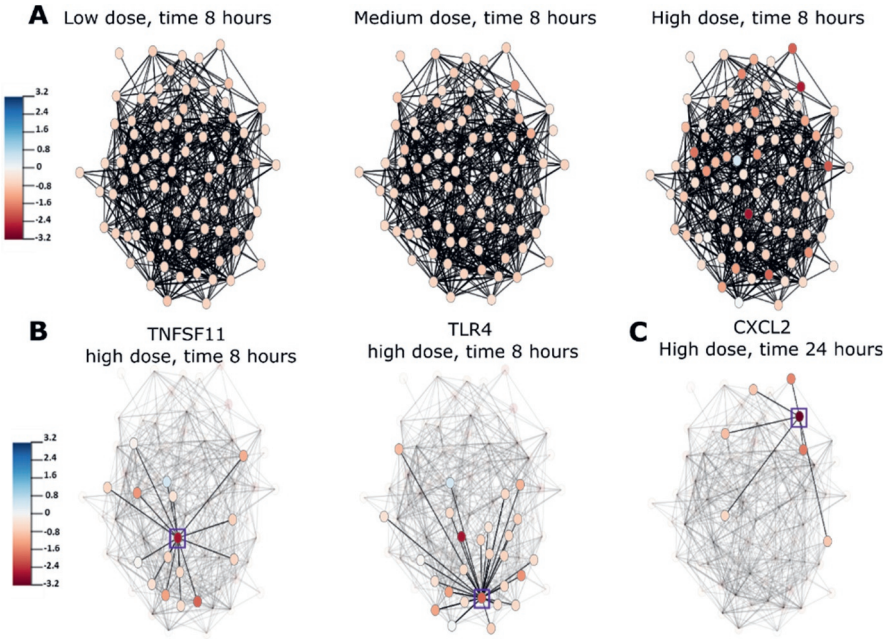


Figure 2.2 Frame-by-frame view of the expression profile in the NF-kB pathway at 8 hours for low dose, medium dose and high dose of acetaminophen (Blue color: upregulation, red color: downregulation). After 8 hours of exposure at high dose, a change in gene expression has been observed (**Panel A**, low, medium and high dose compared). TNFSF11 and TLR4 show downregulation at 8 hours (**Panel B**, node in purple square), whereas CXCL2 shows a strong downregulation at 24 hours at high dose (**Panel C**, node in purple square).

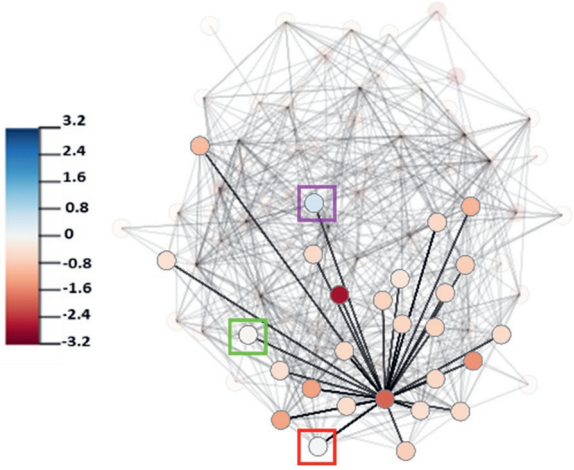


Figure 2.3 The highlighted selection of the first degree neighbors of TLR4 (central node in the sub-selection) at time point of 8 hour and high dose. CCL4 (green square), GADD45B (purple square), TNFSF13B (red square) show upregulation.

Table 2.1 Overview of the different features present in DynOVis, Cytoscape and the different Cytoscape applications.

Tools	DynOVis	Cytoscape	DyNet	CyAnimator	DyNetViewer
Platform	web	Standalone Java application	Cytoscape plugin	Cytoscape plugin	Cytoscape plugin
Input Network	List of interactions, Adjacency matrix, .sif file	List of interactions, Adjacency matrix, .sif file, URL or database	Cytoscape input types + 2 equal networks	Cytoscape input	Cytoscape input
Input data	Text file, CSV file, excel workbook	Text file, CSV file, Excel workbook, URL or database	Text file, CSV file, Excel workbook, URL or database	Text file, CSV file, Excel workbook, URL or database	Text file
Built-in database	Yes	No	No	No	No
Network layout	Force directed layout	Multiple different layouts	Cytoscape Layouts	Cytoscape Layouts	Cytoscape Layouts
Weighted edge visualization	No	Yes	Yes	Yes (from Cytoscape Core function)	No
Dynamic node expression	Yes	No	Yes	Yes	No
Dynamic node interactions	Yes	No	Yes	No	Yes
3D visualization	Yes	No	No	No	No
2D Line graph expression values	Yes	No	No	No	No
Pathway histogram	Yes	No	No	No	No

TLR4 plays an important role in pathogen recognition and can activate the innate immune system [22]. It has been found in previous research that downregulation of *TLR4* is related to liver cirrhosis [23]. Highlighting the neighbors of *TLR4* using DynOVis (Figure 2.3) demonstrates that a number of genes interacting with *TLR4*, show the same expression patterns and thus may play the same role in the response. *TNFSF11* has an interaction with *TLR4* (downregulated at 8 hours and high dose, dark red node in Figure 2.3) and is involved in the regulation of the T cell-dependent immune response.

While TLR4 shows a downregulation, three genes show an upregulation: C-C motif chemokine ligand 4 (CCL4), growth arrest and DNA damage inducible beta (GADD45B) and tumor necrosis factor ligand superfamily member 13b (TNFSF13B) (light blue nodes in figure 2.3, CCL4 green square, GADD45B purple square, TNFSF13B red square). GADD45B belongs to the GADD nuclear protein family that is associated with DNA damage [24]. These results can also be saved in a 2D line graph, to show the changes in gene expression in a static way (Figure 2.4). CXCL2 is one of the chemokines that leads to an influx of inflammatory cells including macrophages which are known for wound healing [25]. Downregulation of CXCL2 is observed at 24 hours and therefore could play an important role in the toxic effect of acetaminophen on the liver. In this case study, DynOVis did help to highlight the most important genes in a high-density network as well as to identify important gene-gene interaction effects.

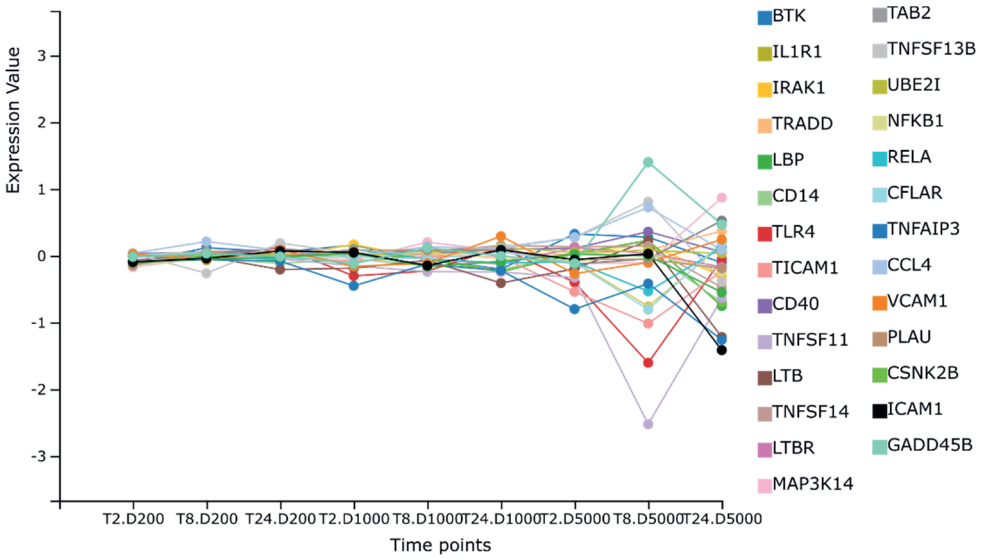


Figure 2.4 2D line graph of expression profiles of TLR4 and the first degree selected neighbors at different time and dose points. Time is abbreviated by T followed by 2, 8 or 24 hours. Dose is abbreviated by D followed by 200 μM (low dose), 1000 μM (medium dose) and 5000 μM (high dose). Data is derived from human in vitro samples exposed to acetaminophen.

In the second example, DynOVis is applied to a network with changing interactions over time as a response to valproic acid (VPA) exposure [26]. In this study, primary human hepatocytes are exposed to VPA for a period of three days with a follow-up washout (WO) period of three days. A selection of genes has been made by Wolters et al by applying the induced network module of ConsensusPathDB [12] to create multiple gene-gene interaction networks to unravel the dynamic changes over time after VPA exposure. From those results, only the high-confidence protein interactions and gene regulatory interactions are exported. We have used the edge

interaction function of DynOVis to create an animation of the dynamic interaction changes over time that highlights important network structure changes (Figure 2.5).

From figure 2.5, it becomes clear that the number of interactions grows over time during VPA exposure, with the highest number of interactions on day 3, which is in line with the observations of Wolters et al [26]. The increase of edge is data-driven, due to higher number of differentially expressed genes at each time point [26]. More differentially expressed genes means more possible interactions and thus more nodes and edges in the network. A number of persistent gene-gene interactions are identified after the WO period. To fully understand the relationship between perturbations in gene expression and gene-gene interactions, DynOVis is used to create an animation that combines node expression with edge interactions.

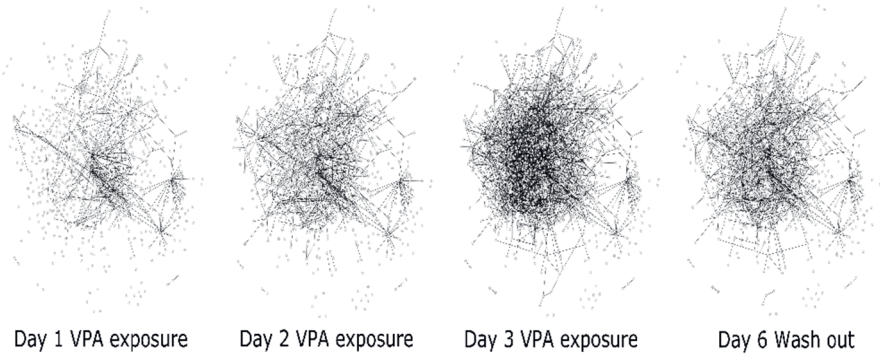


Figure 2.5 Dynamic molecular interaction changes over time after VPA exposure after 1, 2 and 3 days, as well as after the 3 day wash out period.

Here we will discuss only one gene of interest: Fibronectin 1 (*FN1*) (Figure 2.6, network of *FN1*). *FN1* is initially connected to 10 neighbors (outer degree: 3, inner degree: 7), but develops to 71 interactions (outer degree: 22, inner degree: 49) with other genes after a 3-day treatment with VPA and maintains in total 50 of the 71 interactions (outer degree: 15, inner degree: 35) after the 3 day washout period. The changes in gene-gene interactions between *FN1* and its neighbors, as well as the changes in gene expression, are of great interest for various reasons. Knock-down of *FN1* has been shown to play a role in mitochondrial-dependent apoptosis [27] as well as increased fibrosis in mice [28]. Accumulating cancer research showed that fibronectin expression in various tumors is highly correlated with malignant phenotypes and poor prognosis [29–31].

By identifying the dynamic interactions over time, we found both persistent and non-persistent interactions. The non-persistent interactions between *FN1* with amyloid P component serum (*APCS*), and filamin A (*FLNA*) could be associated with VPA exposure since the interactions disappear after the WO period. The expression pattern of *APCS* and *FLNA* are highly similar to *FN1*, which could indicate that the

downregulation of those three genes plays a role in hepatotoxicity. Several interactions disappear over a 2 day VPA exposure, including the interactions between *FN1* and inter-alpha-trypsin inhibitor heavy chain 2 (*ITIH2*) and proliferating cell nuclear antigen (*PCNA*). These different dynamic gene-gene interactions are of interest because they may play a role in the same biological processes that lead to the adverse outcome of VPA exposure [32]. PCNA is a nuclear protein involved in DNA-synthesis and repair and decreased expression of PCNA has been experimentally shown in tumors of VPA treated mice [33]. ITIH2 belongs to the family of plasma serine protease inhibitors involved in the stabilization and prevention of tumor metastasis. Further investigation of the relationship between these genes is needed but is behind the scope of this case study. In the second case study, DynOVis gives us a quick overview of the growing gene-gene interactions over VPA exposure, which helps to identify important time points in the exposure series and directs us towards the most important genes to study.

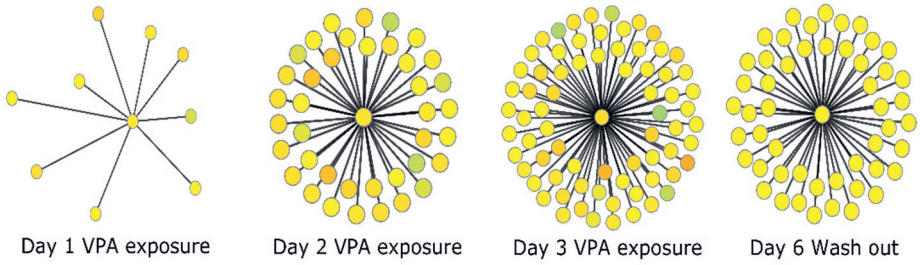


Figure 2.6 Subnetwork for Fibronectin 1 at the different days of VPA exposure. There appears a strong time effect of VPA exposure, due to the increasing interactions over time. The center node of each small subnetwork represents Fibronectin 1. At day 1 of exposure, there are 9 genes that have a gene-gene interaction with Fibronectin 1, this number increases over the following two days to 72 gene-gene interactions. The number of interactions decreases after the washout period.

Conclusion

With DynOVis we offer the implementation of dynamic network visualization, by providing the users with functionalities to highlight node expression changes and dynamic edges. The addition of biological information, such as pathway or disease association, helps to further understand the role of different nodes in the network. It is possible to study dynamic effects without the need for multiple networks while the tool immediately provides the user with information about gene function, associated pathways, and diseases. Although Cytoscape is more powerful in analyzing a static network, DynOVis has some advantages over the current Cytoscape applications regarding dynamic visualization. DynOVis allows studying both dynamic node expression changes and edge interaction changes simultaneously, whereas the current Cytoscape tools focus more on one topic. With the provided case studies, we have shown that with the dynamic network visualization it becomes less complicated to identify important causal events. Further development of the tool will be carried out, in order to enable the integration

of multiple omics platforms that will provide an even more detailed explanation of the cellular response to perturbations. These updates will be integrated into the tool and updated regularly

Availability

Project name: DynOVis
Project home page: https://tjmkuijpers.shinyapps.io/dynovistool/ and
https://bitbucket.org/mutgx/dynovis/src/master/ for source code
Operating system(s): Operating system independent (web service)
Programming Language: R and JavaScript
Other requirements: The DynOVis interface uses HTML5 features that are not supported by the current version of Internet Explorer (IE v11); i.e. Internet Explorer v11 does not fully support DynOVis. Therefore, we recommend using DynOVis on a different browser (Google Chrome or Mozilla Firefox).
License: BSD license
Authors' contributions: TK designed and wrote the code for DynOVis. DJ and JK supervised the design of the tool. JW performed the analysis and interpretation of the VPA case study.TK wrote the original draft, JW, DJ and JK reviewed and edited the original draft. All authors read and approved the final manuscript.

References

1. Bar-joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nat Publ Gr. 2012;13:552–64. doi:10.1038/nrg3244.
2. Heijne WH, Kienhuis AS, van Ommen B, Stierum RH, Groten JP. Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. Expert Rev Proteomics. 2005;2:767–80. doi:10.1586/14789450.2.5.767.
3. Novère N Le. Quantitative and logic modelling of molecular and gene networks. Nat Publ Gr. 2015;16:146–58. doi:10.1038/nrg3885.
4. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol cell Biol. 2008;9:770.
5. Nachman I, Regev A, Friedman N. Inferring quantitative models of regulatory networks from expression data. Bioinformatics. 2004;20 Suppl 1:i248–56. doi:10.1093/bioinformatics/bth941.
6. Bansal M, Gatta GD, di Bernardo D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics. 2006;22:815–22. doi:10.1093/bioinformatics/btl003.
7. Kim SY. Inferring gene networks from time series microarray data using dynamic Bayesian networks. Brief Bioinform. 2003;4:228–35. doi:10.1093/bib/4.3.228.
8. Hendrickx DM, Souza T, Jennen DGJ, Kleinjans JCS. DTNI: a novel toxicogenomics data analysis tool for identifying the molecular mechanisms underlying the adverse effects of toxic compounds. Arch Toxicol. 2017;91:2343–52. doi:10.1007/s00204-016-1922-5.
9. Theocharidis A, van Dongen S, Enright AJ, Freeman TC. Network visualization and analysis of gene expression data using BioLayout Express3D. Nat Protoc. 2009;4:1535–50. doi:10.1038/nprot.2009.177.
10. Kincaid R, Kuchinsky A, Creech M. VistaClara: an expression browser plug-in for Cytoscape. Bioinformatics. 2008;24:2112–4. doi:10.1093/bioinformatics/btn368.
11. Morris JH, Vijay D, Federowicz S, Pico AR, Ferrin TE. CyAnimator: Simple Animations of Cytoscape Networks. F1000Research. 2015. doi:10.12688/f1000research.6852.2.
12. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013;41:D793–800. doi:10.1093/nar/gks1055.
13. CTD. Curated chemical-disease data were retrieved from the Comparative Toxicogenomics Database (CTD), North Carolina State University, Raleigh, NC and Mount Desert Island Biological Laboratory, Salisbury Cove, Maine. World Wide Web. 31-08-2017. 2017. http://ctdbase.org/.
14. Maglott D. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2004;33 Database issue:D54–8. doi:10.1093/nar/gki031.
15. Zotenko E, Mestre J, O'Leary DP, Przytycka TM. Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Comput Biol. 2008;4:e1000140. doi:10.1371/journal.pcbi.1000140.

16. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008;6:e264. doi:10.1371/journal.pbio.0060264.
17. Goenawan IH, Bryan K, Lynn DJ. DyNet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics. 2016;32:2713–5. doi:10.1093/bioinformatics/btw187.
18. Li M, Yang J, Wu F-X, Pan Y, Wang J. DyNetViewer: a Cytoscape app for dynamic network construction, analysis and visualization. Bioinformatics. 2018;34:1597–9. doi:10.1093/bioinformatics/btx821.
19. Schivo S, Scholma J, van der Vet PE, Karperien M, Post JN, van de Pol J, et al. Modelling with ANIMO: between fuzzy logic and differential equations. BMC Syst Biol. 2016;10:56. doi:10.1186/s12918-016-0286-z.
20. Vinken M. Adverse Outcome Pathways and Drug-Induced Liver Injury Testing. Chem Res Toxicol. 2015;28:1391–7. doi:10.1021/acs.chemrestox.5b00208.
21. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. 2015;43:D921–7. doi:10.1093/nar/gku955.
22. Broering R, Lu M, Schlaak JF. Role of Toll-like receptors in liver health and disease. Clin Sci. 2011;121:415–26. doi:10.1042/CS20110065.
23. Manigold T, Böcker U, Hanck C, Gundt J, Traber P, Antoni C, et al. Differential expression of toll-like receptors 2 and 4 in patients with liver cirrhosis. Eur J Gastroenterol Hepatol. 2003;15:275–82. doi:10.1097/01.meg.0000050010.68425.cb.
24. Zhang N, Ahsan MH, Zhu L, Sambucetti LC, Purchio AF, West DB. NF-κB and Not the MAPK Signaling Pathway Regulates GADD45β Expression during Acute Inflammation. J Biol Chem. 2005;280:21400–8. doi:10.1074/jbc.M411952200.
25. Luedde T, Schwabe RF. NF-κB in the liver—linking injury, fibrosis and hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol. 2011;8:108–18. doi:10.1038/nrgastro.2010.213.
26. Wolters JEJ, van Breda SGJ, Grossmann J, Fortes C, Caiment F, Kleinjans JCS. Integrated 'omics analysis reveals new drug-induced mitochondrial perturbations in human hepatocytes. Toxicol Lett. 2018;289:1–13. doi:10.1016/j.toxlet.2018.02.026.
27. Wu D. Knockdown of Fibronectin Induces Mitochondria-Dependent Apoptosis in Rat Mesangial Cells. J Am Soc Nephrol. 2005;16:646–57. doi:10.1681/ASN.2004060445.
28. Kawelke N, Vasel M, Sens C, von Au A, Dooley S, Nakchbandi IA. Fibronectin Protects from Excessive Liver Fibrosis by Modulating the Availability of and Responsiveness of Stellate Cells to Active TGF-β. PLoS One. 2011;6:e28181. doi:10.1371/journal.pone.0028181.
29. Cheng H-C, Abdel-Ghany M, Elble RC, Pauli BU. Lung Endothelial Dipeptidyl Peptidase IV Promotes Adhesion and Metastasis of Rat Breast Cancer Cells via Tumor Cell Surface-associated Fibronectin. J Biol Chem. 1998;273:24207–15. doi:10.1074/jbc.273.37.24207.
30. Huang L, Cheng H-C, Isom R, Chen C-S, Levine RA, Pauli BU. Protein Kinase Cϵ Mediates Polymeric Fibronectin Assembly on the Surface of Blood-borne Rat Breast Cancer Cells to Promote Pulmonary Metastasis. J Biol Chem. 2008;283:7616–27. doi:10.1074/jbc.M705839200.
31. Cheng H-C, Abdel-Ghany M, Pauli BU. A Novel Consensus Motif in Fibronectin Mediates Dipeptidyl Peptidase IV Adhesion and Metastasis. J Biol Chem. 2003;278:24600–7. doi:10.1074/jbc.M303424200.
32. Gilbert-Diamond D, Moore JH. Analysis of Gene-Gene Interactions. In: Current Protocols in Human Genetics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2011. doi:10.1002/0471142905.hg0114s70.
33. Sang Z, Sun Y, Ruan H, Cheng Y, Ding X, Yu Y. Anticancer effects of valproic acid on oral squamous cell carcinoma via SUMOylation in vivo and in vitro. Exp Ther Med. 2016;12:3979–87. doi:10.3892/etm.2016.3907.

Chapter 2.2: GINBuilder: a python frame to build Genomic Interaction Networks to study –gene-gene and DNA methylation – gene interactions

T.J.M. Kuijpers[1], J.C.S. Kleinjans[1] and D.G.J. Jennen[1]

[1] Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

Abstract

To gain a better understanding of the relationships between features in large omics data sets, we propose *GINBuilder*. *GINBuilder* is a python framework designed to construct a transcriptome – epigenome interaction network with the option to integrate curated knowledge on disease and compound interactions. Integration of knowledge databases with omics patterns into genomic interaction networks can help us to deeper understand different phenotypic endpoints, such as disease/control or exposure studies. *GINBuilder* provides the user with associations that cannot only help to improve our understanding of relationships within omics layers but also highlights certain non-omics association. The constructed genomic interaction networks will be a valuable tool to unravel the relation between genomic alterations and phenotypic endpoints.

Introduction

The generation and analysis of large omics data sets can help us unravel different biological processes involved in understanding toxicity or disease development. To extract potential important information from omics data, feature selection techniques are widely applied to identify latent features in the data [3–5]. To gain a better understanding of the different inter- and intra-omics relationships, we have to highlight the biological interactions between those features. Biological networks are a great tool to study the relationships between DNA methylated regions and genes. To translate the obtained features into a network, we have to define the interactions between those entities. These interactions will help us to understand the relationships between the biological components and can help us further unravel an observed phenotype.

We can add different types of interaction to a biological network, depending on the hypothesis we want to investigate. For example, if we would like to study the biological mechanisms in a cancer cell, we can add protein-protein interactions, gene-gene interactions as well as DNA methylated regions linked to their corresponding gene. This will provide us with information on physical connections between gene products (protein-protein), relationships between genes (gene-gene interactions) and can reveal key function modules. However, adding all these interactions will create a complex network, with a high number of interactions (edges) between each biological entity (node). This makes it very difficult to validate our hypothesis. We could apply a classical network biology method to extract nodes with a high degree, or network modules with members that share a high number of connections. However, we will look into a different approach, in which we integrate existing knowledge from databases, to help us identify genes related to a certain disease phenotype or chemical compound exposure. We hypothesize that integration of knowledge databases with omics patterns into genomic interaction networks can help us to understand different phenotypic endpoints, such as disease/control or exposure studies.

Therefore, to study the relationship between epigenetics and transcriptomics, we propose GINBuilder, a python framework to construct genomic interaction networks, which store the interactions between genes and DNA methylation – gene relationships.

Application

GINBuilder is a python framework designed to construct a transcriptome – epigenome interaction network that the user can expand with disease or compound interactions (Figure 2.7). GINBuilder uses a list of genes and (hypo/hyper-) methylated genes as input to define the seeding nodes of the network. In this list, methylated genes extracted from the epigenome do not necessarily need to overlap with the list of genes extracted from the transcriptome. The rationale behind this choice is that in some cases, there are changes on one layer that are not

necessarily detected on the other layer, due to limitations of the omics integration techniques.

We map every seeding node against various databases to retrieve potential interesting interactions. There are different steps to construct the genomic interaction network and multiple possibilities to add information onto the genomic interaction network. Here, we will discuss the options to translate a list of genes and methylated DNA regions into a complete genomic interaction network.

To start building the genomic interaction network, the user launches the Jupyter notebook, which guides the user through the different steps. A python object for the Genomic interaction network will be created by calling the function "*CreateGenomicInteractionNetwork()*". This is an empty network, but we populate the network with gene nodes (via "*set_genes_as_nodes()*") and CpG nodes (via "*get_cpg_gene_interactions()*").

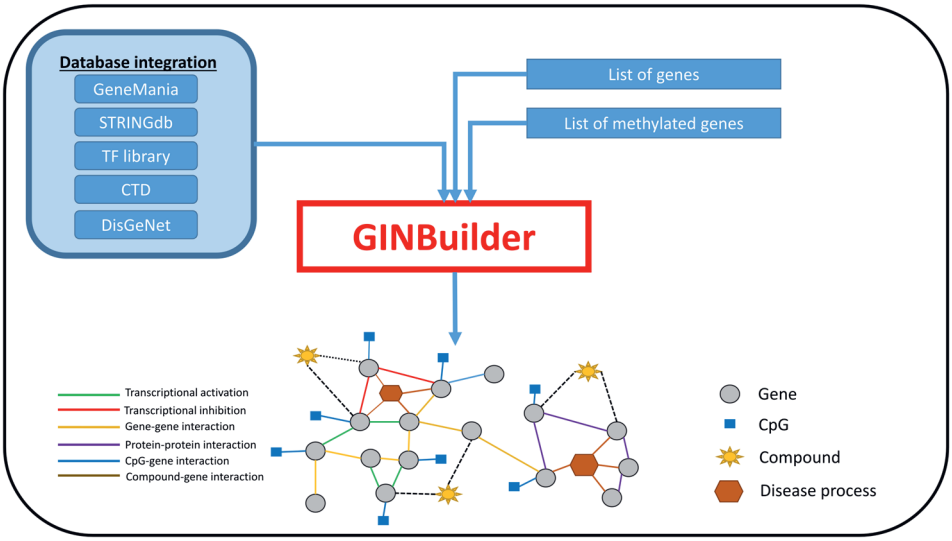


Figure 2.7 Genomic Interaction Network (GIN) Builder: a graphical representation of the connection between feature genes and methylated genes and knowledge integration.

Now we can add interactions between the nodes and start to fill the genomic interaction network with edges. For each of the CpG nodes, GINBuilder uses an internal database to search for the gene belonging to the CpG node. If the mapping returns a gene name, GINBuilder will define an interaction between the CpG node and the gene, defined by "CpG – Gene interaction".
After defining the CpG – gene interactions, the next step is to extract known interactions from knowledge databases. We have added an API functionality to retrieve information from various databases and since each of them stores specific information, we will briefly explain their use:

Transcription factor library
Transcription factors tightly regulate the expression of their target genes via transcriptional activation or inhibition. Alterations of transcription factors are important factors leading to disease onset [6, 7] and therefore transcription factor – gene target interactions are vital to study. Each gene, defined as a starting node, is mapped against the in-house developed transcription database by Souza et al [8]. We use the function "get_transcription_factor_regulation()" to get the transcription factor – target interactions, which are returned in the format TF – target – activation/inhibition. This function identifies all the transcription factors in the network and their transcriptional target. By default, we only include a transcription factor and its target, if both are a node in our network. If the user is only interested in some transcription factors, it is possible to search for those transcription factors by defining a custom list of genes and pass this list as an argument to the function.

OmniPath
Disruption of cell signaling cascades is believed to be an important onset in diseases such as cancer [9]. OmniPath [10] stores signaling interactions, kinase-substrate interactions, transcriptional factor – target interactions as well as miRNA – mRNA interactions. GINBuilder uses the function "get_gene_gene_interactions()" to retrieve all the interactions based on the nodes in the network so no input parameters have to be defined.

STRINGdb
The STRING database (STRINGdb) [11] is a collection of known and predicted protein-protein interactions (PPIs) that can be either direct (physical) interactions or indirect (functional) interactions. GINBuilder users the function "get_protein_protein_interactions()" to retrieve the interactions between the nodes in the network. STRINGdb uses different sources to construct PPIs including databases, experiments, and text mining. Therefore, GINBuilder sets the confidence score to 0.7 (medium confidence) to remove all low confident PPIs. This improves the quality of the knowledge added to the network since low confident PPIs can be misleading.

GeneMANIA
GeneMANIA [12] is an interactive database that stores gene-gene interactions based on co-expression, co-localization, pathway, predicted interactions, physical interactions, and shared protein domains. The user can retrieve all interactions through the function "get_gene_interactions()" but by default GINBuilder only searches for the physical interactions and shared protein domains. We believe that interactions such as the co-expressed interactions are more sensitive to experimental conditions and therefore it is better to derive them from your original data set.

Comparative toxicogenomics database
The comparative toxicogenomics database (CTD) [13] stores literature-based manually curated associations between chemicals, genes, phenotypes, and diseases. In exposure studies, it is not only of interest to identify new chemical – gene interactions, but also to better understand known chemical – gene interactions by studying downstream effects on gene perturbations. To add compound – gene interactions, we call the function "get_compound_gene_interactions()". This function uses a number of parameters that have to be specified by the user: list of genes to use in the search (genes_to_subset), input type of the compound (input_type_compound), list of compounds to use in search (input_terms_compound), and if curated, inferred or all associations should be reported (report_only_parameters).

DisGeNET
DisGeNET is a knowledge platform that stores associations between genes and human diseases. It stores 17 549 genes and 24 166 diseases, making up for a total of 626 685 gene-disease associations [14]. Here, the term 'disease' is an umbrella term for not only actual diseases but also disease symptoms and abnormal phenotypes relevant in human genomics. Each association holds a score: curated, animal model, literature, and inferred. By default, GINBuilder sets the association score to curated interactions to increase the confidence of the retrieved interactions. To search the database for a disease – gene interaction, we use the function "find_disease_associated_with_genes()". This function takes all the genes in our genomic interaction network and adds a disease-gene interaction if present in the database. However, it is also possible to specify a disease of interest by using "find_genes_associated_with_disease(disease)" and use your disease of interest as a parameter.

Conclusion
GINBuilder provides a set of easy-to-use functionalities to connect REST APIs and local databases to build a genomic interaction network. Biological networks play an important role in the understanding of complex biological systems and the combination of a data-to-knowledge integration with a network is important to decrease this complexity. We believe GINBuilder is an addition to the current network analysis methods and can increase our biological understanding of such networks.

Availability and implementation
GINBuilder is freely available at https://github.com/TJMKuijpers/GINBuilder to download the source code and Jupyter notebook. This Jupyter notebook guides the user to build a genomic interaction network and can be used as a template for future work.
License: MIT

References

1. Panagiotou G, Taboureau O. The impact of network biology in pharmacology and toxicology. SAR QSAR Environ Res. 2012;23:221–35. doi:10.1080/1062936X.2012.657237.
2. Piñero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. Sci Rep. 2016;6:24570. doi:10.1038/srep24570.
3. Canzler S, Schor J, Busch W, Schubert K, Rolle-Kampczyk UE, Seitz H, et al. Prospects and challenges of multi-omics data integration in toxicology. Arch Toxicol. 2020;94:371–88. doi:10.1007/s00204-020-02656-y.
4. Conrad T, Kniemeyer O, Henkel SG, Krüger T, Mattern DJ, Valiante V, et al. Module-detection approaches for the integration of multilevel omics data highlight the comprehensive response of Aspergillus fumigatus to caspofungin. BMC Syst Biol. 2018;12:88. doi:10.1186/s12918-018-0620-8.
5. Lee B, Zhang S, Poleksic A, Xie L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. Front Genet. 2020;10. doi:10.3389/fgene.2019.01381.
6. Golson ML, Kaestner KH. Fox transcription factors: from development to disease. Development. 2016;143:4558–70. doi:10.1242/dev.112672.
7. Kitamura Y, Shimohama S, Ota T, Matsuoka Y, Nomura Y, Taniguchi T. Alteration of transcription factors NF-κB and STAT1 in Alzheimer's disease brains. Neurosci Lett. 1997;237:17–20. doi:10.1016/S0304-3940(97)00797-0.
8. Souza TM, Rieswijk L, Beucken T van den, Kleinjans J, Jennen D. Persistent transcriptional responses show the involvement of feed-forward control in a repeated dose toxicity study. Toxicology. 2017;375:58–63. doi:10.1016/j.tox.2016.10.009.
9. Sever R, Brugge JS. Signal Transduction in Cancer. Cold Spring Harb Perspect Med. 2015;5:a006098–a006098. doi:10.1101/cshperspect.a006098.
10. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13:966–7. doi:10.1038/nmeth.4077.
11. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13. doi:10.1093/nar/gky1131.
12. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38 suppl_2:W214–20. doi:10.1093/nar/gkq537.
13. CTD. Curated chemical-disease data were retrieved from the Comparative Toxicogenomics Database (CTD), North Carolina State University, Raleigh, NC and Mount Desert Island Biological Laboratory, Salisbury Cove, Maine. World Wide Web. 31-08-2017. 2017. http://ctdbase.org/.
14. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2019. doi:10.1093/nar/gkz1021.

Chapter 3

Transcriptome and epigenome integration of *in vitro* cancer data

T.J.M. Kuijpers[1,]
J.C.S. Kleinjans[1]
D.G.J. Jennen[1]

1 Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

Chapter 3.1: Integrating omics layers through multi-layer nonnegative matrix factorization with gene – CpG methylation interaction networks to identify genomic cancer profiles

Abstract

Background: To unravel the complexity of cancer biology, it has become clear that we have to integrate the different layers of molecular information. Here, we propose a multi-layer Nonnegative Matrix Factorization method based on the Kullback-Leibler divergence to create biosignatures, which can be used to create genomic interaction networks to study interactions between biological entities. First, the algorithm will be tested against a simulated data set to validate performance. Second, the NCI60 cancer cell line data set will be used to test the algorithm against a real multi-omics biological data set.

Results: The results of the simulated data set showed that the algorithm can detect predesigned groups in conditions with and without noise. In the NCI60 data set, the algorithm proposed four clusters to classify the cancer cell lines. Here, we study the melanoma cluster in more depth, to understand specific cellular changes. In the genomic interaction network, important genes are identified based on their gene expression, CpG islands methylation status, or position in the network. Different genes are found to play a pivotal role in melanoma, including MITF, IRF4, and OCA2, but also hypermethylation of HES family genes, including HES5, which might be a potentially interesting target in melanoma.

Conclusion: The creation of genomic networks through multi-layer Nonnegative Matrix Factorization adds to the current package of tools for analyzing multi-omics data sets. These genomic profiles and interaction networks can be further used to investigate characteristics of the different cancer cell lines or could be of value for potential biomarker selection.

Supplementary data available at: https://github.com/TJMKuijpers/PhDThesis

Introduction

Cancer is one of the leading causes of death in the world, with 1.93 million deaths in Europe [1] in 2018 and an estimated of 606.880 cancer deaths in the United States in 2019 [2]. Unfortunately, due to the heterogenic nature of cancer, it is a major challenge to design effective treatment [3, 4]. During the last decade, it has become known that cancer is as well a genetic as an epigenetic disease [5, 6]. Research shows that different cancer types have integrated patterns of gene expression, DNA methylation, but also point mutations and copy number changes [6, 7]. Moreover, DNA methylation is not only associated with gene repression but also with gene activation, splicing regulation, and the recruitment of transcription factors [8] and therefore an important regulator of gene expression. To unravel the complexity of cancer biology, it has become clear that we have to integrate different layers of molecular information instead of looking at one layer of information.

The ongoing development of high-throughput technologies enables the generation of large and complex multi-omics data sets, such as mRNA and microRNA expression, but also DNA methylation and protein expression. More importantly, these data sets create the opportunity for an integrated analysis approach of different omics layers. At the same time, the complexity of the multi-omics data raises challenges concerning the initial data analysis [9–11]. First, it is well known that biological systems are not homogeneous and therefore the different omics layers are heterogeneous. Different layers contain different information that follows different distributions. For instance, whereas microarray-measured mRNA is expressed with a log fold change, CpG methylation is expressed in M or β values. Thus, an appropriate mathematical method has to be developed that takes into account the underlying heterogenic nature of these data sets. Second, as most data sets consist of a large number of biological variables and a relatively low number of biological samples [10] the method should deal with this unbalanced ratio of variables versus samples. Third, the designed method should handle unsupervised learning, to explore and find hidden patterns in data sets with so far unknown relationships.

In the present study, Nonnegative Matrix Factorization (NMF) has been advanced to perform unsupervised clustering and to identify latent features that explain these clusters. Therefore, the original NMF [12] approach will be adapted for multi-layer omics. NMF is a powerful tool for data reduction and exploration, which has been widely used to extract relevant biological information [13–15]. It has been frequently applied to identify biomarkers or molecular patterns [13, 16] and recently adapted to enable cross omics data analysis [15, 17, 18]. These cross-omics adaptations either use a joint integration or NMF algorithms based on the Euclidean distance. Here, we propose an integrated NMF method based on the Kullback-Leibler divergence, which is believed to outperform the Euclidean NMF approach [19]. Moreover, we aim for an intermediate NMF integration, which is designed to find a

solution across multiple omics layers simultaneously. We hypothesize that by updating the multi-layer NMF method with the multi-layer Kullback-Leibler update rules, we can identify different feature profiles, to explain the molecular differences between clusters of samples. These feature profiles will then be used to build gene – CpG island interaction networks, which may help to further understand the different relationships between the omics layers.

Method

From single layer NMF to multi-layer NMF

The single-layer NMF algorithms perform the dimension reduction and grouping of samples based on only one layer of information (X in equation 1) to obtain W, containing the latent features and H, holding the coefficients. Here, layer X could be gene expression data, W would hold the genes that explain the clusters that can be identified from H. To find the local optimal solution for the problem defined in equation 1, a cost function (Equation 2) has to be minimized. Here, the cost function defined by the Kullback-Leibler (KL) divergence is set (Equation 2), which is a measure for the divergence between two matrices. This divergence between the two matrices should be minimal to find the most optimal solution to equation 1.

$$X \approx WH \quad (1)$$

$$KL\ divergence = X * \sum \left(\log \left(\frac{X}{WH} \right) - X + WH \right) \quad (2)$$

We propose an integrated NMF strategy that aims to identify H based on the information in the different omics layers X_i but also based on the features from the layers W_i. The underlying hypothesis is that the solution for H is depending on the combination of features W_i that explain the clusters formed from the different data sources. Therefore, equation (1) is updated to a new equation (3), where H depends on all n data layers. For each data layer, the original data matrix X_i is estimated by the product HW_i (Equation 3). To find a local optimal solution, matrices W_i and H are updated by their update rules (Equations 4 and 5 respectively). For H we consider the effect of the different omics layers via X_i and W_i whereas for W_i we take into the effect the omics layers via X_i and the sample clustering via H. In the end, n matrices W are obtained that store the latent features and one coefficient matrix H that stores the clustering coefficients.

$$\sum_{i=1}^{n} X_i \approx \sum_{i=1}^{n} W_i H \quad (3)$$

$$W_{w+1} = W * \frac{\frac{X_i}{W_i H} H^T}{\sum H} \quad (4)$$

$$H_{H+1} = H * \frac{\sum \frac{X_i}{W_{iH}}^T}{\sum W_i^T} \quad (5)$$

$$KL\ divergence = X_i * \sum \left(\log \left(\frac{X_i}{W_i H} - X_i + W_i H \right) \right) \quad (6)$$

Choosing the right dimension k

An important characteristic of NMF is that it reduces the dimensionality of the original data matrix to a much smaller matrix with dimension k. This dimension k relates to the number of metagenes/encoding coefficients found in the data and it is, therefore, nontrivial for choosing the optimal value for k. For each value of k, a connectivity matrix C (dimension MxM) is calculated based on the sample assigned to each cluster. If two samples i and j belong to the same clusters, then the entry C_{ij} and C_{ji} will be 1, otherwise, it is 0. The measure of dispersion [13, 20] is then calculated for each connectivity matrix by:

$$\rho = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4 * \left(C_{ij} - \frac{1}{2} \right)^2 \quad (7)$$

Here, a value of 1 for ρ will indicate a perfect consensus matrix, whereas a value between 0 and 1 will indicate a scattered matrix. Besides the dispersion coefficient, the silhouette coefficient is calculated for each connectivity matrix. For each value of k, the silhouette score and dispersion score can be used to identify the optimal k for which the silhouette score is maximum.

Identifying features for each cluster

To analyze the difference in methylation and gene expression profile of each cluster, each matrix W_i is scored by using the method proposed by Kim et al [20]. For each cluster, the entities are selected as features, if those entities have a high score for the selected cluster and a low score for the other clusters. This will ensure that the different features for each cluster are unique to that cluster. One should note that it is also possible to choose more stringent criteria, which will reduce the number of features, but at the same time will select the more important ones.

Integrated features network to resemble biological events

The obtained feature matrices W_i obtained after solving equation (3) are used to construct multi-layer biological networks. Biological networks are important tools to study interaction changes after perturbation of the system [21], or to identify relevant interaction changes between different conditions [3, 4]. For the different features identified by NMF, within platform interactions will be added for each feature. Furthermore, interactions between entities in different platforms will be created by mapping CpG islands to the associated gene and used to build a feature-to-feature multi-platform interaction network. Network characteristics will be calculated to determine entities with a high degree of outgoing edges, representing those entities that influence a high number of neighbors. It is believed that perturbations of entities with a high degree of interactions are vital to study and help to understand the overall effect of the perturbation on the system [22, 23].

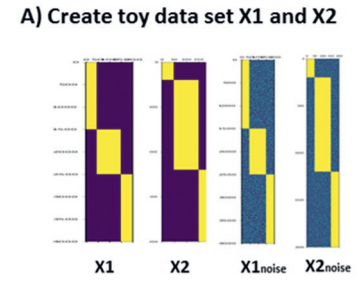
Simulated data set
To test the method in a controlled environment, two data matrices X1 and X2 have been built consisting only of ones and zeros, where three distinct profiles have been created for three different groups. Three different scenarios have been simulated: i) noise-free data, ii) data with added noise and iii) randomized data with noise. In all three cases, three group sizes have been implemented within the data sets, group 1 of 50 samples, group 2 of 100 samples, and group 3 of again 50 samples.

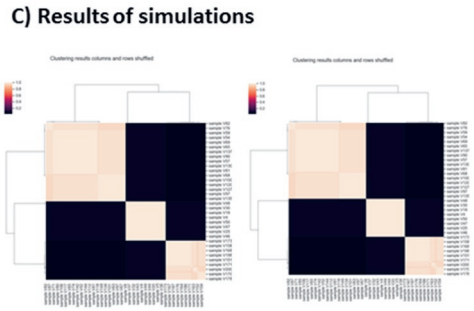
NCI60 cancer cell line data set
The NCI60 data cancer cell line data set is a large collection of 60 human cancer cell lines that are used to screen over 100,000 chemical compounds and natural products. The NCI60 data set includes tissues ranging from the prostate to the central nervous system, as well as samples collected from leukemia and melanoma. From the CellMiner interface [24], microarray mRNA expression values (Whole Human Genome Microarray 4 x 44K, log2 normalized) and CpG Island methylation (Illumina 450K platform, β value) have been downloaded. Both data sets have been filtered to eliminate the low variant genes and CpG Islands. This processing step has been applied since low variant entities do explain the differences between groups and therefore are not of interest. After this processing step, the two data sets were given as an input to the workflow to obtain clusters with their corresponding gene and methylation profile.

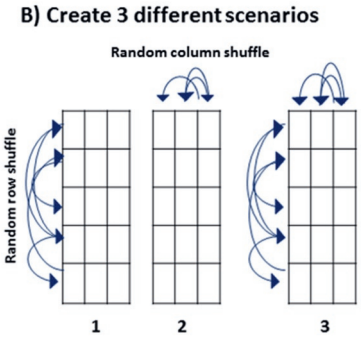
Results
The proposed integration workflow has been tested on two different data sets: one toy data set and the NCI60 data set. The toy data set will be generated with two different conditions, with and without the presence of noise. Furthermore, four different scenarios have been designed to test the performance of the method. These four scenarios include non-randomized data, column-randomized data, row-randomized data, and total randomized data.

Toy data set: noise-free and noisy conditions
A toy data set has been created to evaluate the performance of the new update rules for the multi-layer NMF approach. First, a data set without noise has been used to determine if the designed groups of samples could be retrieved (Figure 3.1A, left). The results show that for the different scenarios (Figure 3.1B), the algorithm is able to cluster all the samples into the three designed groups (Figure 3.1C). Furthermore, if only the columns or rows, or both rows and columns are shuffled, the predicted clusters still contain the samples that correspond to their predesigned group. Second, the data set has been converted to a data set with noise (Figure 3.1A, right). Therefore, random numbers between 0 and 1 have been added to random locations in the data file. The predicted clusters indeed contain the samples that belong to the same designed group. In conclusion, with the proposed update rules to create a multi-layer NMF model based on the KL

divergence, we can reconstruct the different predesigned groups with their associated features. This result validates that our approach is capable of recognizing patterns that built different groups even in the presence of noise.







Clustering results		
scenarios	noise free data set	noisy data set
no shuffle	100%	100%
1	100%	100%
2	100 %	100%
3	100 %	100 %

Figure 3.1 Results of the toy data set. **A**: Different datasets used to evaluate method, with and without noise added. **B**: Different scenarios applied, 1) rows are randomly shuffled, 2) columns are randomly shuffled and 3) all rows and columns are randomly shuffled. **C**: Consensus matrix for the dataset (X_1, X_2) on the left and the noise dataset ($X_{1,noise}$, $X_{2,noise}$) on the right. For each of the two data sets the samples in each cluster are identified and a 100% score is obtained when the predefined clusters are obtained as the final solution.

NCI60 data set

The NCI60 cancer cell line data set [25] has been used to evaluate the workflow for integrating multiple omics layers. The identified clusters with their feature transcripts and CpG islands can be used to evaluate if the features that are proposed, are known to play a role in the different clusters. To identify the number of k clusters, multiple simulations have been performed to calculate the cluster dispersion and silhouette score for k values in the range of 1 to 15. From figure 3.2A, the maximum silhouette score is found for k=4, indicating that it is highly likely that the data set contains four clusters. Furthermore, the consensus map shows four stable clusters (Figure 3.2B), and for the four different clusters, most of the silhouette scores per sample e above the average silhouette score (Figure 3.2C).

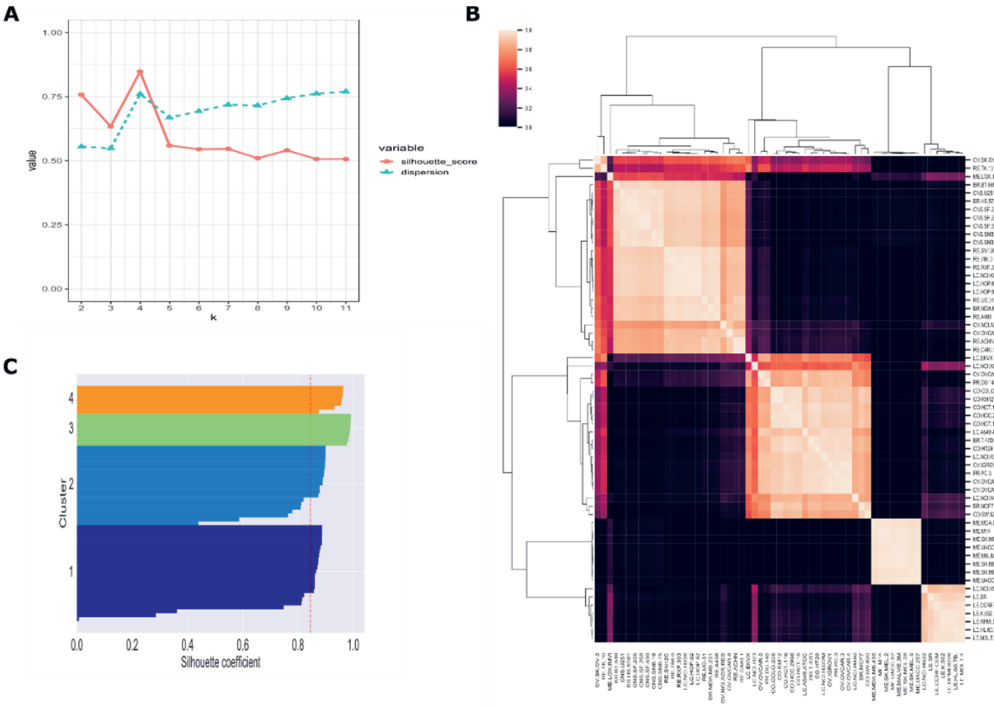


Figure 3.2 Cluster metrics to evaluate optimal k. **A**: Silhouette score and dispersion for k in range 2 to 11. **B**: Consensus map for the NMF clustering with k=4. **C**: Silhouette score per sample for k=4, red vertical line indicating the average silhouette score.

The four different clusters contain a different number of samples (Figure 3.3A), but more interesting, clusters 3 and 4 are quite homogeneous concerning their cell types (Figure 3.3B). Cluster 3 contains 8 samples of which all are melanoma samples, whereas cluster 4 contains 6 leukemia samples out of 7 cluster members. Cluster 1 and 2 both have a mixture of different cell types, where cluster 2 only contains epithelial cells.

For each of the clusters, the features can be identified after scores the feature matrices W_{gene} and W_{CpG}. After scoring the feature matrices, the different microarray probes are converted to their gene identifier, whereas the CpG probe names are converted to the CpG island identifier which can be connected to their associated gene. Here, the different feature transcripts and CpG islands for cluster 3 will be further investigated. This will give us a better understanding of the role between the transcriptome and epigenome in melanoma but also if our approach can give is new insights.

Different feature transcripts are identified from genes that play a role in biological processes associated with melanoma (Supplementary table 3.1.1), including melanin biosynthetic process (p-value: 1.75E-11, FDR: 1.38E-07), melanocyte

differentiation (p-value: 9.65E-06, FDR:8.03E-03), but also more general processes such as cell differentiation (p-value:9.20E-06, FDR: 8.09E-03) and developmental processes (p-value: 2.88E-07, FDR:4.15E-04).

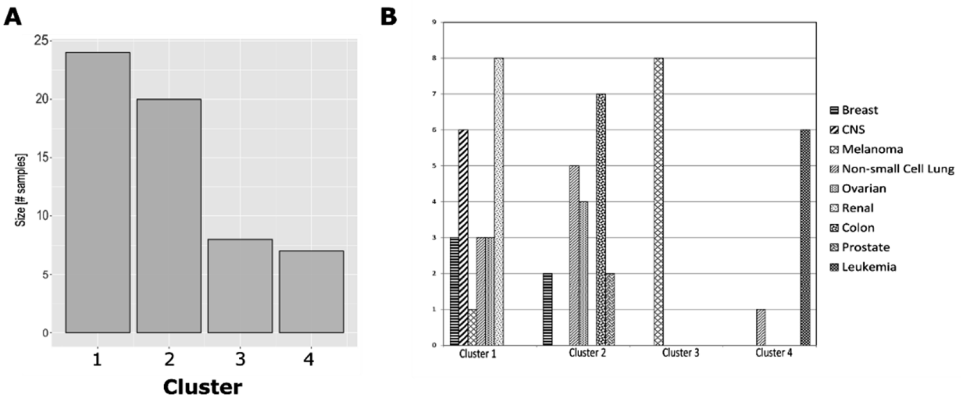


Figure 3.3 Cluster information: **A**: The size of each cluster: n=24 for cluster 1, n=20 for cluster 2, n=8 for cluster 3 and n=7 for cluster 4. **B**: The different cell types in each cluster. Cluster 1 and 2 contain a mixture of cell types, whereas cluster 3 only contains melanoma samples. A high prevalence of leukemia cell types is observed in cluster 4, along with one non-small cell lung sample.

These processes are of interest since genes involved in melanocyte development are also implicated in the development of melanoma [26], whereas genes involved in melanin biosynthesis are known to play a role in malignant melanocytes and up-regulated melanin synthesis has been observed [27]. Moreover, researchers have proposed that inhibition of melanin synthesis can improve radiotherapy of melanoma [28], which highlights the importance of melanin. Furthermore, different CpG islands features are identified that play a role in regulation of developmental processes (p-value: 8.05E-05), neurogenesis (p-value:2.07E-04), regulation of cell development (p-value:5.09E-14), negative regulation of pro-B cell differentiation (7.96E-03) (Supplementary table 3.1.2). We identified CpG islands related to HES family genes, where HES proteins regulate the expression of genes involved in Notch-dependent cell-fate determination such as apoptosis, proliferation but also differentiation [29] and are often deregulated in melanoma [30]. Our results show a general pattern of hypermethylation in the HES CpG islands and low expression levels of the HES genes in melanoma (Supplementary figure 3.1.1).

To construct a genomic interaction network, the top 10 percent feature genes and CpG island are selected and used as an input for GeneMANIA [31] to retrieve genetic interaction as well as shared protein domains between genes. Activation or inhibition reactions between transcription factors and their target are added, if present in the network, as well as CpG Island to gene connections. Here, a more stringent cutoff of 10 percent for the feature genes has been applied, to extract only the most relevant feature genes. This results in a genomic interaction network for the melanoma samples in cluster 3 (Figure 3.4A), with their associated feature genes (Figure 3.4B, top 40) and CpG islands (Figure 3.4C, top 20). From the gene

and CpG profiles, one can see that the difference is more distinguishable on the gene expression platform. Since this might indicate that the transcriptome could be enough to cluster the cancer cell lines, we will perform two single-omics NMF simulation and compare the results (supplementary table 3.1.3 for the clusters on transcriptomic data, supplementary table 3.1.4 for the clusters on methylation, comparison in discussion).

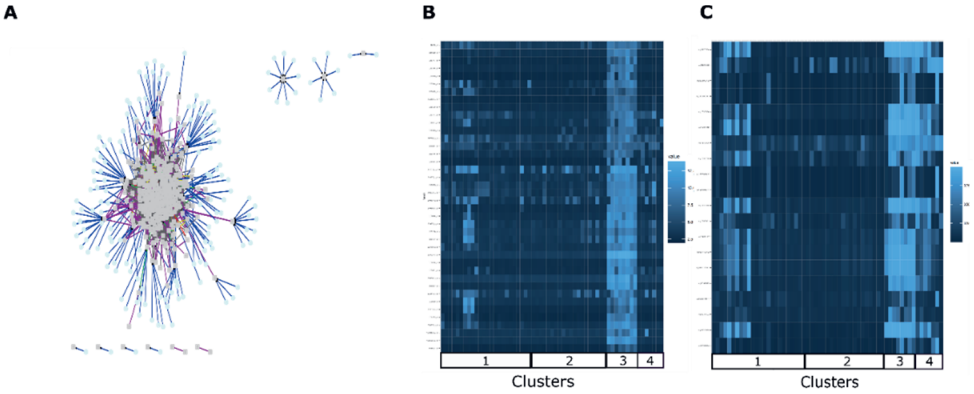


Figure 3.4 A: Genomic interaction network for melanoma cluster 3 build to study the interaction between the different features genes and CpG islands. **B:** Expression profiles of the feature genes of cluster 3 in comparison to the other clusters. **C:** Expression profile of CpG islands associated with the melanoma cell lines of cluster 3 in comparison to the other clusters.

In our genomic interaction network (Figure 3.4A), we have identified genes with a central position in the network and at the same time have high expression values. This includes *IRF4, MITF, RXRG*, and *PAX3*, which are all transcription factors and have targets associated with melanoma. Transcription factors might be one of the most interesting to study in more detail because the interaction between transcription factors and their targets is a key determinant in cellular processes and is frequently found to be changed in disease [32]. Transcription factors regulate promotor activity, often through interactions with gene regulatory regions [33].

MITF is known to drive towards a drug-tolerant environment of melanoma [34] and is linked to innate resistance [35]. Moreover, *MITF* is claimed to be a potential drug target because inhibition of *MITF* improves melanoma cells to BRAF inhibitors [36]. *PAX3* regulates important processes as cell survival, proliferation, migration, and its function is suggested to be retained in melanoma cells by Medic et al [37]. *IRF4* is a regulator of transcriptional targets involved in cell development, oncogenesis, and immune response [38].

Besides transcription factors, CpG islands are another important regulator of transcription. In most cases, CpG islands are in their hypomethylated state, even in promotor regions or when genes are inactive. It is now well established that hypermethylation of CpG islands in the promoter region of tumor suppressor genes leads to gene inactivation [39–41]. CpG islands can be in different regions of the

DNA, such as at the 3'UTR or 5'UTR region, but also have different transcription start sites. Most CpG islands are not methylated if they are located at the transcription start site, but methylation of those CpG islands is associated with long-term silencing [39]. From the feature CpG islands of our melanoma cluster, we identified those CpG islands in the TSS regions on the DNA that are also present in the local neighborhood of *IRF4, MITF, RXRG*, and *PAX3*. This resulted in potential interesting candidates that could influence the *MITF* and *PAX3* expression levels. Here we have selected *HES5* (4 hypermethylated CpG islands in the TSS200 region, 3 hypermethylated islands in the TSS1500 region), *RNF207* (4 CpG islands in the TSS1500 region, 1 CpG islands in TSS200 region), and *ICMT* (hypermethylation in the TSS1500 region) as potential interesting methylated genes, with β-values above 0.7. Since these patterns of hypermethylation are mostly observed in the melanoma clusters (Figure 3.5A, 3.5B), they might play a driving role in melanoma. Moreover, Hypermethylation of *HES5* and *RNF207* leads to low expression of *HES5* and *RNF207* (Figure 3.6).

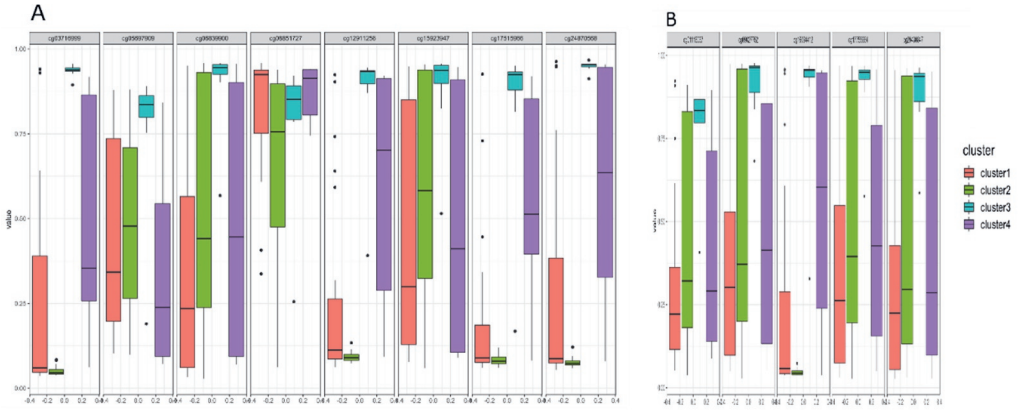


Figure 3.5 A: Methylation values of different CpG island probes associated with the transcription start site 1500 for the genes RNF207 (cg24870568, cg17515966, cg12911256, cg03716999), HES5 (cg15923947, cg06839900, cg05697909) and ICMT (cg08851727). **B:** Methylation values of the CpG islands probes associated with the transcription start site 200 for genes RNF207 (cg16094412) and HES5 (cg20430847, cg17755964, cg09827752, cg01116067).

It is of great interest to further investigate these interesting features mentioned before. Especially the role of *MITF* and *PAX3* in melanoma: *MITF* is associated with pro-survival of melanoma cells [42] whereas *PAX3* is associated with worse survival rates in melanoma patients [43]. Both *HES5* and *MITF* have a genetic interaction with *ABCB5*, a promoter of melanoma metastasis [44] and maintaining melanoma-initiating cells [45]. Hypermethylation of *HES5*, involved in the notch pathway, could have a role in increased *MITF* and *ABCB5* expression.

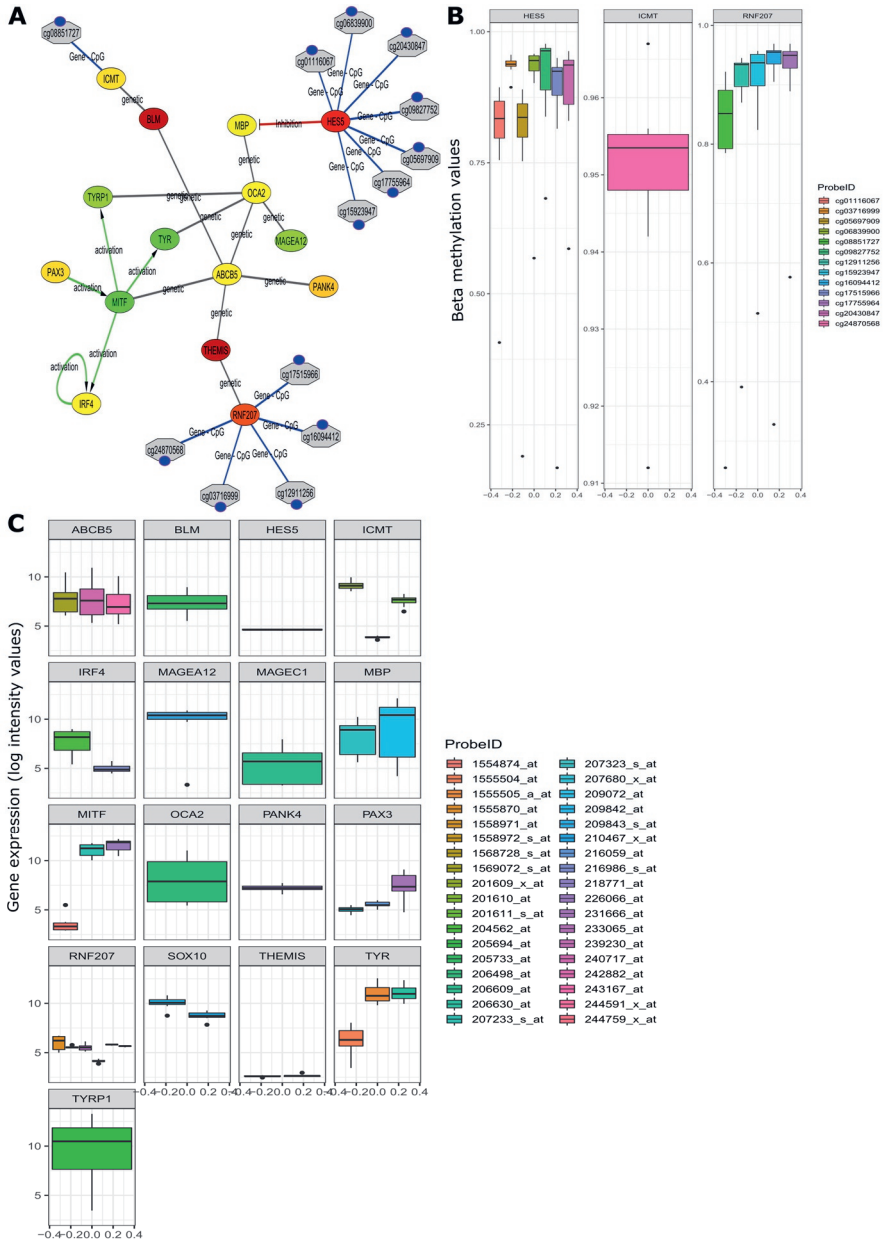


Figure 3.6 A: Local network around focusing around MITF, ABCB5, and IRF4. Nodes: circles are genes, hexagons are CpG islands. Node color of genes represents low expression (red, log intensity of 5) and high expression (green, log intensity 11). Edge color represent: CpG – Gene (blue), transcription activation (green) and genetic interaction (black). CpG hypermethylation is visualized by a small blue circle at the CpG island. **B:** Methylation values of the different CpG islands associated with the three genes that are in the close neighborhood of MITF. **C:** Gene expression values for all the genes in the network.

Discussion

One of the major challenges of this research is to integrate different omics layers into a genomic network that can be used to understand cellular processes. Here, we have updated the Kullback-Leibler single-layer update rule to a multi-layer update rule, instead of working with the Euclidean update rule proposed in the original NMF approach [12]. The Kullback-Leibler (KL) objective function (Equation 6) differs from the Euclidean objective function by assuming a Poison distribution which has a higher resemblance to biology [46, 47]. In previous work, the KL divergence has been shown to outperform the Euclidean distance when analyzing microarray data. Here we show that the KL divergence can be used to extend NMF to integrate multiple platforms. Our results of the simulated toy data set show that the updated rules for multi-layer NMF can identify known groups of samples, even in the presence of noise. In the case of a real biological data set, it was capable of detecting four clusters, of which one was cell-specific and another one cell type enriched.

To evaluate the performance of the multi-layer NMF approach versus the single layer NMF approach, two simulations have been performed to cluster the gene expression data and the methylation data. Here, 4 clusters are observed on each platform and for each cluster the members are identified (supplementary table 3.1.3 for the clusters on transcriptomic data, supplementary table 3.1.4 for the clusters on methylation). The results show a melanoma-specific cluster on both the transcriptomic and methylome platforms, however, the size of the melanoma-specific cluster on the gene expression data is different from the melanoma-specific cluster on the methylation data. It did not detect the high similarity of ME.MDA.ME.435 with the other melanoma cell lines as the multi-layer NMF approach did. One of the advantages of the multi-layer approach is that it can identify for each cluster, different sets of feature genes and CpG islands simultaneously. If one performs a single-layer NMF approach on both data set and tries to join the different findings, based on the overlap between samples in the same methylation and genes clusters, one can only identify a small overlap. This would reduce the number of features and does not imply intra-related features in each platform. The multi-layer approach clusters samples based on each platform and, therefore, is believed to yield inter-platform-related features. Thus the complementary information on each platform improved the clustering of the cancer cell lines.

The workflow addresses the different challenges of omics integration, as previously mentioned. First, multi-layer NMF is an unsupervised clustering approach, which does not require any phenotypic endpoint data as a label. With these unsupervised approaches, it becomes possible to study unknown relationships between the different omics layers in a wide applicable field of research. Second, due to the design of the updated algorithms, where H is dependent on X_i and W_i, it becomes possible to cluster samples based on multiple layers of information. Furthermore,

due to the regularization of H, which makes sure that H only holds values between 0 and 1, the values of W_i depend only on the corresponding omics layer. For instance, W_{gene} is updated based on the values of H and X_{gene} and does not depend on $X_{methylation}$. Therefore, the distribution of the methylation data does not influence the outcome of W_{gene}. With this implementation, it becomes possible to integrate data sets with different distributions, such as methylation and gene expression data.

With our workflow, we were able to generate different genomic interaction networks that can be used to study the interactions between genes, but also between genes and CpG Islands. These interaction networks might be used to identify new biomarkers or potential genes for drug targets. Our results show a strong cluster with only melanoma samples, for which different genes are identified that play a role in melanoma. One should note that of all the melanoma samples, only ME-LOX-IMVI is not present in cluster 3 and therefore is hypothesized to follow a different expression pattern. Although ME-LOX-IMVI is a skin cancer cell line it presents different surface antigen and gene expression profiles [48]. Our results indeed identify a different expression profile in both the epigenome and transcriptome and therefore places ME-LOX-IMVI into a different cluster.

From our genomic interaction network and the melanoma cluster 3 expression profile, different genes and CpG islands can be identified that could play an important role in melanoma. *PAX3, IRF4*, and *MITF* are all three transcription factors with supporting evidence to be tissue enhanced for melanoma (tissue enrichment information from the Human Protein Atlas [49, 50]). The melanoma cells show high expression levels of *MITF*, as well as expression values above log-intensity 7 for *IRF4* and *PAX3* (Figure 3.6). These characteristics are of interest because *PAX3* is positively regulation *MITF* expression levels [51]. However, there is also contradicting evidence of an independent role for both *PAX3* and *MITF* [52], as well as *PAX3* being a repressor of *MITF* [53]. This *PAX3-MITF* signature of the NCI60 melanoma cell lines, in cluster 3, is of importance since in other melanoma cell lines the *PAX3-MITF* switch is proposed [53] to be active, which regulates melanoma fate. Therefore, the effect of knockdown experiments, on cell lines with high *PAX3* and *MITF* expression could differ from cell lines with a *PAX3-MITF* switch (low *PAX3* vs high *MITF* or high *PAX3* vs low *MITF*) and is an important cell line characteristic.

Besides the role of *PAX3* and *MITF*, our results also indicate a role for *IRF4* in melanoma, which has been proposed in other research [54]. *MITF* is transcriptional regulating *IRF4* expression and therefore the increased expression of *MITF* could further increase the expression of *IRF4*. *IRF4* encodes for B-cell proliferation and differentiation and consequently can affect the immune response [55]. This is highly relevant since tumor-associated B-cells induce tumor heterogeneity and increase therapy resistance [56]. We have identified different CpG islands related to genes

that are associated with Hematopoiesis, but also *HES5*, which is involved in the negative regulation of pro-B cell differentiation. Our results show that several TSS1500 and TSS200 CpG Islands are hypermethylated (Figure 3.5A, 3.5B) with low expression values of *HES5* (Figure 3.6B). Therefore, increasing the expression levels of *HES5* by reversing the hypermethylation of CpG islands (Figure 3.5A, 3.5B) could be a potential candidate to disrupt the effect of B-cells and increased *IRF4* expression that leads up to melanoma survival and therapy resistance.

From the genomic interaction network, different genes are identified which share a connection with *MITF*, including *ABCB5, OCA2, MBP, MAGEA12*, and *MAGEC1*. Especially *ABCB5* is of interest, due to its central position in the network (Figure 3.6A). Moreover, *ABCB5* is associated with clinical tumor progression but also with chemotherapeutic resistance [57] and recurrence in melanoma patients [45]. *MAGE* genes are expressed in various tumors, including melanoma. *MAGEA12* is a repressor of tumor-suppressor genes and therefore plays an important role in melanoma [58, 59]. *MAGEA* genes can inhibit p53 function, by directly binding to the p53 DNA binding domain, which will prevent p53 transcription [60]. *OCA2* is involved in melanin biosynthesis, although its role in melanoma is not yet fully understood. Melanin is important for UV-protection but can also regulate epidermal homeostasis, thus can affect melanoma behavior [61]. It is known that melanoma cells do not excrete pigment [62], therefore an increase in melanin biosynthesis could lead to even more pigment in melanoma. The inhibition of melanogenesis is hypothesized to be a good therapeutic target for melanoma therapy [63]. Because of the role of *OCA2*, this would be an important characteristic of the cluster 3 melanoma cells and could, therefore, be a potentially interesting marker to investigate and predict the outcome of melanin inhibitor drugs. Further research is needed to investigate the genetic reactions between those genes to determine if increasing the expression levels of *HES5* will lead to a lower expression of *ABCB5, PAX3, MITF*, and *IRF4*. Increased expression of *HES5* can inhibit *MPB* expression levels and potentially, via genetic interactions,*OCA2*,*ABCB5*,and *MITF*.

In the melanoma cluster, low expression values of *RNF207* correlate with hypermethylation of different *RNF207* TSS1500 regions. The role of *RNF207* in cancer remains unclear, with no evidence suggesting a role of *RNF207* in tumor growth or suppression. Therefore, although hypermethylation is observed in melanoma, it might not be a potentially interesting event.

Although the integrated networks help us to link important changes in methylation to gene expression, but also gene regulatory effects, there are still some remaining challenges. For instance, the role of the transcription start site should be further investigated, especially to determine the importance of the methylation status of both TSS200 and TSS1500 associated with the same gene, to understand the effect of hypermethylation of either one or both TSS regions on gene silencing. It is of great value to understand if one region has a greater effect on gene silencing, or

if both regions have to be hypermethylated before gene silencing occurs. Computational methods can improve if they 'learn' to identify TSS methylation concerning gene expression. The clustering of the different NCI60 cell lines also resulted in two clusters with a mixture of different cell types. Here, different factors could play a role, which could relate to the nature of the cancer cell lines themselves. Multiple cell lines could have general expression patterns with small local tissue-specific patterns. Current computational tools for omics integration should be further updated to extract those local patterns. However, the multi-layer NMF approach already shows that it can analyze biological data sets and identify important features.

Conclusion

The creation of genomic networks through multi-layer NMF adds to the current tools to analyze multi-omics data sets. In this research, we identified important characteristics of the NCI60 melanoma cell lines. Hypermethylation of *HES* family genes leads to a low expression of these genes, which could further affect melanoma behavior. *HES5* shows potential interesting interactions with melanoma-associated genes *IRF4, PAX3, MITF*, and *ABCB5*. Here, we hypothesize that reversing the methylation status of HES5 could lead to a decreased expression of *PAX3, MITF*, and *ABCB5*. This would be of great relevance since *MITF* and *ABCB5* both play a role in developing towards a drug-resistant state. Furthermore, *HES5* regulates *MBP* transcription and therefore could also affect *MAGE12* and *OCA2* expression levels via *MBP*. *OCA2* is involved in melanin biosynthesis and therefore could serve as a potentially interesting marker for melanoma inhibition. These findings show that a multi-layer approach with genomic interaction networks can be a way to analyze data and propose biomarkers or therapeutic targets.

References

1. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer. 2018;103:356–87. doi:10.1016/j.ejca.2018.07.005.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69:7–34. doi:10.3322/caac.21551.
3. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501:328–37. doi:10.1038/nature12624.
4. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer. 2013;108:479–85. doi:10.1038/bjc.2012.581.
5. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. Nat Rev Cancer. 2011;11:726–34. doi:10.1038/nrc3130.
6. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun. 2018;9:4453. doi:10.1038/s41467-018-06921-8.
7. Sun W, Bunn P, Jin C, Little P, Zhabotynsky V, Perou CM, et al. The association between copy number aberration, DNA methylation and gene expression in tumor samples. Nucleic Acids Res. 2018;46:3009–18. doi:10.1093/nar/gky131.
8. Tirado-Magallanes R, Rebbani K, Lim R, Pradhan S, Benoukraf T. Whole genome DNA methylation: beyond genes silencing. Oncotarget. 2017;8. doi:10.18632/oncotarget.13562.

9. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC Syst Biol. 2014;8 Suppl 2:I1. doi:10.1186/1752-0509-8-S2-I1.
10. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17:S15. doi:10.1186/s12859-015-0857-9.
11. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017;8. doi:10.3389/fgene.2017.00084.
12. Lee DD, Seung H.S. Algorithms for non-negative matrix factorization. Adv Neural Inf Proccessing Syst. 2000;:556–62.
13. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci. 2004;101:4164–9. doi:10.1073/pnas.0308531101.
14. Devarajan K, Ebrahimi N. Class Discovery via Nonnegative Matrix Factorization. Am J Math Manag Sci. 2008;28:457–67. doi:10.1080/01966324.2008.10737738.
15. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012;40:9379–91. doi:10.1093/nar/gks725.
16. Wilson TJ, Lai L, Ban Y, Ge SX. Identification of metagenes and their Interactions through Large-scale Analysis of Arabidopsis Gene Expression Data. BMC Genomics. 2012;13:237. doi:10.1186/1471-2164-13-237.
17. Fujita N, Mizuarai S, Murakami K, Nakai K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Sci Rep. 2018;8:9743. doi:10.1038/s41598-018-28066-w.
18. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2015;:btv544. doi:10.1093/bioinformatics/btv544.
19. Yang Z, Zhang H, Yuan Z, Oja E. Kullback-Leibler Divergence for Nonnegative Matrix Factorization. In: Honkela T, Duch W, Girolami M, Kaski S, editors. Artificial Neural Networks and Machine Learning -- ICANN 2011. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 250–7.
20. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007;23:1495–502. doi:10.1093/bioinformatics/btm134.
21. Santolini M, Barabási A-L. Predicting perturbation patterns from the topology of biological networks. Proc Natl Acad Sci. 2018;115:E6375–83. doi:10.1073/pnas.1720589115.
22. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008;6:e264. doi:10.1371/journal.pbio.0060264.
23. Zotenko E, Mestre J, O'Leary DP, Przytycka TM. Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Comput Biol. 2008;4:e1000140. doi:10.1371/journal.pcbi.1000140.
24. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. Cancer Res. 2012;72:3499–511. doi:10.1158/0008-5472.CAN-12-1370.
25. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6:813–23. doi:10.1038/nrc1951.
26. Uong A, Zon LI. Melanocytes in development and cancer. J Cell Physiol. 2010;222:38–41. doi:10.1002/jcp.21935.
27. Riley PA. Melanogenesis and Melanoma. Pigment Cell Res. 2003;16:548–52. doi:10.1034/j.1600-0749.2003.00069.x.
28. Brożyna AA, Jóźwicki W, Roszkowski K, Filipiak J, Slominski AT. Melanin content in melanoma metastases affects the outcome of radiotherapy. Oncotarget. 2016;7. doi:10.18632/oncotarget.7528.
29. Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet. 2012;13:654–66. doi:10.1038/nrg3272.
30. Bedogni B. Notch signaling in melanoma: interacting pathways and stromal influences that enhance Notch targeting. Pigment Cell Melanoma Res. 2014;27:162–8. doi:10.1111/pcmr.12194.
31. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38 suppl_2:W214–20. doi:10.1093/nar/gkq537.
32. Wilkinson AC, Nakauchi H, Göttgens B. Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. Cell Syst. 2017;5:319–31. doi:10.1016/j.cels.2017.07.004.

33. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci. 2014;111:6131–8. doi:10.1073/pnas.1318948111.
34. MITF Drives a Reversible Drug-Tolerant State in Melanoma. Cancer Discov. 2016;6:OF11–OF11. doi:10.1158/2159-8290.CD-RW2016-053.
35. Smith MP, Brunton H, Rowling EJ, Ferguson J, Arozarena I, Miskolczi Z, et al. Inhibiting Drivers of Non-mutational Drug Tolerance Is a Salvage Strategy for Targeted Melanoma Therapy. Cancer Cell. 2016;29:270–84. doi:10.1016/j.ccell.2016.02.003.
36. Aida S, Sonobe Y, Tanimura H, Oikawa N, Yuhki M, Sakamoto H, et al. MITF suppression improves the sensitivity of melanoma cells to a BRAF inhibitor. Cancer Lett. 2017;409:116–24. doi:10.1016/j.canlet.2017.09.008.
37. Medic S, Rizos H, Ziman M. Differential PAX3 functions in normal skin melanocytes and melanoma cells. Biochem Biophys Res Commun. 2011;411:832–7. doi:10.1016/j.bbrc.2011.07.053.
38. Wang L, Yao ZQ, Moorman JP, Xu Y, Ning S. Gene Expression Profiling Identifies IRF4-Associated Molecular Signatures in Hematological Malignancies. PLoS One. 2014;9:e106788. doi:10.1371/journal.pone.0106788.
39. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92. doi:10.1038/nrg3230.
40. Yang M, Park JY. DNA Methylation in Promoter Region as Biomarkers in Prostate Cancer. 2012. p. 67–109. doi:10.1007/978-1-61779-612-8_5.
41. Curradi M, Izzo A, Badaracco G, Landsberger N. Molecular Mechanisms of Gene Silencing Mediated by DNA Methylation. Mol Cell Biol. 2002;22:3157–73. doi:10.1128/MCB.22.9.3157-3173.2002.
42. Hartman ML, Czyz M. Pro-Survival Role of MITF in Melanoma. J Invest Dermatol. 2015;135:352–8. doi:10.1038/jid.2014.319.
43. Liu Y, Cui S, Li W, Zhao Y, Yan X, Xu J. PAX3 is a biomarker and prognostic factor in melanoma: Database mining. Oncol Lett. 2019. doi:10.3892/ol.2019.10155.
44. Wang S, Tang L, Lin J, Shen Z, Yao Y, Wang W, et al. ABCB5 promotes melanoma metastasis through enhancing NF-κB p65 protein stability. Biochem Biophys Res Commun. 2017;492:18–26. doi:10.1016/j.bbrc.2017.08.052.
45. Wilson BJ, Saab KR, Ma J, Schatton T, Putz P, Zhan Q, et al. ABCB5 Maintains Melanoma-Initiating Cells through a Proinflammatory Cytokine Signaling Circuit. Cancer Res. 2014;74:4196–207. doi:10.1158/0008-5472.CAN-14-0582.
46. Vikalo H, Hassibi B, Hassibi A. A statistical model for microarrays, optimal estimation algorithms, and limits of performance. IEEE Trans Signal Process. 2006;54:2444–55. doi:10.1109/TSP.2006.873716.
47. Tu Y, Stolovitzky G, Klein U. Quantitative noise analysis for gene expression microarray experiments. Proc Natl Acad Sci U S A. 2002;99:14031–6. doi:10.1073/pnas.222164199.
48. Koo H-M, VanBrocklin M, McWilliams MJ, Leppla SH, Duesbery NS, Woude GF V. Apoptosis and melanogenesis in human melanoma cells induced by anthrax lethal factor inactivation of mitogen-activated protein kinase kinase. Proc Natl Acad Sci. 2002;99:3052–7. doi:10.1073/pnas.052707699.
49. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas—a tool for pathology. J Pathol. 2008;216:387–93. doi:10.1002/path.2440.
50. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28:1248–50. doi:10.1038/nbt1210-1248.
51. Yang G, Li Y, Nishimura EK, Xin H, Zhou A, Guo Y, et al. Inhibition of PAX3 by TGF-β Modulates Melanocyte Viability. Mol Cell. 2008;32:554–63. doi:10.1016/j.molcel.2008.11.002.
52. He S, Li CG, Slobbe L, Glover A, Marshall E, Baguley BC, et al. PAX3 knockdown in metastatic melanoma cell lines does not reduce MITF expression. Melanoma Res. 2011;21:24–34. doi:10.1097/CMR.0b013e328341c7e0.
53. Eccles MR, He S, Ahn A, Slobbe LJ, Jeffs AR, Yoon H-S, et al. MITF and PAX3 Play Distinct Roles in Melanoma Cell Migration; Outline of a “Genetic Switch” Theory Involving MITF and PAX3 in Proliferative and Invasive Phenotypes of Melanoma. Front Oncol. 2013;3. doi:10.3389/fonc.2013.00229.
54. Praetorius C, Grill C, Stacey SN, Metcalf AM, Gorkin DU, Robinson KC, et al. A Polymorphism in IRF4 Affects Human Pigmentation through a Tyrosinase-Dependent MITF/TFAP2A Pathway. Cell. 2013;155:1022–33. doi:10.1016/j.cell.2013.10.022.
55. Han J, Qureshi AA, Nan H, Zhang J, Song Y, Guo Q, et al. A Germline Variant in the Interferon Regulatory Factor 4 Gene as a Novel Skin Cancer Risk Locus. Cancer Res. 2011;71:1533–9. doi:10.1158/0008-5472.CAN-10-1818.
56. Somasundaram R, Zhang G, Fukunaga-Kalabis M, Perego M, Krepler C, Xu X, et al. Tumor-associated B-cells induce tumor heterogeneity and therapy resistance. Nat Commun. 2017;8:607. doi:10.1038/s41467-017-00452-4.

57. Chartrain M, Riond J, Stennevin A, Vandenberghe I, Gomes B, Lamant L, et al. Melanoma Chemotherapy Leads to the Selection of ABCB5-Expressing Cells. PLoS One. 2012;7:e36762. doi:10.1371/journal.pone.0036762.
58. Yanagi T, Nagai K, Shimizu H, Matsuzawa S-I. Melanoma antigen A12 regulates cell cycle via tumor suppressor p21 expression. Oncotarget. 2017;8. doi:10.18632/oncotarget.19497.
59. ZHAO G, BAE JY, ZHENG Z, PARK HS, CHUNG KY, ROH MR, et al. Overexpression and Implications of Melanoma-associated Antigen A12 in Pathogenesis of Human Cutaneous Squamous Cell Carcinoma. Anticancer Res. 2019;39:1849–57. doi:10.21873/anticanres.13292.
60. Marcar L, MacLaine NJ, Hupp TR, Meek DW. Mage-A Cancer/Testis Antigens Inhibit p53 Function by Blocking Its Interaction with Chromatin. Cancer Res. 2010;70:10362–70. doi:10.1158/0008-5472.CAN-10-1341.
61. Slominski RM, Zmijewski MA, Slominski AT. The role of melanin pigment in melanoma. Exp Dermatol. 2015;24:258–9. doi:10.1111/exd.12618.
62. Lazova R, Pawelek JM. Why do melanomas get so dark? Exp Dermatol. 2009;18:934–8. doi:10.1111/j.1600-0625.2009.00933.x.
63. Slominski A, Zbytek B, Slominski R. Inhibitors of melanogenesis increase toxicity of cyclophosphamide and lymphocytes against melanoma cells. Int J Cancer. 2009;124:1470–7. doi:10.1002/ijc.24005.

Chapter 3.2: From multi-omics integration towards novel genomic interaction networks to identify key cancer cell line characteristics

Abstract
Cancer is a complex disease where cancer cells express epigenetic and transcriptomic mechanisms to promote tumor initiation, progression, and survival. To extract relevant features from the 2019 Cancer Cell Line Encyclopedia (CCLE), a multi-layer nonnegative matrix factorization approach is used. We used relevant feature genes and DNA promoter regions to construct genomic interaction network to study gene-gene and gene – DNA promoter methylation relationships.

Here, we identified a set of gene transcripts and methylated DNA promoter regions for different clusters, including one homogeneous lymphoid neoplasms cluster. In this cluster, we found different methylated transcription factors that affect transcriptional activation of *EGFR* and downstream interactions. Furthermore, the hippo-signaling pathway might not function properly because of DNA hypermethylation and low gene expression of both *LATS2* and *YAP1*. Finally, we could identify a potential dysregulation of the *CD28-CD86-CTLA4* axis.

Characterizing the interaction of the epigenome and the transcriptome is vital for our understanding of cancer cell line behavior, not only for deepening insights into cancer-related processes but also for future disease treatment and drug development. Here we have identified potential candidates that characterize cancer cell lines, which give insight into the development and progression of cancers.

Published in Scientific Reports
Supplementary data available at: https://github.com/TJMKuijpers/PhDThesis
[1] Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

Introduction

Different hallmarks of cancer have been identified that contribute to the development and propagation of tumors [1]. These hallmarks include sustaining proliferative signaling, evading growth suppressors, resisting cell death, and activating invasion and metastasis. Evading growth suppressors is achieved by the inhibition of the expression of certain genes, called tumor suppressor genes [1]. Tumor suppressor genes regulate important processes such as preventing unrestrained cellular growth, DNA repair promotion, and cell cycle checkpoint activation [2]. Besides tumor suppressor genes, oncogenes play a crucial role in regulating cellular growth, division, and survival [2]. Tumorigenesis is likely to be driven by events that result in the gain of an oncogene or the loss of the suppressor gene, and tumor maintenance often depends on continued oncogene activity [3]. However, the order in which both events happen differ per tumor type. Most hematopoietic cancers and soft-tissue sarcomas are initiated by oncogene activation, followed by alterations in tumor-suppressor genes and other oncogenes [4]. Whereas some carcinomas are initiated by first, a loss of function of a tumor-suppressor gene, and second, alterations in oncogenes and additional tumor-suppressor genes [4]. Although mutations in tumor suppressor genes are important, it is not the only mechanism responsible for alternated gene expression [5]. Genomic instability plays a major part in the activation of oncogenes and subsequently, the inhibition of tumor suppressor genes, thus suggesting a role for epigenomics. For example, inactivation of *BRCA1* in sporadic breast cancer is not due to a mutation but promoter hypermethylation [6].

Almost all cancer cells show genomic instability [7]. In healthy cells, chromatin and associated epigenetic mechanisms ensure stable gene expression and cellular states. Cancer cells show important alterations in these epigenetic mechanisms, which represent one of the fundamental characteristics of nearly all human cancers [8]. A large number of cancer cells show an increase in methylation of normally unmethylated CpG islands and promoter regions of tumor suppressors and DNA repair genes [9]. It has been shown that the increase in DNA methylation increases genomic instability by causing genetic mutations in the DNA sequence [10].

DNA methylation alterations are also associated with drug treatment sensitivity, for example, hypermethylation of DAPK in colon and breast cancer [11]. These findings suggest that aberrations DNA methylation might affect certain pathways that prevent cancer cells from advancing towards apoptosis or other cell death-related mechanisms, as well as towards the development of drug resistance.

Although we know that epigenetic and transcriptional mechanisms play an important role in tumor development, there are still gaps in our current knowledge. DNA hypermethylation is specifically and locally augmented at CpG islands of tumor suppressor genes but its role in tumorigenesis is controversial [12]. DNA hypermethylation of tumor suppressor genes or genes involved in cell cycle

processes are more frequent than their mutation in cancer cells. Consequently, we observe hundreds of methylated DNA regions in cancer cell lines, whereas we only find a few mutated genes that drive tumor onset. Different genes and DNA methylation regions play a role in different types of cancers, and therefore it is even harder to get a clear view of the interplay between DNA methylation and gene expression in carcinogenesis. Identifying key characteristic profiles of DNA methylated regions and alterations in gene expression in cancer cell lines is therefore of major relevance for understanding epigenome/transcriptome interactions in human tumors.

In the present study, to better understand the interplay between the epigenome and the transcriptome, we propose a systems biology framework that allows us to i) classify samples of cancer cell lines based on their epigenetic and transcriptomic signature, and ii) extract relevant features from these clusters to construct a cross-omics interaction network.

Therefore, we apply a multi-layer Nonnegative Matrix Factorization (multi-layer NMF) to obtain a set of transcriptome/epigenome clusters with their corresponding biological features. Nonnegative matrix factorization has already been successfully applied to distinguish between different types of cancers by extracting relevant genomic features and has been applied to investigate the relationship between omics data [13]. Expanding the workflow with the construction of the genomic interaction networks allows us, to not only study the effect of DNA methylation on one gene but could be used to study how one alternation in that specific gene can influence other genes. This could potentially give new insight into the interplay between epigenetic and transcriptomic alterations in cancer cells.

Method

Gene expression and DNA methylation data

For this study, normalized gene expression data and DNA promoter methylation data have been downloaded from the Cancer Dependency Portal (DepMap). Gene expression data is downloaded as Log2(TPM+1) expression values. Gene expression levels have been measured through RNA-sequencing on the Illumina HiSeq 2000 or HiSeq 2500 instruments with sequence coverage of no less than 100 million paired 101 nucleotides-long reads per sample. RNA-seq reads were aligned to the GrCH37 using STAR 2.4 [14].

DNA methylation is measured by Reduced Representation Bisulfite Sequencing (RRBS) analysis to assess promoter methylation. RBBS utilized the MspI cutting pattern to digest DNA to enrich for CpG dinucleotides [15]. The fragments are sequenced on an Illumina HiSeq 2000 and aligned to the hg19 genome using MAQ [15]. A fixed window size of 1000 bp upstream of the transcription starting site for each gene is used to calculate a coverage- weighted average of CpG methylation.

RRBS yielded robust coverage of 17,182 gene promoter regions with average coverage greater than 5 reads for the 843 cell lines.

Multi-layer Nonnegative Matrix Factorization

The original data matrix X_i is estimated by the product HW_i for each data layer (Equation 1). To find a local optimal solution, matrices W_i and H are updated by their update rules (Equations 2 and 3 respectively) and minimizing the Kullback-Leibler divergence (Equation 4). For H we take into account the effect of the different omics layers via X_i and W_i whereas for W_i we take into the effect the omics layers via Xi and the sample clustering via H. In the end, n matrices W are obtained that store the latent features and one coefficient matrix H that stores the clustering coefficients.

$$\sum_{i=1}^{n} X_i \approx \sum_{i=1}^{n} W_i H \qquad (1)$$

$$W_{w+1} = W * \frac{\frac{X_i}{W_i H} H^T}{\sum H} \qquad (2)$$

$$H_{H+1} = H * \frac{\sum \frac{X_i}{W_{iH}}^T}{\sum W_i^T} \quad (3)$$

$$KL\ divergence = \sum \left(X_i * \log \left(\frac{X_i}{W_i H} - X_i + W_i H \right) \right) \quad (4)$$

Feature extraction from NMF results

To analyze the difference in methylation and gene expression profile of each cluster, each matrix W_i is scored by using the method proposed by Kim et al [16]. For each cluster, the entities are selected as features, if those entities that have a high probability of explaining a cluster.

Genomic interaction networks

The biological features obtained for each cluster are used to create a genomic interaction network. These networks consist of DNA promoter region – Gene interactions, to study the relationship between DNA methylation and gene expression, to identify transcription factor – target interactions, Gene – Gene interactions, and protein-protein interactions as well as to gather information about cell line-specific genes related to cancer. In the genomic interaction network, we allow connections if both genes are in the feature list or if expressed genes from the feature list are connected by one seeding node.

Gene – Gene interactions have been downloaded from OmniPath [17] for each in the extracted feature list. Transcription factor – target interactions are added to the network from a transcription factor library built by Souza et al [18], while protein-protein interactions have been downloaded from STRINGdb [19].

Tissue or cancer specific genes and CpG regions
The Human Protein Atlas (HPA) database is used to download information on tissue specificity for lymphoid tissue. Here, we selected those genes that are enriched, meaning their expression levels in lymphoid tissues are at least four times higher compared to other tissues. Cancer or disease specific genes are identified if there is evidence that their protein form is disease or cancer related.

Results
To estimate the number of clusters in the data set, multiple simulations with different cluster sizes k have been performed to get the silhouette score for every proposed multi-NMF solution. Here, we picked values for k in the range of 6 to 11 due to the fact that earlier research suggested at least 6 clusters [20]. Our method predicts the most optimal solution for 8 clusters in our data (Figure 3.7A). The solution for k=8 is above the threshold of 0.7 for a cluster to be regarded stable, but more important, visual inspection of the consensus map of shows multiple stable clusters, as well as a few clusters that contain some noisy samples (Figure 3.7C). We observe some clusters that express a strong signal and appear stable across all simulations, while some samples tend to occasionally shift between clusters.

For each of these clusters, we have identified the number of samples (Figure 3.7B), as well as the cancer types of each sample in that particular cluster (Figure 3.8A). From figure 3.8A, it becomes clear that we have been able to identify genomic profiles, by the combination of DNA promotor site methylation and gene expression, which results in two homogenous and six heterogeneous clusters. Three clusters (clusters 3, 5, and 8) show a high diversity of cancer types, including carcinomas, sarcomas, and blastomas. However, there are two clusters (cluster 1 and cluster 7) that are very homogenous and consist of carcinomas (cluster 1) and lymphoid neoplasms (cluster 7). These two clusters can be of interest for further investigation, to analyze whether these cancer cell lines consist of a generic DNA promoter methylation and gene expression profile.

We extracted unique feature genes and DNA promoter regions for each cluster that explain the observed clustering (Figure 3.8B). Between the clusters, there exists a different weight of the importance of DNA promoter region features versus gene features driving the classification. In clusters 3, 4 and 6 there are no DNA promoter regions found that explained the classification and the cluster is only defined by a set of transcripts. On the contrary, the formation of cluster 7 can be explained by evaluating the combination of the different DNA promoter regions and genes. Here, we have applied two scoring functions: a method by Kim et al [16] as introduced in

the method section and a more stringent cutoff for features that score >0.95 for the transcriptome layer and >0.80 for epigenome layer regions (threshold determined based on the distribution of the scoring functions).

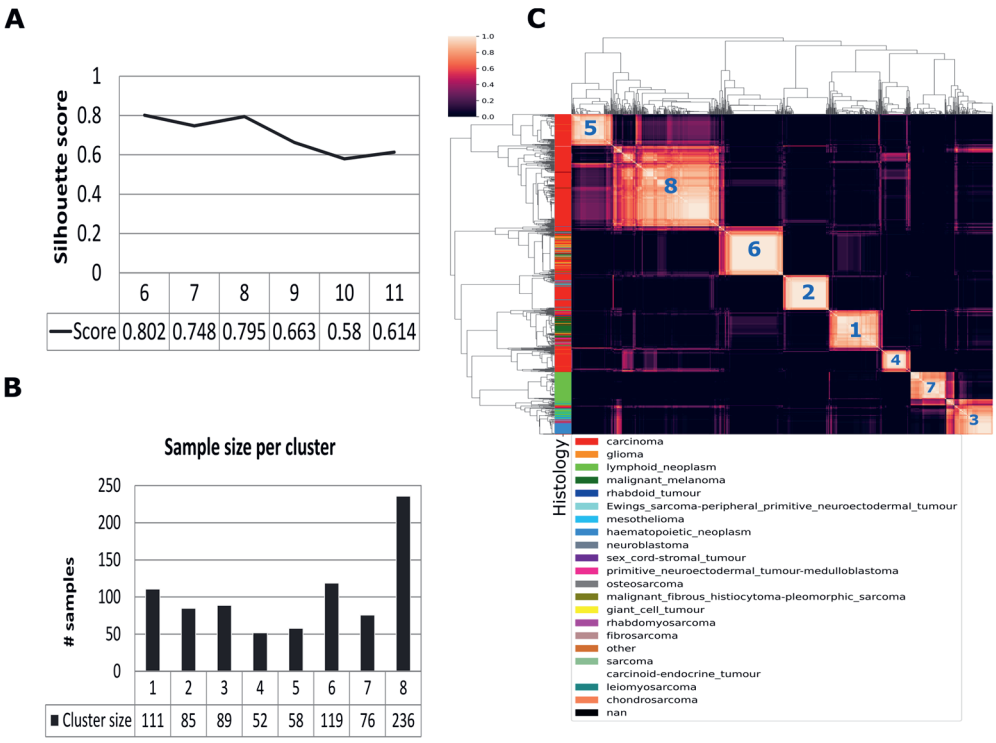


Figure 3.7 A: Silhouette score for cluster sizes 6 to 10. **B:** Number of samples in each cluster. **C:** Consensus plot for cluster size 8.

To further investigate the observed heterogeneity within particular clusters, we have looked into various factors that may explain the clustering, including tissue type, TP53 mutation, Race, and Sex (Supplementary file 3.2.1). Here, it can be seen that there is a different distribution of tissues over the clusters. Cluster 7 again shows a homogenous distribution of only lymphoid tissue (lymphoid neoplasms), whereas other tissues are distributed across multiple clusters. This may indicate that for those tissue types different genomic profiles are driving the clustering.

Next, we investigate why cancer cell lines are separated into different clusters, and have analyzed the features for cluster 5 and 8. Both clusters are selected because of the overlap of the tissue types in cluster 5 in cluster 8 (Figure 3.9). This enabled

the identification of genomic features that are different between the tissue types (Figure 3.10).

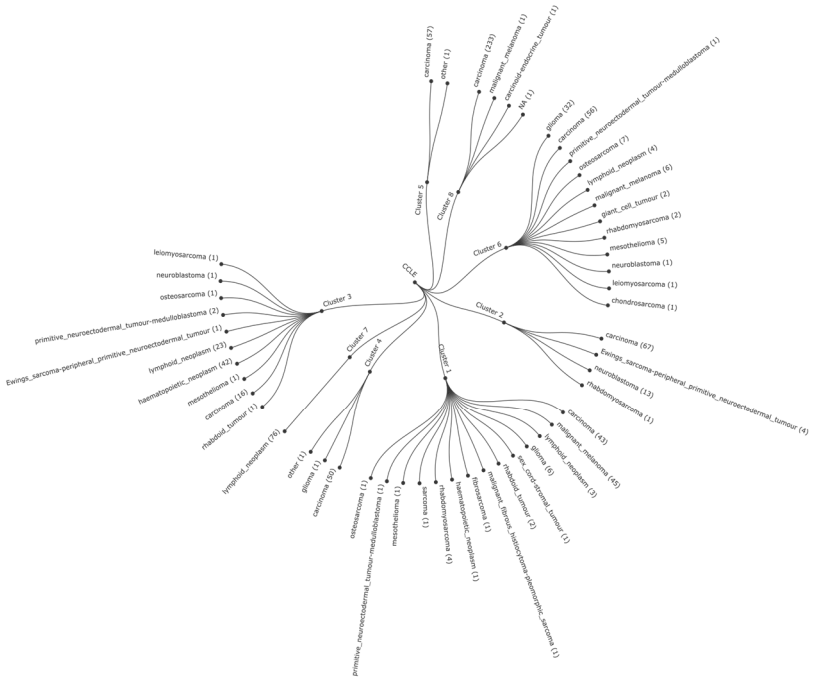
A major difference between clusters 5 and 8 is the DNA promoter region methylation of *ZEB1* and *VAV3* (Figure 3.10B). A strong difference in methylation patterns across all cell lines is observed between the clusters. Within the transcriptome, a different expression pattern of a set of genes is visible between clusters 5 and 8. For some genes, there is no expression in cluster 5 whereas these are expressed in cluster 8 and vice versa (Figure 3.10A, median Log2(TPM+1) value). This includes *FN1, CALD1, THBS1, TAGLN, AXL, HEXB, REG4*, and *LGALS4*.

Cluster 7 appears to contain a large number of features from both the transcriptome and epigenome platforms. Interestingly, this may point towards a genomic profile that is overlapping between cancer cell lines even though they are histologically different. Therefore, we select this cluster for further investigating the underlying genomic profile.

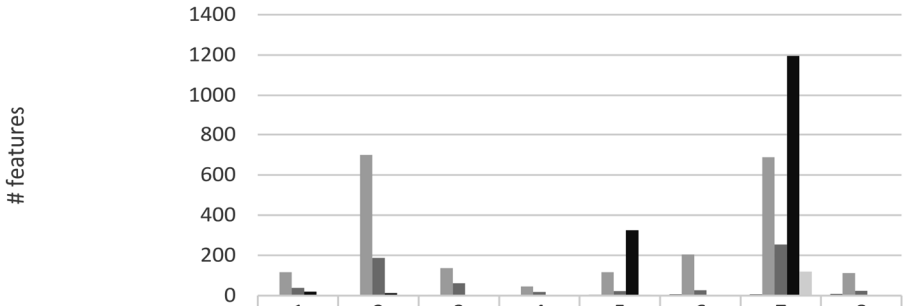
The features of cluster 7 are mapped against the PantherDB to extract the gene ontology biological processes. Several expressed genes are related to immune response, including the adaptive immune response (p-value 1.28E-09), innate immune response (p-value 2.66E-04) complement activation (p-value 1.44E-02), and immunoglobulin-mediated response (p-value 3.76E-03). For each of the DNA promoter regions, the corresponding gene ID is mapped to identify the biological processes. Here, we found signaling processes such as regulation of signaling (p-value 8.97E-09), negative regulation of signaling (p-value 2.26E-09), regulation of signal transduction (p-value 5.16E-09), regulation of cell communication (p-value 6.97E-08) and Hippo signaling pathway (p-value 1.02E-02).

To identify which genes are specific for blood and lymphoid tissue, we have mapped feature genes to the Human Protein Atlas (HPA) Database [21], a database that can be used to categorize genes based on expression level and tissue distribution. We have identified 19 enriched genes for blood and lymphoid tissue, as well as genes disease- or cancer-associated genes. For each DNA promoter region, we mapped the associated gene against HPA. Although there are no known lymphoid tissue-enriched genes in the DNA promoter region feature list, there are some known cancer-related genes (see Table 3.1).

A



B



	1	2	3	4	5	6	7	8
Cluster	1	2	3	4	5	6	7	8
# Feature genes (Kim)	117	699	136	45	117	205	688	112
# Feature genes (0.95)	39	188	62	18	23	27	254	24
# Feature CpG (Kim)	19	13	0	0	324	0	1196	0
# Feature CpG (0.80)	0	0	0	0	5	0	119	0

Figure 3.8 A: The histology of each cluster member as defined by the 2019 CCLE metadata. Here, it becomes apparent that there are a number of clusters with mixed cancer types, but more importantly, there are clusters that show strong heterogeneity. **B**: Histogram for the number of feature genes and DNA promoter regions with Kim score and a more stringent feature score.

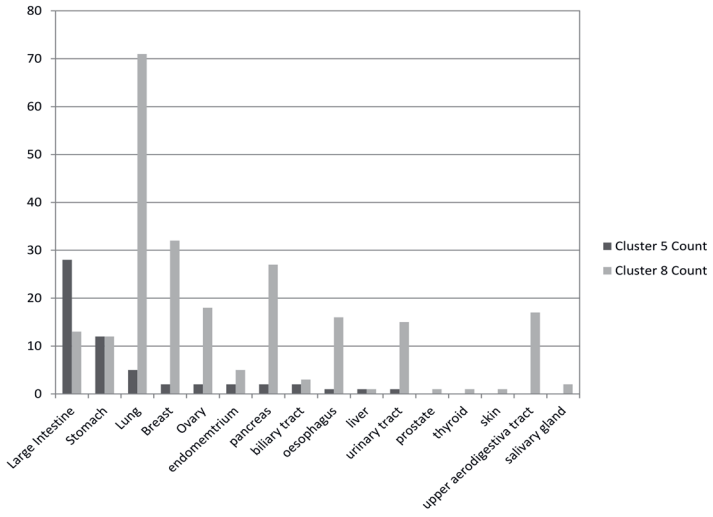


Figure 3.9 Distribution of the cancer tissues in cluster 5 and cluster 8. All cancer tissues in cluster 5 are also present in cluster 8 but cluster 8 also contains some unique cancer tissues.

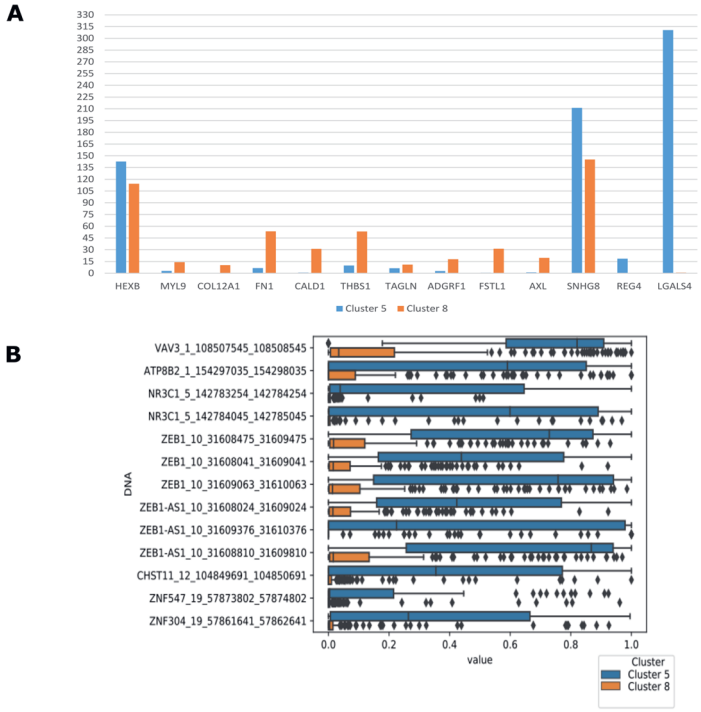


Figure 3.10 A: Median Log2(TPM+1 values of genes in cluster 5 and cluster 8. This shows the main drivers behind the stratification of cluster 5 and cluster 8. **B**: DNA promoter region of cluster 5 and cluster 8. Although cluster 5 and cluster 8 both contain the same cancer tissues, a different methylation pattern is observed for certain regions.

Finally, we used the DNA promoter regions and expressed genes to construct genomic interactions, to study interactions between and within the transcriptome and epigenome. Here, we focused on the genomic interaction network, because this cluster gave a strong homogeneous signal for lymphoid neoplasms. From the total genomic interaction network (Supplementary figure 3.2.1), we have identified potential interesting network neighborhoods based on genes that have a high degree, genes that are transcription factors or genes mentioned in Table 3.1. Figure 3.11A shows the subnetwork of genes that are located around the epidermal growth factor (*EGFR*), a gene with a high inner and outer degree. *EGFR* can be transcriptionally activated by two feature methylated genes *KLF5* and *CREBPD*. Furthermore, *EGFR* shares protein-protein interactions with the oncogene *FGR* and *PTK2* and therefore this subnetwork can be important to study in more detail.

The second cluster of genes of interest is located in the neighborhood of *LATS2* and *YAP1* (Figure 3.11B). These two genes play a role in the Hippo signaling pathway, a pathway believed to play a pivotal role in cancer [22]. Finally, we have identified a third subnetwork centralized around the lymphoid tissue enriched genes *CD28* and *CD86*, which form the *CD28-CD86* pathway (Figure 3.11C). The two subnetworks are of interest because of their role in the signaling pathways.

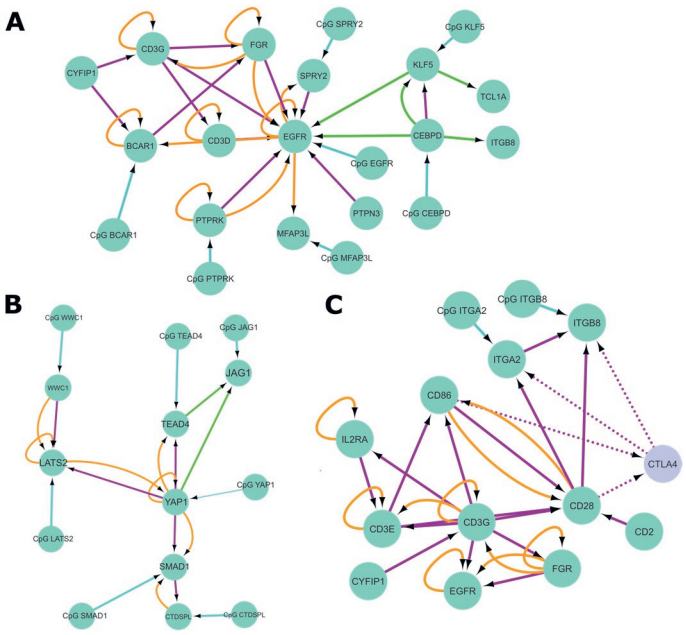


Figure 3.11 Genomic interaction network modules for cluster 7 (lymphoid neoplasm). **A**: Subnetwork for the genes connected to EGFR. **B**: Subnetwork for the region of genes connected to YAP1, TEAD4, JAG1, and SMAD1. **C**: Subnetwork for the set of genes connected with CD28, CD86, and CTLA4. CTLA4 is a seeding node highlighted by the light color and the dotted interactions. Gene-gene interactions are shown in orange, DNA promoter region - gene interactions in cyan, protein-protein interactions in purple, and transcription interactions in green (activation) or red (inhibition).

For each of the genes in the three subnetworks, we compared the gene expression and methylation values for the different feature genes and DNA promoter regions. Here we can see that hypermethylation of the DNA promoter region (Supplementary figure 3.2.2, plot A-H) corresponds with low gene expression for *EGFR, CEBPD, KLF5, YAP1, LATS2, NFIB, LRCC49*, and *ARHCAP29* (Supplementary figure 3.2.3: plots A-H).

Table 3.1 List of genes and DNA promoter region-associated genes that are either tissue-specific, cancer-related or disease-related

Genes		Methylated genes
Blood and lymphoid tissue enriched	**Cancer or Disease- related**	**Cancer or Disease-related genes**
SASH3,CD48	*CREB1,FGR*	*ACVR2A,SPRY2*
FGR,SEPT1	*TNFSF13B,CD79A*	*SMAD1,WDFY3*
CD86,IGLL5	*CD79A,CD19*	*ITGA2,CRY1*
LY9,TNFSF13B	*CCR7,CTLA4*	*PTPRL,LRIG3*
CD79A,CD19	*TNFSF13,FCRL3*	*PTK2,BCL2L2*
LAX1,VPREB1		*NFIB,MYO1E*
TNFSF13,IGJ		*NRP1,TGIF1*
CD28,CCR7		
FCRL3,IGLL1		
FCRL1		

Discussion

Cancer is one of the most complex diseases and the same types of tumors can exhibit different genomic traits. The challenge here is to discern whether similar aberrations in different histologies (cross-cancer similarity) have a comparable biological significance. There are five histology classes of cancer: carcinoma, sarcoma, myeloma, leukemia, and lymphoma. Each class has different subclasses according to the origin of the cancer cell. However, there is a shift in the importance of histology as a marker for cancer types. More and more cancers are found to share a set of genetic features, even if they do not belong to the same subclass. It is more important to identify key genomic similarities shared by subgroups of cancer since they present an opportunity to design tumor treatment strategies among tumors regardless of the tissue of origin [23]. Our genomic interaction networks as reported in this study for lymphoid neoplasms could help to identify and further investigate the key genomic characteristics.

Upon having integrated epigenomics and transcriptomics data across a wide range of cancer cell lines, our results demonstrate clusters that contain a mixture of different cancer cell line samples, therefore also a mixture of cancer types. When we look into the genomic features of a given cluster, a set of transcripts and DNA promoter regions is identified that may explain the separation of the same cancer tissues. In cluster 5, which contains the same type of cancer tissues as cluster 8, a different methylation profile is observed in the DNA promoter regions of some key genes. This is of great interest since this might point towards the fact that the same cancer tissues have different epigenetic and transcriptomic alterations.

There are different transcripts and methylated DNA promoter regions that explain the clustering of the different cancer tissues in cluster 5 or 8. One major difference is the role of certain DNA promoter regions in cluster 5, whereas there is no methylation effect predicted to play a role in cluster 8 (Results in figure 3.10B). If we take into account only the DNA promoter regions with a high probability of explaining cluster 5, it is appears that *ZEB1*, *ZEB1-AS*, and *VAV3* are the most important genes that are hypermethylated. Highly expressed *ZEB1* is associated with malignancy of various cancers, and it plays an important role in cancer transformation [24]. *VAV3* is involved in cell signaling and tumorigenesis [25] and is a prognostic factor of poor prognosis in breast cancer patients [26] as well as an important driver of prostate cancer [27]. The hypermethylation of both *ZEB1* and *VAV3* might indicate that those genes do not play a role in the development and progression of the different cancer cell lines in cluster 5.

Besides the different methylation features, several gene transcripts explain the differences between clusters 5 and 8. In cluster 5, a member of the regenerating gene (REG) family members, *REG4*, is predicted to be discriminative transcriptomic features. The REG family members are small secreted lectin-like proteins involved in hepatic, pancreatic, gastric, and intestinal cell proliferation and differentiation [28]. Aberrant expression of *REG4* is associated with tumor growth, survival, adhesion but also resistance to apoptosis [28]. Elevated expression of *FN1* cluster 8 is of interest, because *FN1* is an important gene involved in the development of various cancer types driving proliferation [29, 30]. *AXL* expression is associated with various processes in cancer, including proliferation, survival, metastasis and resistance to cancer therapy [31]. Due to the role of *AXL* it has been proposed as target for cancer therapy [31, 32]. Due to the absent expression of *AXL* in cluster 5, it might not be an effective strategy in those cancer cell lines. *LGALS4* encodes for the protein galectin 4. Galectins are associated with various diseases including cancer and regulate tumor cell adhesion and migration [33]. Moreover, galectin 4 serves as a strong prediction for metastatic potential of adenocarcinomas [34], a type of carcinoma.

Although heterogeneity seems to play an important role in the clustering of the cancer cell lines, there is one cluster that shows homogeneity towards a class of

cancers. Cluster 7 shows a strong intensity for lymphoid neoplasms. This cluster could give us more insight into the underlying epigenetic and transcriptomic changes in lymphoid neoplasms.
The samples in cluster 7 are from a group of disorders that originate from the neoplastic transformation of lymphocytes. Normally, lymphoid stem cells develop into lymphoid blasts that differentiate towards B or T lymphocytes. Recent research has shown that chronic lymphocytic leukemia and multiple myeloma have a shared biological basis [35]. Furthermore, follicular lymphomas and diffuse large B cell lymphomas show shared gene expression patterns associated with immune escape mechanisms [36]. These current insights show that B and T cell lymphomas potentially share genomic alterations. Lymphoma and leukemia originate from white blood cells, thus potentially share the same genomic alterations leading to the development of normal white blood cells towards cancer cells. Therefore, it is of interest to deeper investigate this genomic interaction network.

It is of no surprise that pathways analysis of the nodes in the genomic interaction network shows several genes involved in immune response regulation. It is known that lymphoid neoplasms is a disease associated with immunological ignorance and immune evasion [37].

In the genomic interaction network for cluster 7, various nodes can be identified that are of potential interest. Here we have made a selection based on methylation status, gene expression, and the role of a specific gene in the established genomic interaction network (Supplementary figure 3.2.1). Methylated promotor regions of *NFIB, ARHGAP29*, and *LRRC49* are predicted to be a feature of cluster 7 meaning that there are drivers of cluster formation. For most samples in cluster 7, the promoter region of *NFIB, ARGHAP29*, and *LRRC49* is hypermethylated. In these samples, the genes *NFIB* and *LRRC49* both have low expression values, whereas *ARHGAP29* is not expressed at all. *ARHGAP29* is one of the protein-coding genes for Rap1 that regulates Rho GTPase signaling. Dysregulation of Rap1 activation is responsible for the development of malignancy [38]. Furthermore, RAP1 interacts with many members of the DNA damage response pathway but RAP1-depleted cells show reduced interaction between DNA ligase IV and DNA-pk and are impaired in DNA ligase IV recruitment to enable efficient repair of damaged chromatin [39].

NFIB, with an increased DNA promoter methylation in cluster 7 cell lines, is a transcription factor regulating the maturation of megakaryocytes, a platelet precursor [40]. Megakaryopoiesis is the developmental process of bone marrow progenitor cells into mature megakaryocytes and is required for normal hemostasis. From the genomic interaction network, we can identify possible interactions between *NFIB* and other genes. *NFIB* shares genetic interactions with *FGR* and *CD28*. FGR is a proto-oncogene of the Src family of tyrosine kinases expressed in immune cells [41]. Src family kinases are most of all best known for their role in

tumor development and progression [42]. *FGR* is not only connected to *NFIB*, but *FGR* also shares a protein-protein interaction with *IGLL5* and a gene-gene interaction with *FCRL1*. *FCRL1* expressed in a majority of chronic lymphocytic leukemia, follicular lymphoma, hairy cell leukemia, and mantle cell lymphoma [43] and might play an important role in the onset of these malignancies. It is therefore of interest to investigate whether the hypermethylation of *NFIB* can be reversed and whether *NFIB* is capable of downregulating *FCRL1* via genetic interactions with *FGR*.

One of the central nodes in the network is *EGFR*, a gene responsible for controlling cellular proliferation, apoptosis, angiogenesis, and metastatic spread in a variety of cell types and tissues [44]. In cluster 7, it is evident that *EGFR* is hypermethylated and consequently is not expressed (expression level of 0 TPM). Because of the hypermethylation of the promoter region, the transcription factors *CEBPD* and *KLF5* cannot transcriptionally activate *EGFR* expression (Figure 3.11A). Even if the DNA promoter region of E*GFR* would be hypomethylated, transcription activation of *EGFR* might not occur since both *CEBPD* and *KLF5* are not expressed in the cancer cell lines of cluster 7. The combination of hypermethylation of *EGFR* and the inactivity of *CEBPD* and *KLF5* is interesting since *EGFR* shares different gene-gene and protein-protein interactions with *FGR, CD3D, CD3G, BCAR*, *PTK2*, and *PTPN3*. The inactivity of *EGFR* could be of importance since this may alter the interactions with *FGR* and *PTK2* and potentially disrupt the functioning of these oncogenes. *EGFR* expression is still a subject of debate in leukemia [45] but in lymphomas, it has been demonstrated to increase drug resistance [46]. Our results show low expression of *EGFR* which could potentially mean that *EGFR* cannot contribute to drug resistance and highlight the mechanism of low *EGFR* expression in these cancer cell lines.

A second local neighborhood of interest is defined around *YAP1*, a gene believed to be involved in the regulation of the hematopoietic system [47]. The role of *YAP1* is important, since in solid tumors it emerges as an oncogene, whereas *YAP*1 seems to exert a tumor-suppressive function in multiple myeloma and leukemia [47]. In our network, *YAP1* can regulate the transcription of *JAG1* and might interact with *LATS2, TEAD4*, and *SMAD1* via protein-protein and gene-gene interactions (Figure 3.11B). These possible interactions and transcriptional activation might be altered because of the methylation status of *YAP1*, which shows a trend towards a higher methylated DNA promoter region, and as a possible effect, there is no *YAP1* expression observed in the cancer cell lines in cluster 7. This result is in agreement with previous research, where downregulation or deletion of *YAP1* in multiple myeloma and leukemia is reported [48]. Due to the inactivity of *YAP1*, it will be of interest to determine whether *JAG1* and *TEAD4* are expressed. *TEAD4* is low expressed in the cell lines of cluster 7, which could be favorable since *TEAD4* expression is associated with tumor onset and progression [49]. *JAG1* is involved in the NOTCH signaling pathway and downregulation of *JAG1* has been proposed

as a target for treatment, since *JAG1* can function as an oncogene in the different lymphoid neoplasms [50]. Similar to *TEAD4, JAG1* is lowly expressed in cluster 7, which could be because *YAP1* is not expressed and therefore cannot activate *JAG1* transcription. Although *YAP1* is proposed as a potential tumor suppressor gene [47], increasing *YAP1* expression might lead to transcriptional activation of the oncogene *JAG1* (Figure 3.11B). In our genomic interaction network, there is also an interaction between *LATS2* and *YAP1*. This interaction is actually of interest since *LATS2* and *YAP1* are two genes involved in the hippo signaling pathway [22]. As mentioned before, *YAP1* has a low gene expression due to DNA hypermethylation and therefore we believe that this protein-protein interaction is affected. Furthermore, *LATS2* is low expressed in cluster 7 in comparison with the other clusters, which could be a consequence of the increased methylation of the DNA promoter region of *LATS2*. This could indicate that in cluster 7 the hippo-signaling pathway might not function properly because of DNA hypermethylation and low gene expression of both *LATS2* and *YAP1*.

A third region of interest emerged while studying the local neighborhood of the genes *CD28, CD86*, *CD80, ITGA2*, and *CTLA4*. *CD28* and *CD86* are both lymphoid tissue enriched genes [21]. The two genes form a co-stimulatory pair and upon CD86-activation, *CD28* can carry out different functions involved in the Th1 differentiation pathway [51], cytokine production, and downstream signaling events of the B cell receptor through the activation of *NFkB* [52]. In the gene interaction network with *CD28* and *CD86*, the dotted interactions around the seeding node *CTLA4* are of relevance (Figure 3.11C). As a seeding node, *CTLA4* does not belong to the features for cluster 7 but its absence is of interest. It becomes clear that *CTLA4* is not expressed in any of the cancer cell lines, whereas *CD28* and *CD86* are expressed only in cluster 7 (Supplementary figure 3.2.3: plot I-K). *CTLA4* is an inhibitor of the CD28 – CD86 activation pathway and humans that carry any *CTLA4* mutations are found to suffer from profound autoimmunity [53]. *CD86* shows elevated expression in cluster 7 in comparison to the other clusters, which could not only indicate that *CD86* is specific for lymphoid neoplasms, but also that the signaling pathway of *CD86-CD28* is perturbed leading to CD28 stimulation. The inactivity of *CTLA4* might result in a loss of the inhibition of the signaling pathway of CD86-CD28 which impacts the differentiation of blood cells (Th1 and B cells) but more interesting, the association of *CTLA4* with autoimmunity might point towards a hypothesis that lymphoid neoplasms might share the same alterations as autoimmune diseases [54]. The absence of *CTLA4* might have other implications, due to the protein-protein interactions with *ITGB8* and *ITGA2*. Dysregulation of the CD28 – CD86 pathway could propagate to *EGFR* and *FGR* expression via the different *CD3* genes as shown in the network.

The previously discussed features are of interest because their changes in expression do not occur for all lymphoid neoplasms. The lymphoid neoplasm samples in clusters 3 and 6 do not have an increased expression of both *CD28* and

CD86 (Supplementary figure 3.2.3). Cluster 6 shows increased *LRRC49* expression whereas this gene is low expressed in all lymphoid neoplasms samples in cluster 7. Furthermore, the hypermethylated DNA regions in cluster 7 are hypomethylated in cluster 3 and cluster 6 (Supplementary figure 3.2.2). The combination of the epigenetic and transcriptomic changes stratify the lymphoid neoplasms in different clusters and might therefore be of relevance.

By integration of the omics layers employing Multi-layer Nonnegative Matrix Factorization, we are capable of separating clusters based on their DNA methylation and gene expression profiles across a wide range of cell lines derived from multiple human cancer types. The combination of these profiles leads to heterogeneous clusters of sarcomas and carcinomas, but also more homogeneous clusters of lymphoid neoplasms. Although our method can extract signals that characterize different cancer types, there is still room for improvement. Heterogeneity remains a problem, which will be difficult to solve. One way to overcome this is by performing omics integration on one class of cancer cell lines. We expect that this would improve the integration and would select more subtype-specific signals. However, our findings from the complete 2019 CCLE clarify that our method is indeed capable of identifying possible important characteristics. We can identify different methylated DNA promoter regions in the same cancer tissues, but we are also able to construct a genomic interaction network for lymphoid neoplasms based on specific genomic features for that cancer type. This genomic interaction network helps us to identify the possible relationship between methylated genes and other genes in the network. We have identified different methylated DNA promoter regions that affect transcriptional activation of *EGFR*, which might impact on the protein-protein interactions with the oncogenes *FGR* and *PTK2*. The DNA hypermethylation of *EGFR* could be of interest since this gene contributes to drug resistance. We showed that hypermethylation of *YAP1* leads to low gene expression and as a consequence no transcriptional activation of *JAG1*. Although *YAP1* has tumor-suppressive characteristics, it is relevant to take into account that this may lead to transcriptional activation of the oncogene *JAG1*. Finally, through the genomic interaction network, we could identify a potential dysregulation of the *CD28-CD86-CTLA4* axis in the different lymphoid neoplasms cancer cell lines.

Conclusion

Characterizing the epigenome and transcriptome is vital for our understanding of cancer cell line behavior, not only for better understanding the cancer-related processes but also for future treatment and anti-cancer drug developments. Here, we have identified potential candidate genes that characterize cancer cell lines of the type for lymphoid neoplasms. Our current insights show that, although assumed different, B and T cell lymphomas potentially share similar genomic alterations. These key alterations are important to study and further understand the development and progression of lymphoid neoplasms cancer cell lines.

References

1. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144:646–74. doi:10.1016/j.cell.2011.02.013.
2. Lee EYHP, Muller WJ. Oncogenes and Tumor Suppressor Genes. Cold Spring Harb Perspect Biol. 2010;2:a003236–a003236. doi:10.1101/cshperspect.a003236.
3. Weinstein IB. CANCER: Enhanced: Addiction to Oncogenes--the Achilles Heal of Cancer. Science (80-). 2002;297:63–4. doi:10.1126/science.1073096.
4. Croce CM. Oncogenes and Cancer. N Engl J Med. 2008;358:502–11. doi:10.1056/NEJMra072367.
5. Sager R. Expression genetics in cancer: Shifting the focus from DNA to RNA. Proc Natl Acad Sci. 1997;94:952–5. doi:10.1073/pnas.94.3.952.
6. Esteller M. Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors. J Natl Cancer Inst. 2000;92:564–9. doi:10.1093/jnci/92.7.564.
7. Andor N, Maley CC, Ji HP. Genomic Instability in Cancer: Teetering on the Limit of Tolerance. Cancer Res. 2017;77:2179–85. doi:10.1158/0008-5472.CAN-16-1553.
8. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002;3:415–28. doi:10.1038/nrg816.
9. Lahtz C, Pfeifer GP. Epigenetic changes of DNA repair genes in cancer. J Mol Cell Biol. 2011;3:51–8. doi:10.1093/jmcb/mjq053.
10. Gonzalo S, Blasco MA. Role of Rb Family in the Epigenetic Definition of Chromatin. Cell Cycle. 2005;4:752–5. doi:10.4161/cc.4.6.1720.
11. Lehmann U, Celikkaya G, Hasemeier B, Länger F, Kreipe H. Promoter hypermethylation of the death-associated protein kinase gene in breast cancer is associated with the invasive lobular subtype. Cancer Res. 2002;62:6634–8. doi:12438260.
12. Choi JD, Lee J-S. Interplay between Epigenetics and Genetics in Cancer. Genomics Inform. 2013;11:164. doi:10.5808/GI.2013.11.4.164.
13. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS One. 2017;12.
14. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov G V., Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8.
15. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. 2012;13:R92. doi:10.1186/gb-2012-13-10-r92.
16. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007;23:1495–502. doi:10.1093/bioinformatics/btm134.
17. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13:966–7. doi:10.1038/nmeth.4077.
18. Souza TM, Rieswijk L, Beucken T van den, Kleinjans J, Jennen D. Persistent transcriptional responses show the involvement of feed-forward control in a repeated dose toxicity study. Toxicology. 2017;375:58–63. doi:10.1016/j.tox.2016.10.009.
19. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13. doi:10.1093/nar/gky1131.
20. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov G V., Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8. doi:10.1038/s41586-019-1186-3.
21. Lindskog C. The Human Protein Atlas – an important resource for basic and clinical research. Expert Rev Proteomics. 2016;13:627–9. doi:10.1080/14789450.2016.1199280.
22. Han Y. Analysis of the role of the Hippo pathway in cancer. J Transl Med. 2019;17:116. doi:10.1186/s12967-019-1869-4.
23. Liu Z, Zhang S. Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. BMC Genomics. 2015;16:503. doi:10.1186/s12864-015-1687-x.
24. Zhang Y, Xu L, Li A, Han X. The roles of ZEB1 in tumorigenic progression and epigenetic modifications. Biomed Pharmacother. 2019;110:400–8. doi:10.1016/j.biopha.2018.11.112.
25. Van Aelst L, D'Souza-Schorey C. Rho GTPases and signaling networks. Genes Dev. 1997;11:2295–322. doi:10.1101/gad.11.18.2295.
26. CHEN X, CHEN S, LIU X-A, ZHOU W-B, MA R-R, CHEN L. Vav3 oncogene is upregulated and a poor prognostic factor in breast cancer patients. Oncol Lett. 2015;9:2143–8. doi:10.3892/ol.2015.3004.

27. Dong Z, Liu Y, Lu S, Wang A, Lee K, Wang L-H, et al. Vav3 Oncogene Is Overexpressed and Regulates Cell Growth and Androgen Receptor Activity in Human Prostate Cancer. Mol Endocrinol. 2006;20:2315–25. doi:10.1210/me.2006-0048.
28. Kawasaki Y, Matsumura K, Miyamoto M, Tsuji S, Okuno M, Suda S, et al. REG4 is a transcriptional target of GATA6 and is essential for colorectal tumorigenesis. Sci Rep. 2015;5:14291. doi:10.1038/srep14291.
29. Sun Y, Zhao C, Ye Y, Wang Z, He Y, Li Y, et al. High expression of fibronectin 1 indicates poor prognosis in gastric cancer. Oncol Lett. 2019. doi:10.3892/ol.2019.11088.
30. Li B, Shen W, Peng H, Li Y, Chen F, Zheng L, et al. Fibronectin 1 promotes melanoma proliferation and metastasis by inhibiting apoptosis and regulating EMT. Onco Targets Ther. 2019;Volume 12:3207–21. doi:10.2147/OTT.S195703.
31. Rankin E, Giaccia A. The Receptor Tyrosine Kinase AXL in Cancer Progression. Cancers (Basel). 2016;8:103. doi:10.3390/cancers8110103.
32. Zhou L, Liu X-D, Sun M, Zhang X, German P, Bai S, et al. Targeting MET and AXL overcomes resistance to sunitinib therapy in renal cell carcinoma. Oncogene. 2016;35:2687–97. doi:10.1038/onc.2015.343.
33. Bartolazzi A. Galectins in Cancer and Translational Medicine: From Bench to Bedside. Int J Mol Sci. 2018;19:2934. doi:10.3390/ijms19102934.
34. Hayashi T, Saito T, Fujimura T, Hara K, Takamochi K, Mitani K, et al. Galectin-4, a Novel Predictor for Lymph Node Metastasis in Lung Adenocarcinoma. PLoS One. 2013;8:e81883. doi:10.1371/journal.pone.0081883.
35. Went M, Sud A, Speedy H, Sunter NJ, Försti A, Law PJ, et al. Genetic correlation between multiple myeloma and chronic lymphocytic leukaemia provides evidence for shared aetiology. Blood Cancer J. 2019;9:1. doi:10.1038/s41408-018-0162-8.
36. Laurent C, Charmpi K, Gravelle P, Tosolini M, Franchet C, Ysebaert L, et al. Several immune escape patterns in non-Hodgkin's lymphomas. Oncoimmunology. 2015;4:e1026530. doi:10.1080/2162402X.2015.1026530.
37. Curran EK, Godfrey J, Kline J. Mechanisms of Immune Tolerance in Leukemia and Lymphoma. Trends Immunol. 2017;38:513–25. doi:10.1016/j.it.2017.04.004.
38. Hattori M. Rap1 GTPase: Functions, Regulation, and Malignancy. J Biochem. 2003;134:479–84. doi:10.1093/jb/mvg180.
39. Khattar E, Maung KZY, Chew CL, Ghosh A, Mok MMH, Lee P, et al. Rap1 regulates hematopoietic stem cell survival and affects oncogenesis and response to chemotherapy. Nat Commun. 2019;10:5349. doi:10.1038/s41467-019-13082-9.
40. Chen L, Kostadima M, Martens JHA, Canu G, Garcia SP, Turro E, et al. Transcriptional diversity during lineage commitment of human blood progenitors. Science (80-). 2014;345:1251033–1251033. doi:10.1126/science.1251033.
41. Kovács M, Németh T, Jakus Z, Sitaru C, Simon E, Futosi K, et al. The Src family kinases Hck, Fgr, and Lyn are critical for the generation of the in vivo inflammatory environment without a direct role in leukocyte recruitment. J Exp Med. 2014;211:1993–2011. doi:10.1084/jem.20132496.
42. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012;40:9379–91. doi:10.1093/nar/gks725.
43. Du X, Nagata S, Ise T, Stetler-Stevenson M, Pastan I. FCRL1 on chronic lymphocytic leukemia, hairy cell leukemia, and B-cell non-Hodgkin lymphoma as a target of immunotoxins. Blood. 2008;111:338–43. doi:10.1182/blood-2007-07-102350.
44. Holbro T, Hynes NE. ErbB receptors : Directing Key Signaling Networks Throughout Life. Annu Rev Pharmacol Toxicol. 2004;44:195–217. doi:10.1146/annurev.pharmtox.44.101802.121440.
45. Mahmud H, Kornblau SM, ter Elst A, Scherpen FJG, Qiu YH, Coombes KR, et al. Epidermal growth factor receptor is expressed and active in a subset of acute myeloid leukemia. J Hematol Oncol. 2016;9:64. doi:10.1186/s13045-016-0294-x.
46. Jin J, Wang L, Tao Z, Zhang J, Lv F, Cao J, et al. PDGFD induces ibrutinib resistance of diffuse large B-cell lymphoma through activation of EGFR. Mol Med Rep. 2020. doi:10.3892/mmr.2020.11022.
47. Donato E, Biagioni F, Bisso A, Caganova M, Amati B, Campaner S. YAP and TAZ are dispensable for physiological and malignant haematopoiesis. Leukemia. 2018;32:2037–40. doi:10.1038/s41375-018-0111-3.
48. Cottini F, Hideshima T, Xu C, Sattler M, Dori M, Agnelli L, et al. Rescue of Hippo coactivator YAP1 triggers DNA damage–induced apoptosis in hematological cancers. Nat Med. 2014;20:599–606. doi:10.1038/nm.3562.

49. Zhou Y, Huang T, Cheng A, Yu J, Kang W, To K. The TEAD Family and Its Oncogenic Role in Promoting Tumorigenesis. Int J Mol Sci. 2016;17:138. doi:10.3390/ijms17010138.
50. Škrtić A, Korać P, Krišto DR, Ajduković Stojisavljević R, Ivanković D, Dominis M. Immunohistochemical analysis of NOTCH1 and JAGGED1 expression in multiple myeloma and monoclonal gammopathy of undetermined significance. Hum Pathol. 2010;41:1702–10. doi:10.1016/j.humpath.2010.05.002.
51. Hünig T, Beyersdorf N, Kerkau T. CD28 co-stimulation in T-cell homeostasis: a recent perspective. ImmunoTargets Ther. 2015;:111. doi:10.2147/ITT.S61647.
52. Riha P, Rudd CE. CD28 co-signaling in the adaptive immune response. Self Nonself. 2010;1:231–40. doi:10.4161/self.1.3.12968.
53. Schubert D, Bode C, Kenefeck R, Hou TZ, Wing JB, Kennedy A, et al. Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations. Nat Med. 2014;20:1410–6. doi:10.1038/nm.3746.
54. Edward B M. Autoimmunity and Lymphoma: A Brief Review. J Rheum Dis Treat. 2018;4. doi:10.23937/2469-5726/1510062

.

Chapter 4

An integrative omics approach in *in vivo* data towards understanding the effect of persistent environmental pollutants

T.J.M. Kuijpers[1,*]
S.A. Kyrtopoulos[2]
P. Georgiadis[2]
H. Kiviranta[3]
T. Lundh[4]
J.C.S. Kleinjans[1]
D.G.J. Jennen[1]

[1] Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands
[2] National Hellenic Research Foundation, Institute of Chemical Biology, Medicinal Chemistry and Biotechnology, 48 Vas. Constantinou Ave., Athens 11635, Greece
[3] National Institute for Health and Welfare, Department of Health Security, P.O. Box 95, 7071, Kuopio, Finland
[4] Division of Occupational and environmental Medicine, Lund University Hospital, Lund, Sweden

Abstract

Introduction: Persistent organic pollutants (POPs) and heavy metals are toxic compounds with established adverse effects on human health. Current research shows that POPs and heavy metal exposure not only affect gene expression profiles but also epigenetic mechanisms. To investigate the integrative molecular effects of chronic human exposure to these compounds, we aimed to apply a cross-omics computational approach to investigate the relation between exposure and alteration in gene expression and/or DNA methylation in European cohort studies.

Methods: We applied a multi-layer nonnegative matrix factorization method to integrate gene expression and DNA methylation profiles with blood levels of POPs including PCBs, DDE, and HCB, as well as of lead and cadmium, from two prospective cohort studies (EPIC Italy and NSHDS, 610 samples in total). This integration strategy generates different genomic profiles for the obtained clusters that potentially explain exposure-induced molecular effects. We combined those genomic profiles through network construction into genomic interaction networks to investigate their biological relevance.

Results: We identified clusters of subjects within the combined cohort (EPIC and NSHDS cohorts) that show distinct exposure profiles for DDE and lead. These genomic profiles demonstrated the involvement of POP and heavy metal exposure in various biological processes including signaling processes, immune-related processes, DNA repair, mRNA translation, and mRNA transcription. We identified interactions between PCBs and these processes, although we could not determine a single PCB responsible for the alterations. The simulation on the NSHDS cohort revealed a potential sex-specific exposure-response of male gamete formation in the Swedish males, as well as a sex-specific exposure-response of cytochrome P450 genes and bile acid synthesis in the Swedish females. In the EPIC cohort, we identified methylation differences for sex-specific CpG sites associated with lead exposure, which might point towards a sex-specific exposure-related effect. These CpG sites are associated with mantle cell lymphoma and thus could connect lead exposure with the onset of this disease.

Conclusion: This study provides insights into exposure-related changes in the transcriptome and epigenome. By combining the EPIC and NSHDS cohorts, we were able to identify communities in our interaction network that relate to POP and lead exposure that could explain the development of certain diseases. In general, we could not identify clear epigenetic alterations that directly associate with observed changes in the transcriptome. However, the combination of the epigenetic and transcriptomic alterations provided us with genomic profiles that relate to the exposure profiles. This shows that we can use both omics layers to cluster samples into distinct groups that have a mixture of exposures.

Supplementary data available at: https://github.com/TJMKuijpers/PhDThesis

Introduction

Persistent organic pollutants (POPs) are toxic chemicals that have an adverse effect on human health and the environment. The most commonly encountered POPs belong to the class of organochlorine pesticides, including polychlorinated biphenyls (PCB), dichlorodiphenyltrichloroethane (DDT), as well as by-products of many industrial processes such as dioxins. Other non-degradable pollutants, the heavy metals lead and cadmium, also play an important role in exposure-related diseases. Chronic POPs and heavy metal exposure are becoming increasingly relevant to human health, because they reflect the exposure of the general population, mainly via food from contaminated animal fats [1–3] or via inhalation of lead-contaminated dust particles [4].

PCBs are biodegradation-resistant and can induce acute and chronic health disorders, depending on the dose, duration of exposure, type of PCB, and degree of chlorination (dioxin-like or non-dioxin-like) [2]. Dioxin-like PCBs show toxic effects similar to dioxins [5] and mediate aryl hydrocarbon receptor (AhR) activation [6]. Non-dioxin-like PCBs mediate the constitutive androstane receptor (CAR) and pregnane X receptor (PXR) in mammalian cells [7–11] but are also believed to either activate or suppress genes regulated by the thyroid hormone, interact with the thyroid hormone receptor [12] and modulate gene expression dependent on estrogen receptors (ER) [13]. The ability of PCBs to affect these different receptors may also explain the observed sex-specific responses [14–17]. The International Agency for Research on Cancer (IARC) reported that there is sufficient evidence in humans for the carcinogenicity of certain PCBs (IARC group 1) [1, 18]. In humans, PCBs have been consistently linked to increased risk of non-Hodgkin lymphoma [19–22] and acute lymphocytic leukemia [23]. Not only has PCB exposure been related to adverse effects on the immune system [24, 25] but also influences the immune system development of unborn children [26].

Besides interacting with multiple receptors and directly altering gene expression, PCB exposure can also indirectly affect gene expression by changing the methylation status of the DNA. Georgiadis et al discovered a distinct DNA methylation profile in blood leukocytes linked to PCB exposure [27]. The link between PCB exposure and DNA methylation changes has also been found in different cohorts [28–30], increasing the evidence that suggests a link between PCB exposure and DNA methylation effects. However, there is also evidence that suggests a lack of persistence in DNA methylation changes induced by PCB exposure [31].

Whereas PCBs are banned 35 years ago [32], the use of DDT (probably carcinogen to humans, group 2A [33]) was already restricted or banned in most developed countries after 1970 [34]. However, due to the chemical stability of DDT and its lipophilic character, it is only slowly eliminated from the environment and most living creatures [35]. DDT and its main metabolite dichlorodiphenyldichloroethylene

(DDE) may have adverse effects on different organs including the liver and kidney, but also the reproductive [36], endocrine [37, 38], and immune system [39]. Like PCBs, DDT and DDE affect certain receptors [38] including the ER [40, 41]. DDE may exert anti-androgenic effects by interfering with the androgen receptor (AR) [42]. DDT is linked to inducing epigenetic transgenerational inheritance of obesity, kidney, testis, and ovary disease [43]. However, the effects of DDT and DDE on non-inheritance epigenetic alterations possibly leading to disease development are still unclear.

Not only is the human population exposed to POPs, but exposure to heavy metals such as lead and cadmium also has a major impact on human health [4, 44–46]. Lead has been classified by IARC as a possible carcinogen (IARC group 2B [47]), based on the relationship between lead and cancer of the stomach, brain, kidney, and lung [48]. Current findings also associate chronic lead exposure to anemia, increased blood pressure, and severe damage to the brain and kidneys [49]. In males, it has even been found to reduce fertility [50]. Lead exposure may also facilitate the carcinogenic effects of other pollutants, by impairing DNA repair in the cell [51]. Several studies have shown that reactive oxygen species (ROS) production and oxidative stress play a key role in the toxicity and carcinogenicity of lead. With respect to the epigenome, lead exposure is associated with alterations in whole-blood methylation. Furthermore, developmental lead exposure is associated with changes in epigenetic regulators mediating the development of Alzheimer's [52].

Cadmium, another toxic metal, is present in various sources including food, air, and water. IARC classified cadmium as a group 1 carcinogen with sufficient evidence for cancers in humans [53]. Several studies indicate a possible role for cadmium in lung [54] and renal [55, 56] cancer. Cadmium exposure is believed to induce oxidative stress, DNA damage, alterations in DNA repair, enhanced proliferation, and downregulated apoptosis [57, 58]. Furthermore, there is evidence showing a link between alterations in DNA methylation and cadmium exposure [59].

Overall, our current understanding of the relation between exposure to complex toxic mixtures comprising agents such as POPs, lead and cadmium, with respect to inducing gene expression, and DNA methylation is not complete. Such exposures may not only affect gene expression profiles [16] while at the same time altering epigenetic mechanisms [60–62] but the underlying mechanisms are still undefined. Therefore, we here aim to apply a multi-layer nonnegative matrix factorization (NMF) approach to investigate the combined effect of multi-compound exposure and alterations in gene expression and DNA methylation. We hypothesize that with an integrative omics strategy, we will be able to stratify subjects into groups that reflect their exposure status and gain insight into their genomic profiles.

Methods

In this study, we use two cohorts, one from the Northern Sweden Health and Disease Study (NSHDS) and the other from the European Investigation into Cancer and Nutrition Study (EPIC). To increase the sample size and thus the power of the data set, we combine the two cohorts into one large multi-omics data set.

Transcriptomics and Epigenetics

The data processing procedures used for DNA methylation and gene expression profiling are described in detail by Georgiadis et al [63]. Gene expression data is derived from blood samples collected from the Northern Sweden Health and Disease Study (NSHDS) and the European Investigation into Cancer and Nutrition (EPIC) study. Samples were hybridized on Agilent 4x44K human whole-genome microarrays for gene expression analysis. The data were normalized using the quantile method from the Limma package as described by Espín-Pérez et al [17]. Methylation data [27, 64] of the CpG sites has been derived from hybridizing samples on Illumina Infinium human 450K methylation arrays as described by Georgiadis et al [27]. For each CpG site probe, the methylation status is expressed as the β-value in the range of 0 (hypomethylated) or 1 (hypermethylated) to a nonnegative data distribution [65]. We selected those subjects from either the NSHDS or EPIC cohort if both the DNA methylation and gene expression measurements are available.

POPs, Cadmium, and Lead exposure

Plasma POP concentrations were measured as described by Kelly et al [66]. An Agilent 6890 gas chromatograph connected to a Waters Autospec ultima high-resolution mass spectrometer has quantified POPs. For quality control purposes, two reagent blanks were added and the average results of the blank samples were subtracted from the results of the cohort samples. Furthermore, two control samples of Standard Reference Material 1589 from the National Institute of Standards and Technology were included in each batch. Lead and Cadmium levels were determined by inductively coupled plasma-mass spectrometry [67] at Lund University Hospital. The analytical accuracy was checked against human blood reference material from the Centre de Toxicologie du Quebec [66].

Input data filtering

To improve the performance of our workflow, we applied a data-filtering step to reduce the effect of confounding variables. Since we were interested in finding a relation between transcriptional interactions, gene-gene interactions, and protein-protein interactions, we selected only the protein-coding genes from the total transcriptomic data set. Second, we performed a Univariate feature selection to reduce the number of probes that did not relate to exposure. Therefore, we used an ANOVA test to identify which genes showed a cohort-specific or sex-specific effect, and a regression classifier to classify genes related to POPs exposure or age. If genes explain the cohort and POPs exposure, we did not remove the genes

from the data set. We only removed genes if they explicitly explained the difference in the cohort. Since we assumed no large biological effect of CpG sites downstream from the transcription starting site, we only kept DNA methylation data around the promoter region.

Multi-layer Nonnegative Matrix Factorization

Nonnegative matrix factorization decomposes a data matrix to two lower based matrices, containing the cluster coefficients H and the basis coefficients W [68]. For an n multi-layer dataset, the data matrix X_n is decomposed to n matrices W_i and one matrix H (Equation 1). To find a local optimal solution (Equation 4), matrices W_i and H are updated by their update rules (Equations 2 and 3 respectively). For H we took into account the effect of the different omics layers via X_i and W_i whereas for W_i we took into account the effect of the omics layers via X_i and the sample clustering via H.

$$\sum_{i=1}^{n} X_i \approx \sum_{i=1}^{n} W_i H \qquad (1)$$

$$W_{w+1} = W * \frac{\frac{X_i}{W_i H} H^T}{\sum H} \qquad (2)$$

$$H_{H+1} = H * \frac{\sum \frac{X_i}{W_{iH}}^T}{\sum W_i^T} \qquad (3)$$

$$KL\ divergence = \sum \left(X_i * \log \left(\frac{X_i}{W_i H} - X_i + W_i H \right) \right) \qquad (4)$$

To extract feature genes and CpG sites, the corresponding matrix W_i was scored by using the method proposed by Kim et al [69]. For each cluster, we selected the biological entities as features, if those entities that had a high score for the selected cluster and a low score for the other.

Genomic interaction network

To study the relationship between the transcriptome and epigenome, we constructed a genomic interaction network from the obtained feature genes and CpG sites. This network will contain the feature genes and CpG sites to understand how different exposures change DNA methylation or gene expression. A feature gene will be represented in the network as a node and interactions between nodes are derived from OmniPath [70] (for gene-gene interactions), the Comparative Toxicogenomics Database [71] (for compound – gene interactions), STRINGdb [72], transcription catalog from Souza et al [73] (transcriptional activation or inhibition of genes), and CpG – gene interactions.

The Louvain method [74] was used to detect communities in the genomic interaction network. Communities are groups of nodes, which share a high degree of interactions with each other compared to nodes in other communities. Initially, a node is placed inside its community and the algorithm will merge small communities; this will increase the modularity of a partition of the network. Low modularity [-0.5] indicates non-modular clustering, whereas maximum modularity (1) indicates a fully modular clustering. These communities are mapped against the Reactome database through the PantherDB [75] interface to identify overrepresented pathways. We combined pathway information with the network information to identify important HUB genes involved in biological processes.

Results

Case study on the combined EPIC + NSHDS cohorts:

The integration of the two different cohorts based on the combined omics data resulted in three clusters (Figure 4.1A). The three identified clusters represent the cohorts (Cluster 1 EPIC, and Cluster 2 and 3 NSHDS) but two clusters also show a sex effect (Cluster 1 mix, cluster 2 male and cluster 3 female). When we look at the exposure levels, a significant difference appears to exist between the three clusters for DDE, lead, and cadmium exposure (Figure 4.1C). Therefore, DDE and lead exposure could play an important role in the classification of the three clusters, especially since the exposure profiles of the PCBs (Figure 4.1B) are not significantly different between clusters. To identify which genomic features are responsible for the difference between the clusters, feature scoring and extraction has been performed on the feature matrix W_i. This results in a set of genes and CpG sites explaining the formatting of the three clusters, which define the genomic profile of each cluster. First, we will analyze these genomic profiles to investigate which genes and CpG sites could present a potential marker for the exposure profiles. Second, we will use those markers to construct a genomic interaction network to unravel the relationships between the different gene and CpG markers.

There are many genes higher expressed in cluster 1 in comparison with cluster 2 and cluster 3 (Figure 4.2A, only showing the top 50 genes). To get a general idea of which pathways these genes are involved in, we mapped the genes against PantherDB. We have identified different pathways overrepresented including positive regulation of endocytosis (FDR 2.57E-02), transmembrane receptor protein tyrosine kinase signaling pathway (FDR 1.64E-02), regulation of intracellular signal transduction (FDR 3.16E-02), positive regulation of nitrogen compound metabolic process (FDR 1.95E-02), and nervous system development (FDR 8.99 E-02).

For the epigenome layer, we have identified several CpG sites showing different methylation patterns. As shown in figure 4.2B, we have identified CpG sites with a strong methylation difference over the three clusters (hypo vs hypermethylated), as well as CpG sites with smaller differences (range of $\Delta\beta$ of 0.1). It is of interest to

note a specific methylation pattern for cluster 2, with some high methylated CpG sites (Figure 4.2B, red color) that could be sex-specific (cluster 2: male), although these patterns are not observed for the EPIC males. Identification of the chromosome location of those CpG sites revealed that the hypermethylated genes in cluster 2 are on the X chromosome. For the genes linked to the CpG sites, we could not identify any overrepresented pathways, which indicates that various CpG sites of genes with different GO biological processes are targeted

Because we observe a strong contrast in DDE and lead exposure between the EPIC and NSHDS cohorts, with lower DDE and lead exposure in the NSHDS cohort, we hypothesized that combining both cohorts would hamper the possibility to identify cohort-specific exposure patterns. Therefore, we have performed additional analysis to investigate a cohort-specific exposure profile. This will allow us to investigate relevant biological patterns present within a cohort data set. However, we should keep in mind that this approach will decrease the sample size and might have an impact on the power to detect expression patterns within the data.

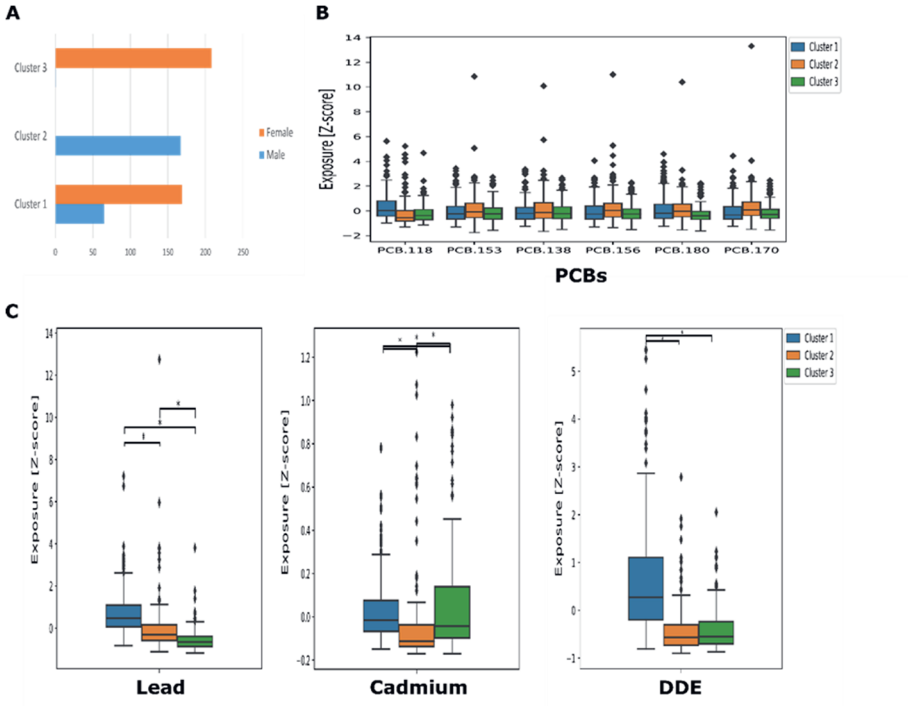


Figure 4.1 Cluster results combined cohort. **A**: Distribution of male and female subjects in each cluster, with cluster 1 representing the EPIC subjects and cluster 2 and 3 the NSHDS subjects. **B**: Distribution of PCB exposure (Z-score) for the three clusters. **C**: Lead, cadmium and DDE exposure (Z-score) profiles with significant mean differences between the three clusters.

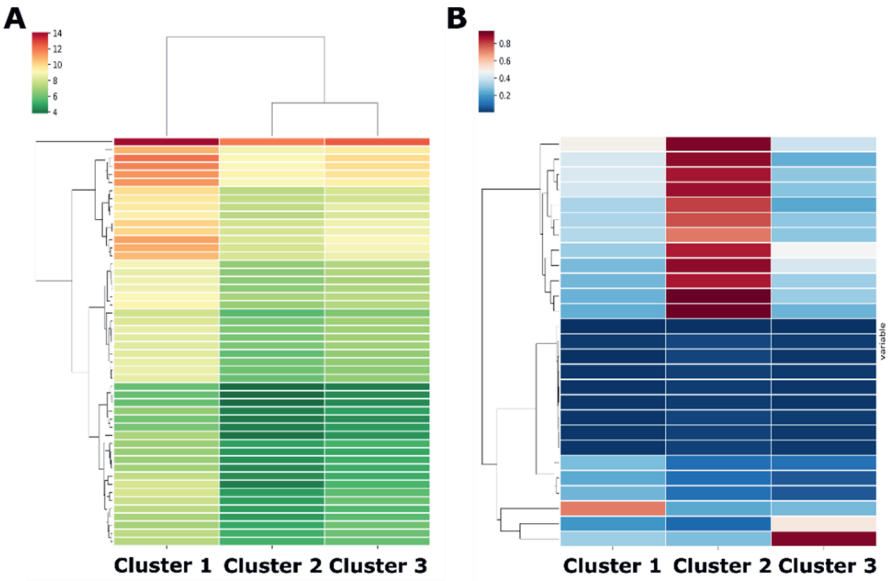


Figure 4.2 Genomic profiles in the combined cohort analysis: **A** Gene expression profiles of the top 50 feature genes (High expression dark red, low expression dark green). **B**: CpG profiles of the top 50 CpG sites (hypermethylation dark red, hypomethylation dark blue).

Case study EPIC cohort

We have performed an additional simulation to combine the transcriptome and the epigenome for the EPIC cohort. Here, we identified two clusters, one cluster with only females (cluster 1) and one cluster containing a mixture of males and females (Figure 4.3A). Cluster 2 shows a significantly different lead exposure profile, whereas the POP and cadmium exposure levels are equal for both clusters (Figure 4.3C). These two clusters might indicate that there are small within-cohort differences in the genomic profiles, which we could not identify in the combined cohort analysis.

The transcriptome profile (Figure 4.4A) only shows some minor differences in gene expression between the clusters. In general, there is a small but significant increase in gene expression in cluster 2, which could be because of a different exposure profile. There is only a small subset of genes with a higher mean expression, including *KDM5D, EIF1A*Y, and *DDX3Y*. These results might indicate the low importance of the transcriptome layer to stratify the EPIC cohort.

The epigenome layer (Figure 4.4B) does indicate a stronger effect with a clear difference in several methylated CpG sites in cluster 2. These results are interesting since we could not identify them in the combined cohort simulation. Those 27 hypermethylated CpG sites belong to different genes: *PPP1R2P9* (4 sites), *FAM47B* (3 sites), *CXorf61* (3 sites), *LUZP4* (2 sites), *TEX11* (2 sites), *DDX53* (2 sites), *DDX53* (2 sites), *FAM47A, MAGEB1, TDGF3, TFHL17, MAMLD1, XIST, ZCCHC13, MAGEB2, FAM9C, GUCY2F*, and *MGC16121* (all 1 site). These CpG

sites are all located on the X chromosome and could indicate that lead might affect the X chromosome. In cluster 2, most of the subjects are male (64 out of 88), whereas in cluster 1 only female subjects are present. On the X chromosome, the CpG sites are located within the TSS200 region. This might indicate a stronger effect due to POP and heavy metal exposure on one of the sex chromosomes. However, we observed no direct change in gene expression for the different genes, which could be for two reasons: the CpG site methylation is not covering the whole TSS200 region and the transcription factor could still bind or the transcription factor binds to the TSS150 region.

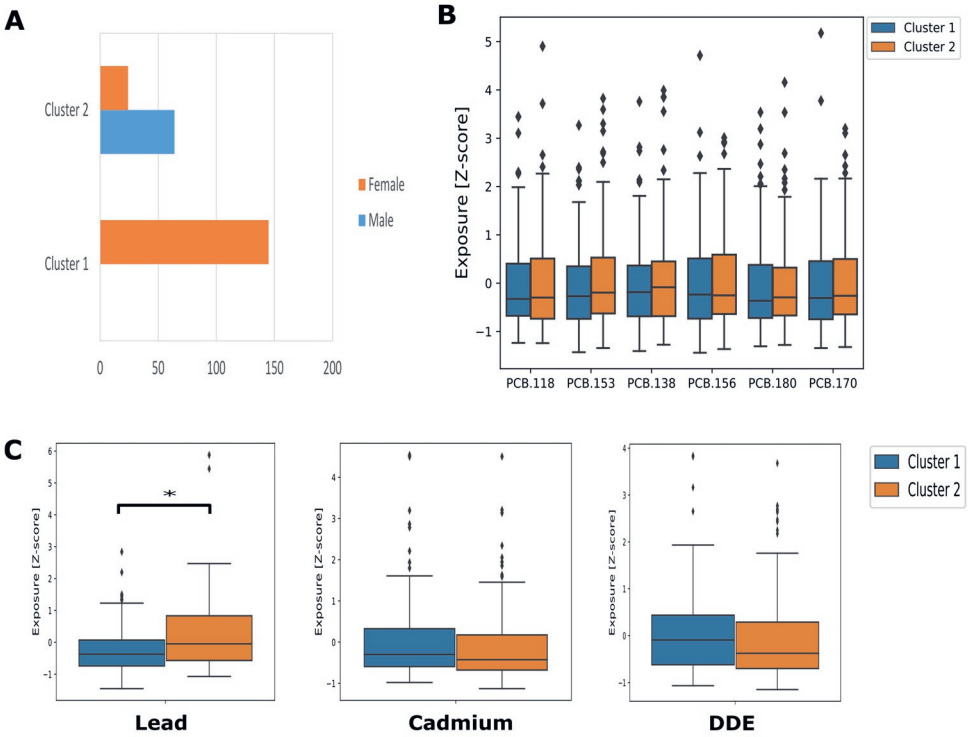


Figure 4.3 Cluster results EPIC cohort**. A:** Distribution of male and female subjects in each cluster, with a strong female component in cluster 1. **B:** Distribution of PCB exposure (Z-score) for the two clusters. **C**: Lead, cadmium and DDE exposure (Z-score) profiles with significant mean differences between the two clusters.

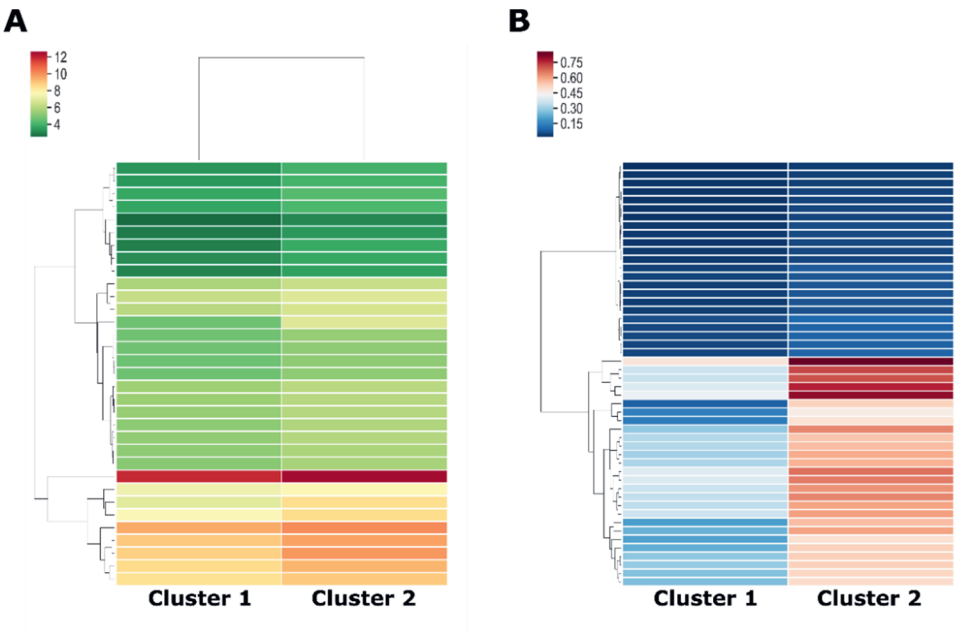


Figure 4.4 Genomic profiles in the EPIC cohort analysis**. A:** Gene expression profiles of the top 50 feature genes. **B:** CpG profiles of the top 50 CpG sites. (Hypermethylation dark red, hypomethylation dark blue.

Case study: NSHDS cohort

As previously indicated for the EPIC cohort, we would also like to identify a cohort-specific exposure effect for the NSHDS cohort. The omics integration for the NSHDS cohort resulted in three clusters (Figure 4.5) of which two clusters contain predominantly females (cluster 1 and cluster 2) and one cluster containing only males (cluster 3). As in agreement with the combined cohort results, the males (cluster 3) have a higher exposure to lead compared to females. These new clusters are of interest because it now becomes possible to identify whether the different transcripts and CpG sites are sex-specific induced (cluster 3 is different from cluster 1 and 3) or are exposure-related (if a CpG site has a different methylation status in cluster 1 in comparison to cluster 2, we can assume it is not sex-specific).

Here we have identified a set of transcripts and CpG genes, which show different patterns between the three clusters (Figure 4.6). On the epigenome, we see two main patterns: 1) cluster 2 (females shows hypomethylation for a set of CpG sites that are methylated in cluster 1 (females) and cluster 3 (males, increased lead exposure), and 2) Cluster 3 (males, increased lead exposure) show strong hypermethylation for CpG sites that are hypomethylated in cluster 1 (females) and cluster 2 (females). This could mean a sex-specific effect of exposure, in particular lead exposure, in the males and females of the NSHDS cohort.

The genes and CpG sites that formed the clusters carry out different biological functions. The feature transcripts carry out roles in various pathways including autonomic nervous system development (FDR 1.58E-02), regulation of glial cell differentiation (3.63E-02), G protein-coupled receptor signaling pathway (4.53E-02), and leukocyte activation (4.11E-02). The genes associated with the CpG sites show a significant overrepresentation of the male gamete generation (FDR 4.23E-02). This pathway can be of interest because of the impact on developmental health or the reproductive system. The CpG sites involved in the male gamete regions are located at various chromosomes, where the most are located at the X chromosome (18 CpG sites), other sites are found on chromosome 1 (7 CpG sites), Chromosome 2 and 3 (2 CpG sites), and chromosome 8 (1 CpG site). The high number of CpG sites at chromosome X can be of interest because this might point towards a sex-specific chromosomal effect.

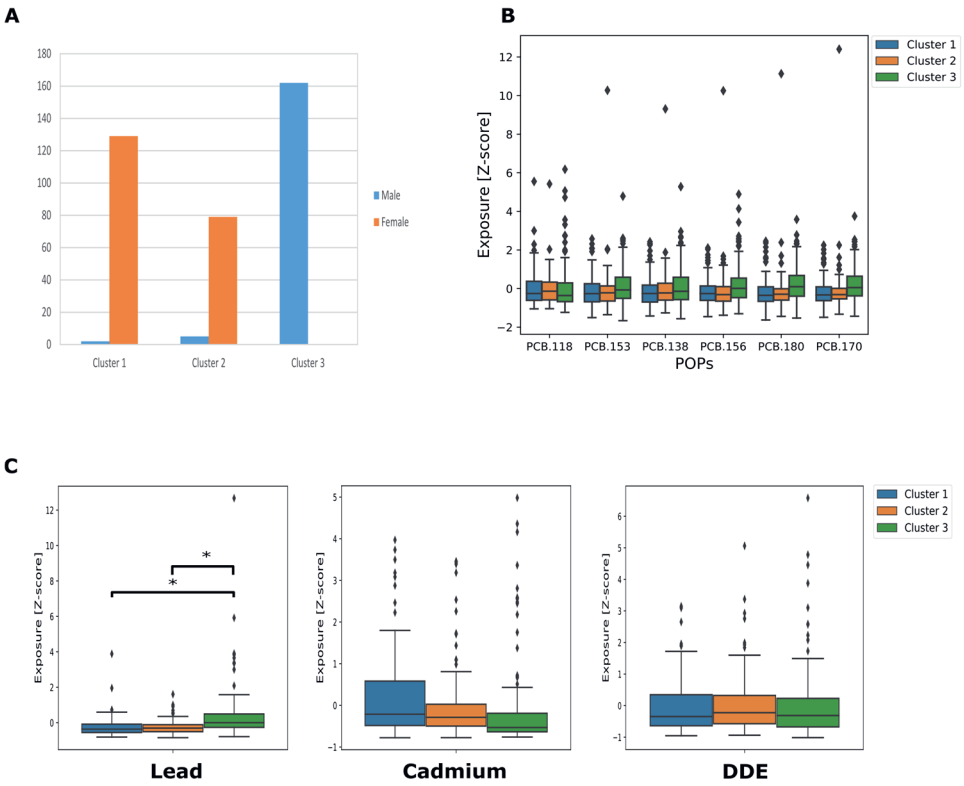


Figure 4.5 Cluster results from NSHDS cohort**. A:** Distribution of male and female subjects in each cluster, with a strong female component in cluster 1 and cluster2, and a strong male component in cluster 3. **B:** Distribution of PCB exposure (Z-score) for the three clusters. **C**: Lead, cadmium, and DDE exposure (Z-score) profiles with significant mean differences between the three clusters for lead.

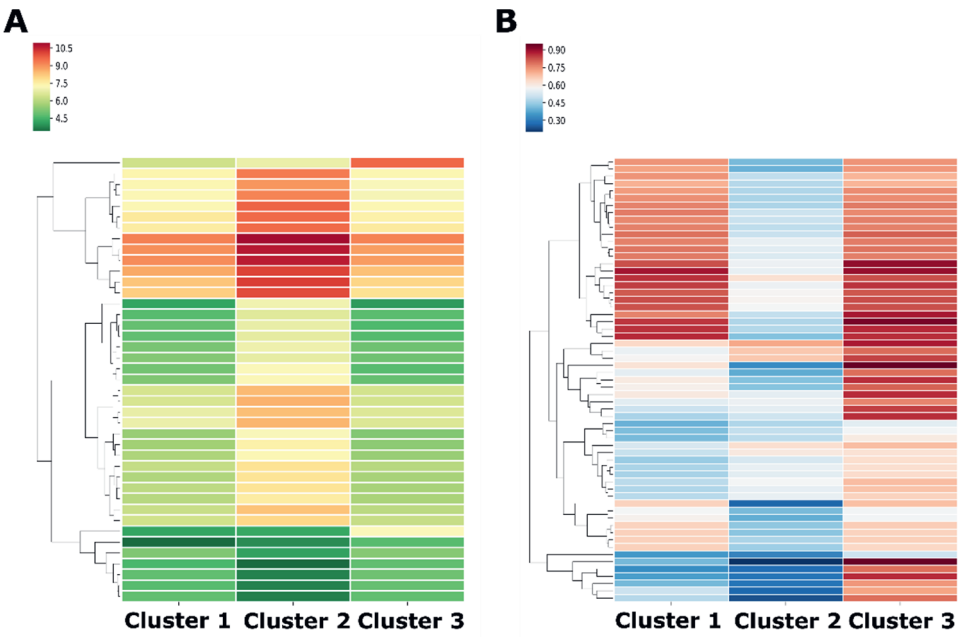


Figure 4.6 Genomic profiles in the NSHDS cohort analysis**. A:** Gene expression profiles of the top 50 feature genes. **B**: CpG profiles of the top 50 CpG sites.

Translation of the genomic profiles to exposure profiles

The previous analyses showed that there is a difference in lead and DDE exposure between the three clusters of the combined cohort data. It is of particular interest to identify which feature genes and CpG sites are not an effect of the mixture of compound exposures but due to a single compound. This would give us more insight into the adverse effect of a compound and could provide us with new biomarkers. Therefore, we performed one extra step before we combined the features into a genomic interaction network. We used a random forest regression model to determine if the gene expression and DNA methylation levels relate to individual compound-exposure levels. With this approach, we could identify within the feature set of each analysis those genes and CpG sites that relate to lead and DDE exposure (Supplementary file 4.1: “Lead and DDE associated features”).

Therefore, we studied the different relationships between the feature transcripts and CpG sites by the creation of genomic interaction networks. Here, we will exclusively study the effects from the combined cohort case study, since the genomic profiles gave stronger signals on both the transcriptome and epigenome layers. To study the relation between the different feature genes and CpG sites, we have combined the features related to DDE and lead exposure. We did not include cadmium because of the low exposure levels in comparison with the other compounds, and we assumed that the exposure effect of DDE and lead would mask the effect of cadmium. The genomic interaction stores transcriptional interactions (activation or inhibition), gene-gene interactions, protein-protein interactions, and

compound-gene interactions. This resulted in a large network in which we detected communities of high-connected transcripts and CpG sites. We used these communities to get a better understanding of the role of POPs and heavy metals in different biological processes.

The communities in the genomic interaction network are associated with various biological pathways (Supplementary table 4.1) and can be grouped by classifying them as signaling processes, immune-related processes, DNA repair processes, mRNA processing processes, and translational processes (Figure 4.7). Here, it becomes clear that lead exposure can be associated with all five categories, whereas the individual POPs are associated with just a subset of the five groups. It is of interest that lead is affecting mRNA transcriptional and translation processes, and thereby does not only influence gene expression (transcription) but also can have an outcome on protein expression (translation). The role of lead in immune-related processes and signaling-related processes is of interest, especially since communities of those two groups also interact with each other. Therefore, a negative effect in either one of the groups can propagate to the other group. Here, we also have to take the role of POPs into account, since they can also exert an effect on the biological processes already altered by lead exposure. For example, DDE, PCB 118, and lead, all target mRNA transcriptional processes and it is vital to further study if this leads to a combined negative effect.

In the genomic interaction network, it became clear that lead interacts with DNA repair genes RAD9A, EXO1, and RBBP8. These genes respond to the detection of DNA damage and as a result stop or reduce the cell cycle rate. As one would expect, the higher lead exposure in cluster 1 (Figure 4.1C) resulted in a higher expression of those DNA repair genes. This might be a direct effect of lead exposure or because of the interaction of *RBBP8* with the signaling-related gene *TFDP1*. As a potential effect, we observed an increase in the signaling gene *TFDP1* and the DNA repair gene *RBBP8*. TFDP1 is a DNA-binding transcription factor part of the NOTCH signaling pathway, involved in the transcription of RNA polymerase 2, and associated with cellular response to stress. This could be a direct effect of the increased lead and DDE exposure in cluster 1 (Figure 4.1C).
We found two communities with genes related to mRNA translational processes, community 9, and community 12. The genes of community 9 are involved in post-translational protein modification and neddylation. The genes of this community, *UBA1, DCUN1D1, UBE2E1, ASB16*, and *FBXl7* showed an increased expression, whereas *RNF7* showed a decreased expression. Within the genomic interaction network, we found that different PCBs (PCB 153, PCB 138, PCB 180) and lead interact with four genes (*RNF7, UBE2E1, NEDD4L,* and *UBA1*) of community 9 and could be potential targets to study lead and PCB exposure effects.

Community 12 consists of multiple methylated DNA regions associated with mRNA translation. Here, we observe an increase in methylation values (thus more hypermethylation) for cluster 3 (NSHDS cohort) compared to cluster 1 (EPIC cohort, increased lead and DDE exposure) and cluster 2 (NSHDS cohort, only males, increased lead exposure). This could be due to higher lead exposure in both clusters 1 and 2 and therefore might indicate that lead does induce changes in mRNA translation. However, the genes associated with the CpG do show a different behavior as one would expect, because we observed a decreased expression for *LAS1L* and *RPL36A* in cluster 1. Our genomic interaction network shows that lead does interact with *LAS1L* and *RPL36A* and this could alter gene expression even if the CpG sites are hypomethylated. However, we should also take into account that the methylation value of a single CpG site of a given gene does not need to be directly related to the expression of this very gene. Furthermore, the interaction of *DDX21* with *LAS1L* may be important since *DDX21* plays a role in translation initiation [76].

PCB exposure might play an important role in changing signaling pathways since PCB 138, 170, 180, and 158 interact with genes of the c-MET signaling pathway (Figure 4.7 community 4). One key member of the c-Met pathway is c-MET, a receptor tyrosine kinase regulating essential cellular processes. PCB 138, 170, 180, and 158, directly influences the gene *MET* (Figure 4.8A), encoding for c-MET, which could have a downstream effect on the protein-protein interactions with *ARF6, LAMA4, LAMC2, PTPN11*, and *PIK3R1* or the gene-gene interactions with *FAS, SHC1*, and *WASL*. Our results show a higher expression of *MET, PTPN11, WASL, LAMC2*, and *LAMA4* in cluster 1 (EPIC cohort, increased lead and DDE exposure), where MET expression could be higher because of the higher expression of its transcriptional activator *STAT5A*. PCB 157 interacts with genes and CpG sites involved in the regulation of the TNFR1 signaling pathway (*XIAP, TAB3, TNFAIP3,* and *SHARPIN*) and the innate immune system (*TAB3, BCL2L1, TNFAIP3, IL1B, RELA, SFTPA1,* and *TREM1*).

A higher lead and DDE exposure is associated with elevated expression levels of both *FOXO3* and *FOXO4*, two members of the Forkhead box O transcription factor family (Figure 4.7 community 2, Figure 4.8B local network). Lead is known to interact with *FOXO3* and affects *FOXO3* mRNA expression [77]. *FOXO3* can transcriptionally activate *STK11*, *TSC22D3*, and transcriptional inhibit *VEGFA*. The community around *FOXO3* (Figure 4.8B) shows that *FOXO3* and *FOXO4* share a protein-protein interaction but also that *EP300* and *RELA* play a role in transcriptional activation *VEGFA*. It becomes clear that although *FOXO3* expression increased in cluster 1, it does not result in transcriptional inhibition of *VEGFA*. On the contrary, we observe an increase in *VEGFA* expression in cluster 1 (EPIC cohort, increased lead and DDE exposure). When we studied the interaction, it becomes clear that lead interacts with both *FOXO3* and *VEGFA*. This

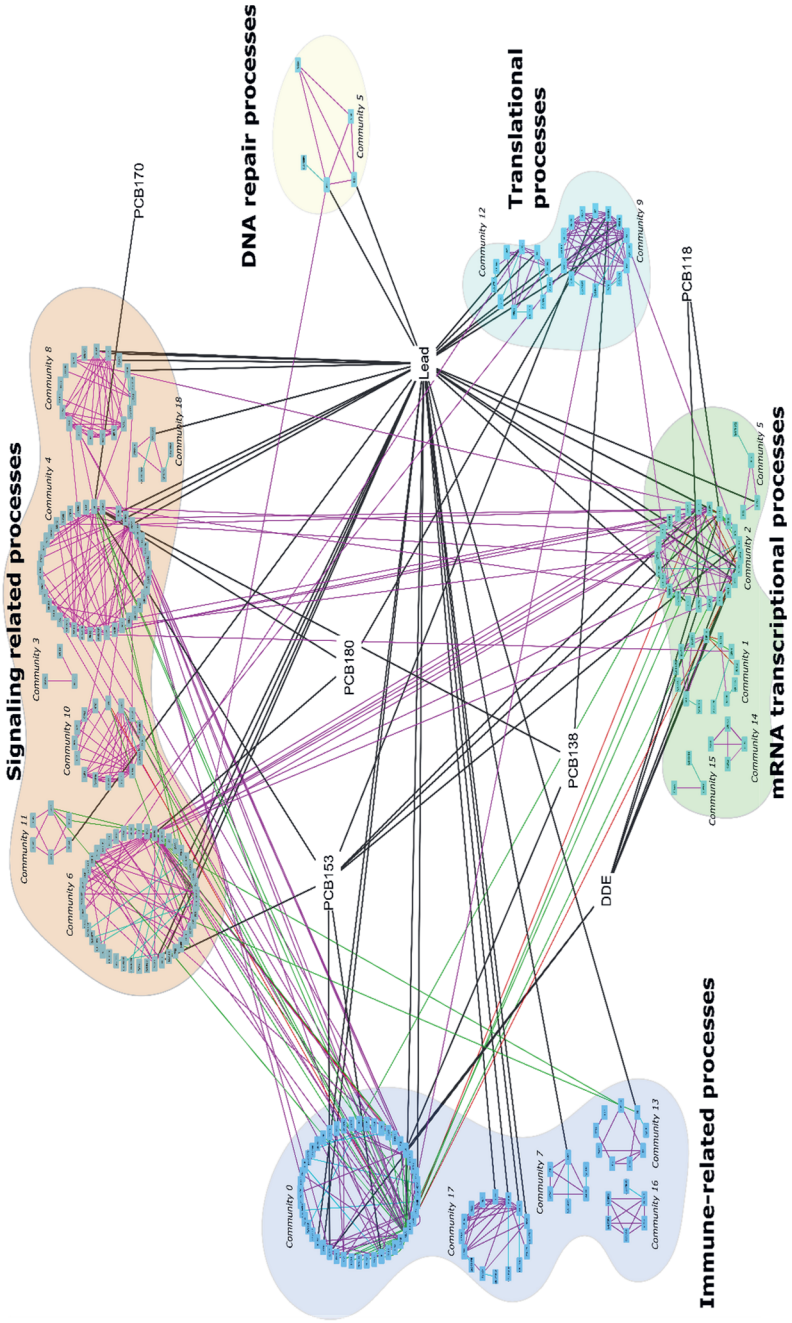


Figure 4.7 Genomic interaction network: Communities grouped by five classes related to immune processes, signal processes, DNA repair, Translational and transcriptional processes. Edge colors represent the type of interaction: transcriptional activation (green), transcriptional inhibition (red), CpG-gene (blue), compound – gene (black), and protein-protein (pink).

might point towards interference of lead in the transcriptional inhibition of *VEGFA* by *FOXO3*.

The interaction between *VEGFA* and its transcriptional activator *RELA* could be another reason why we observed an increase in *VEGFA* expression. Our genomic interaction network showed that *RELA* could be transcriptionally inhibited by *TSC22D3* but in this case, we see an increased expression in cluster 1 (EPIC cohort, higher lead and DDE exposure). The Transcription start sites of *VEGFA* are all hypomethylated in cluster 1 and thus transcription factors (activator or inhibitor) could bind. This observed disruption of transcriptional inhibition of *VEGFA* might indicate that lead exposure does disrupt the interaction between a transcription factor and its target. The previously mentioned interaction between *EP300* and *FOXO3* is of particular interest since *EP300* can acetylate *FOXO3* and consequently might lead to a higher *FOXO3* expression. Thus, as our interaction network showed, the increase in *RELA* expression can lead to more transcriptional activation of *EP300* and consequently can lead to higher *FOXO3* expression.

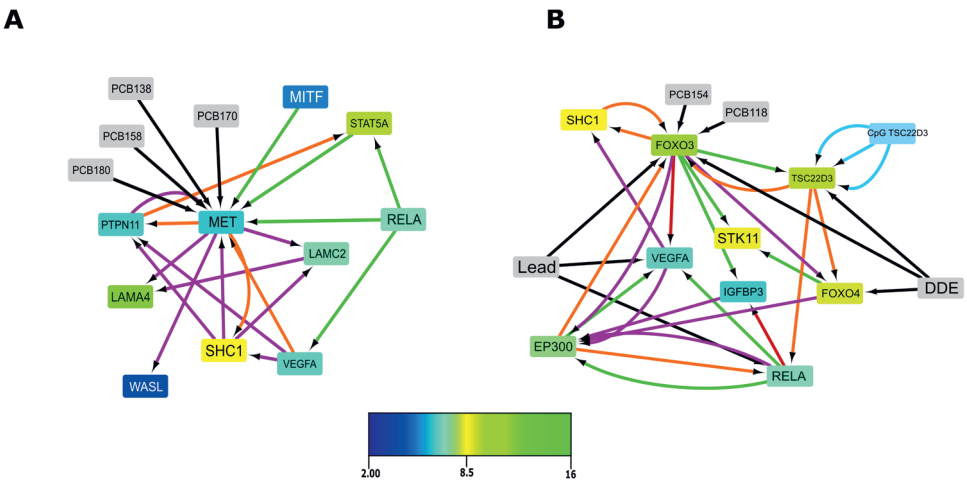


Figure 4.8 Genomic interaction communities. **A**: Genomic interaction network around the genes involved in MET signaling pathway. **B**: Genomic interaction network around the genes involved in FOXO transcription-mediated processes. Edge colors represent the interaction: gene-gene (orange), transcriptional activation (green), transcriptional inhibition (red), CpG-Gene (light blue), and protein-protein (pink).

Discussion

To understand the effect of chronic exposure to POPs and heavy metals on the transcriptome and epigenome, we have performed a multi-omics workflow to cluster our cohort populations based on their genomic profiles. These genomic profiles are then used to construct genomic interaction networks. This genomic interaction network can help us to understand the biological pathways targeted by

the mixture of different compounds and can deepen our knowledge of the specific effects of POPs and lead exposure.

In our results, we found genomic profiles for three clusters within the cohort data that relate to a specific exposure profile. Before we analyzed these exposure-related profiles, we first determined if there is a cohort-specific effect present in our data, as well as a sex-specific effect. In the combined cohort simulation, we found one cluster for the EPIC cohort and 2 clusters for the NSHDS cohort. Therefore, we have performed two additional simulations on each cohort because we hypothesized that the combined cohort data could mask the possibility to identify cohort-specific exposure patterns. The results of each cohort showed no cohort-specific exposure effects and we conclude that the identified profiles in the combined cohort simulation are indeed due to a higher exposure of lead and DDE of the EPIC subjects. To assess if sex-specific exposure effects are relevant, we performed an additional simulation, in which we have removed all sex-specific genes and CpG sites (data not shown). This simulation did result in clusters that do not show any difference in exposure and therefore we conclude that POP and heavy metal exposure induce a sex-specific response. This is in concordance with literature, where sex-specific effects of POPs have been previously reported with relation to DNA methylation [27], developmental effects [78] as well as male reproductive health [79].

Based on the two previous conclusions, we analyzed the results of the combined cohort simulation, since the genomic profiles gave stronger signals on both the transcriptome and epigenome layers. The obtained genomic profiles are of interest because the difference in lead and DDE exposure within a cluster could explain the changes in gene expression and DNA methylation. The genomic interaction network, build from genomic features that explain the clusters, showed different interesting communities, which we could relate to the compounds by integrating the interactions from CTD. Network analysis revealed that several communities belong to biological processes associated with signaling, DNA repair, immune system, mRNA transcriptional, and translational processes. It is of interest to note that these communities do not only interact with communities from the same biological processes but also with communities in other processes. This might indicate that exposure-related alterations in one biological process could have a downstream effect in another process.

The different signaling-related communities are of interest because they play an important role in disease development. Our results show that POP exposure might affect the c-MET signaling pathway. The c-MET signaling pathway is of interest because of its various role in cellular processes, including proliferation, survival, motility, and invasion [80]. As a proto-oncogene, abnormal activation of c-Met can promote the development and progression of various cancers such as breast, lung, liver, and glioblastoma [81, 82]. The increased *MET* expression is accompanied by

increased *LAMC2* and *LAMA4* expression, two genes also involved in the immune response [83]. *LAMC2* is an important factor driving tumorigenesis through its interactions with cell-surface receptors [84]. The identified community of *LAMC2, MET*, and *LAMA4* could therefore be of interest for future studies concerning compound exposure profiles and tumor development. The signaling-related communities do interact with communities related to different processes, such as DNA repair processes. Here, we identified that an increased expression in the signaling gene *TFDP1* results in an increased expression of *RBBP8*, involved in DNA repair and the cell cycle. Overexpression of *RBBP8* activates DNA damage checkpoints leading to DNA damage, suppressing DNA replication at S and G2 phase [85], and could be an important gene altered by lead and DDE exposure.

Transcriptional processes are vital processes in all organisms, and we identified a possible relationship between the increase in certain FOXO transcription factors (*FOXO3, FOXO4*, and *STK11*) and an increased lead or DDE exposure. The FOXO transcription factors, actors in the FOXO-mediated transcription, are active transcription regulators of several processes, including development, differentiation, proliferation, DNA repair, survival, and apoptosis [86–88]. Due to their prominent role in these processes, FOX transcription factors, and the FOX proteins, play a crucial role in the onset of diseases such as acute myeloid lymphoma [88]. It became clear that, although *FOXO3* expression is elevated, not all the transcriptional targets showed an increase in expression. For the clusters with an increase in exposure, we observed a decreased expression of *AMACR* and *ABCA6*. The interactions with lead and DDE might dysregulate the functioning of *FOXO3* and consequently part of the FOXO-mediated transcription. Finally, we found an interesting interaction between *EP300* and *FOXO3* in our genomic interaction network. *EP300* acetylates *FOXO3*, and acetylation of *FOXO3* is linked to oxidative stress, where oxidative stress is another activator of *FOXO3* transcription. Lead is a known toxicant that can induce oxidative stress and the link between the acetylation of *FOXO3* by *EP300* and lead exposure can be an important marker.

Besides the communities related to signaling and transcription processes, our analysis highlighted interesting communities related to mRNA posttranslational processes. These processes play an essential role in gene expression and are important for defining the proteome and maintaining cellular homeostasis [89]. Because of their key role in cellular biology, posttranslational processes are tightly controlled by signaling pathways [90]. One of the communities in our network is associated with neddylation, a process that modulates important biological processes, including tumorigenesis [91], where dysfunction of this process is associated with Alzheimer's disease [92]. Our results showed an increased expression of neddylation genes *ASB16*, *DCUN1D1*, and *FBXL7*, if subjects have a higher DDE and lead exposure. *ASB16*, a member of the Ankyrin Repeat And SOCS Box (ASB) family, is of special interest since it serves as a couple of

suppressors of cytokine (SOCS) proteins and is involved in the ubiquitination of proteins. This process can mark protein targets for degradation or alter cellular localization but also impacts activity by preventing protein interactions [93]. *DCUN1D1* expression may play a role in tumor progression and development of brain metastasis of patients with Non-small cell lung carcinoma [94], a form of lung cancer associated with lead exposure [95]. Research has reported overexpression of *DCUN1D1* in the development of thyroid tumors. Overexpression of *FBXL7*, a member of the F-box protein family, is associated with mitochondrial damage and results in a depolarized and reduced ATP output, and eventually induces apoptosis. F-box proteins also modulate inflammation and innate immunity and are associated with late-onset Alzheimer's disease [96]. This might indicate that the subjects with an increased lead and DDE exposure and increased expression of *ASB16*, *DCUN1D1*, and *FBXL7* are more sensitive to develop late-onset Alzheimer's disease and cancers such as thyroid tumors.

The results of the three case studies showed that POP and heavy metal exposure might induce a sex-specific effect and we can use the transcriptomic and epigenetic profiles to stratify males and females. Within the EPIC cohort, we identified a stronger methylation effect due to POP and heavy metal exposure on the X chromosome. Hypermethylation of certain CpG sites on the X chromosomes occurred in all males and some females that are a member of cluster 2. Some of those CpG sites might be important, since DNA methylated of *XIST*, *MAGEB1*, *MAGEB2*, and *TEX11* are associated with aggressive clinical pathological features in mantle cell lymphoma [97]. The simulation with the NSHDS cohort data showed the strongest separation based on sex. Here, we found that cluster 3 (Male) has the highest lead exposure, although most of the transcriptomic changes are present in cluster 2 (Female). Cluster 3 has many CpG sites related to male gamete generation. The methylation values of those CpG sites show that in both females and males there is a trend towards hypermethylation. Endocrine-disrupting chemicals are known to be a disrupter of the male reproductive systems and lower spermatogenesis and male fertility [98]. In sperm cells, promoters of developmental genes are normally highly hypomethylated [99] but in the case of the NSHDS males, we observe a hypermethylation trend. Low expression levels of testis-specific genes, such as *TSSK4*, in the NSHDS males, might indicate an association between hypermethylation of the CpG sites and gene expression levels in the male gamete generation in association with POP and heavy metal exposure. At the same time, we observed a higher expression of genes involved in bile acid synthesis, phase 1 functionalization of compounds, and cytochrome P450 in cluster 2 (females). Bile acids are signaling molecules, with systemic endocrine function [100]. It is known that DDE and PCBs affect endocrinic gene expression, and more importantly, DDE is known to increase the expression of the CYP genes [71]. This might indicate that this group of females (cluster 2) is more susceptible to certain POP exposure since the exposure levels are comparable for all clusters.

Conclusion

The integrative omics approach resulted in the classification of subjects based on their exposure profiles. By combining the EPIC and NSHDS cohorts, we were able to identify profiles of alterations on the transcriptome and epigenome, which potentially relate to the high Lead and DDE exposure. These changes occur in important biological processes including immune-related, signaling-related, DNA repair, mRNA transcription, and mRNA translational processes. We have identified different features that are associated with disease development and these genes could be of interest to use as a potential marker or as a target to understand the diseases associated with Lead and DDE exposure. By separating the two cohorts, we were able to identify small differences in the transcriptome and the epigenome between two clusters in the EPIC cohort and the NSHDS cohort. In general, we could not identify clear epigenetic alterations that directly associate with changes in the transcriptome. However, the combination of the epigenetic and transcriptomic alterations provided us with genomic profiles that improved the stratification of subjects into subgroups to study the relationship between genomic alterations and exposure to compounds.

References

1. Lauby-Secretan B, Loomis D, Grosse Y, Ghissassi F El, Bouvard V, Benbrahim-Tallaa L, et al. Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. Lancet Oncol. 2013;14:287–8. doi:10.1016/S1470-2045(13)70104-9.
2. Gupta P, Thompson BL, Wahlang B, Jordan CT, Zach Hilt J, Hennig B, et al. The environmental pollutant, polychlorinated biphenyls, and cardiovascular disease: a potential target for antioxidant nanotherapeutics. Drug Deliv Transl Res. 2018;8:740–59. doi:10.1007/s13346-017-0429-9.
3. Antunes Fernandes EC, Hendriks HS, van Kleef RGDM, Reniers A, Andersson PL, van den Berg M, et al. Activation and Potentiation of Human GABAA Receptors by Non-Dioxin–Like PCBs Depends on Chlorination Pattern. Toxicol Sci. 2010;118:183–90. doi:10.1093/toxsci/kfq257.
4. Tchounwou PB, Yedjou CG, Patlolla AK, Sutton DJ. Heavy Metal Toxicity and the Environment. 2012. p. 133–64. doi:10.1007/978-3-7643-8340-4_6.
5. Baars A., Bakker M., Baumann R., Boon P., Freijer J., Hoogenboom LA., et al. Dioxins, dioxin-like PCBs and non-dioxin-like PCBs in foodstuffs: occurrence and dietary intake in The Netherlands. Toxicol Lett. 2004;151:51–61. doi:10.1016/j.toxlet.2004.01.028.
6. Van den Berg M, Birnbaum LS, Denison M, De Vito M, Farland W, Feeley M, et al. The 2005 World Health Organization Reevaluation of Human and Mammalian Toxic Equivalency Factors for Dioxins and Dioxin-Like Compounds. Toxicol Sci. 2006;93:223–41. doi:10.1093/toxsci/kfl055.
7. Al-Salman F, Plant N. Non-coplanar polychlorinated biphenyls (PCBs) are direct agonists for the human pregnane-X receptor and constitutive androstane receptor, and activate target gene expression in a tissue-specific manner. Toxicol Appl Pharmacol. 2012;263:7–13. doi:10.1016/j.taap.2012.05.016.
8. Schuetz EG, Brimer C, Schuetz JD. Environmental Xenobiotics and the Antihormones Cyproterone Acetate and Spironolactone Use the Nuclear Hormone Pregnenolone X Receptor to Activate the CYP3A23 Hormone Response Element. Mol Pharmacol. 1998;54:1113–7. doi:10.1124/mol.54.6.1113.
9. Rustana C, Nossen JO, Christiansen EN, Drevon C a. Eicosapentaenoic acid reduces hepatic synthesis and secretion of triacylglycerol by decreasing the activity of acyl-coenzyme A:1,2-diacylglycerol acyltransferase. J Lipid Res. 1988;29:1417–26. http://www.ncbi.nlm.nih.gov/pubmed/2853717.
10. Honkakoski P, Sueyoshi T, Negishi M. Drug-activated nuclear receptors CAR and PXR. Ann Med. 2003;35:172–82. doi:10.1080/07853890310008224.
11. Tabb MM, Kholodovych V, Grün F, Zhou C, Welsh WJ, Blumberg B. Highly chlorinated PCBs inhibit the human xenobiotic response mediated by the steroid and xenobiotic receptor (SXR). Environ Health Perspect. 2004;112:163–9. doi:10.1289/ehp.6560.

12. Zoeller RT, Dowling ALS, Vas AA. Developmental Exposure to Polychlorinated Biphenyls Exerts Thyroid Hormone-Like Effects on the Expression of RC3/Neurogranin and Myelin Basic Protein Messenger Ribonucleic Acids in the Developing Rat Brain1. Endocrinology. 2000;141:181–9. doi:10.1210/endo.141.1.7273.
13. Salama J, Chakraborty TR, Ng L, Gore AC. Effects of polychlorinated biphenyls on estrogen receptor-beta expression in the anteroventral periventricular nucleus. Environ Health Perspect. 2003;111:1278–82. doi:10.1289/ehp.6126.
14. Hochstenbach K, van Leeuwen DM, Gmuender H, Gottschalk RW, Lovik M, Granum B, et al. Global Gene Expression Analysis in Cord Blood Reveals Gender-Specific Differences in Response to Carcinogenic Exposure In Utero. Cancer Epidemiol Biomarkers Prev. 2012;21:1756–67. doi:10.1158/1055-9965.EPI-12-0304.
15. De Coster S, van Leeuwen DM, Jennen DGJ, Koppen G, Den Hond E, Nelen V, et al. Gender-specific transcriptomic response to environmental exposure in flemish adults. Environ Mol Mutagen. 2013;54:574–88. doi:10.1002/em.21774.
16. Espín-Pérez A, de Kok TMCM, Jennen DGJ, Hendrickx DM, De Coster S, Schoeters G, et al. Distinct genotype-dependent differences in transcriptome responses in humans exposed to environmental carcinogens. Carcinogenesis. 2015;36:1154–61. doi:10.1093/carcin/bgv111.
17. Espín-Pérez A, Hebels DGAJ, Kiviranta H, Rantakokko P, Georgiadis P, Botsivali M, et al. Identification of Sex-Specific Transcriptome Responses to Polychlorinated Biphenyls (PCBs). Sci Rep. 2019;9:746. doi:10.1038/s41598-018-37449-y.
18. Monographs I, Evaluation ONTHE, Risks OFC, Humans TO. Polychlorinated Biphenyls and Polybrominated Biphenyls. IARC Monogr Eval Carcinog risks to humans. 2016;107:9–500.
19. Freeman MD, Kohles SS. Plasma Levels of Polychlorinated Biphenyls, Non-Hodgkin Lymphoma, and Causation. J Environ Public Health. 2012;2012:1–15. doi:10.1155/2012/258981.
20. Colt JS, Severson RK, Lubin J, Rothman N, Camann D, Davis S, et al. Organochlorines in Carpet Dust and Non-Hodgkin Lymphoma. Epidemiology. 2005;16:516–25. doi:10.1097/01.ede.0000164811.25760.f1.
21. De Roos AJ, Hartge P, Lubin JH, Colt JS, Davis S, Cerhan JR, et al. Persistent Organochlorine Chemicals in Plasma and Risk of Non-Hodgkin's Lymphoma. Cancer Res. 2005;65:11214–26. doi:10.1158/0008-5472.CAN-05-1755.
22. Engel LS, Lan Q, Rothman N. Polychlorinated Biphenyls and Non-Hodgkin Lymphoma. Cancer Epidemiol Biomarkers Prev. 2007;16:373–6. doi:10.1158/1055-9965.EPI-07-0055.
23. Ward MH, Colt JS, Metayer C, Gunier RB, Lubin J, Crouse V, et al. Residential Exposure to Polychlorinated Biphenyls and Organochlorine Pesticides and Risk of Childhood Leukemia. Environ Health Perspect. 2009;117:1007–13. doi:10.1289/ehp.0900583.
24. Androutsopoulos VP, Hernandez AF, Liesivuori J, Tsatsakis AM. A mechanistic overview of health associated effects of low levels of organochlorine and organophosphorous pesticides. Toxicology. 2013;307:89–94. doi:10.1016/j.tox.2012.09.011.
25. Serdar B, LeBlanc WG, Norris JM, Dickinson LM. Potential effects of polychlorinated biphenyls (PCBs) and selected organochlorine pesticides (OCPs) on immune cells and blood biochemistry measures: a cross-sectional assessment of the NHANES 2003-2004 data. Environ Heal. 2014;13:114. doi:10.1186/1476-069X-13-114.
26. Hertz-Picciotto I, Park H-Y, Dostal M, Kocan A, Trnovec T, Sram R. Prenatal Exposures to Persistent and Non-Persistent Organic Compounds and Effects on Immune System Development. Basic Clin Pharmacol Toxicol. 2008;102:146–54. doi:10.1111/j.1742-7843.2007.00190.x.
27. Georgiadis P, Gavriil M, Rantakokko P, Ladoukakis E, Botsivali M, Kelly RS, et al. DNA methylation profiling implicates exposure to PCBs in the pathogenesis of B-cell chronic lymphocytic leukemia. Environ Int. 2019;126:24–36. doi:10.1016/j.envint.2019.01.068.
28. van den Dungen MW, Murk AJ, Kampman E, Steegenga WT, Kok DE. Association between DNA methylation profiles in leukocytes and serum levels of persistent organic pollutants in Dutch men. Environ Epigenetics. 2017;3. doi:10.1093/eep/dvx001.
29. Kim K-Y, Kim D-S, Lee S-K, Lee I-K, Kang J-H, Chang Y-S, et al. Association of Low-Dose Exposure to Persistent Organic Pollutants with Global DNA Hypomethylation in Healthy Koreans. Environ Health Perspect. 2010;118:370–4. doi:10.1289/ehp.0901131.
30. Rusiecki JA, Baccarelli A, Bollati V, Tarantini L, Moore LE, Bonefeld-Jorgensen EC. Global DNA Hypomethylation Is Associated with High Serum-Persistent Organic Pollutants in Greenlandic Inuit. Environ Health Perspect. 2008;116:1547–52. doi:10.1289/ehp.11338.
31. Pittman GS, Wang X, Campbell MR, Coulter SJ, Olson JR, Pavuk M, et al. Polychlorinated biphenyl exposure and DNA methylation in the Anniston Community Health Survey. Epigenetics. 2019;:1–21. doi:10.1080/15592294.2019.1666654.

32. Downie D leonard, Fenge T. Combatting Toxic Threats in the Arcticitle. McGill-Queen's University Press; 2003.
33. International Agency fro Research on Cancer (IARC). World Health Orbanization. DDT, Lindane, and 24-D. 2018.
34. BEARD J. DDT and human health. Sci Total Environ. 2006;355:78–89. doi:10.1016/j.scitotenv.2005.02.022.
35. IARC / WHO. Occupational exposures in insecticide application, and some pesticide. Lyon; 1991.
36. De Jager C. Reduced Seminal Parameters Associated With Environmental DDT Exposure and p,p'-DDE Concentrations in Men in Chiapas, Mexico: A Cross-Sectional Study. J Androl. 2006;27:16–27. doi:10.2164/jandrol.05121.
37. Turusov V, Rakitsky V, Tomatis L. Dichlorodiphenyltrichloroethane (DDT): ubiquity, persistence, and risks. Environ Health Perspect. 2002;110:125–8. doi:10.1289/ehp.02110125.
38. LI J, LI N, MA M, GIESY J, WANG Z. In vitro profiling of the endocrine disrupting potency of organochlorine pesticides. Toxicol Lett. 2008. doi:10.1016/j.toxlet.2008.10.002.
39. Harada T, Takeda M, Kojima S, Tomiyama N. Toxicity and Carcinogenicity of Dichlorodiphenyltrichloroethane (DDT). Toxicol Res. 2016;32:21–33. doi:10.5487/TR.2016.32.1.021.
40. Soto AM, Sonnenschein C, Chung KL, Fernandez MF, Olea N, Serrano FO. The E-SCREEN assay as a tool to identify estrogens: an update on estrogenic environmental pollutants. Environ Health Perspect. 1995;103 suppl 7:113–22. doi:10.1289/ehp.95103s7113.
41. Oien CW, Hurd C, Vorojeikina DP, Amold SF, Notides AC. Transcriptional activation of the human estrogen receptor by DDT isomers and metabolites in yeast and MCF-7 cells. Biochem Pharmacol. 1997;53:1161–72. doi:10.1016/S0006-2952(97)00097-X.
42. Kelce WR, Wilson EM. Environmental antiandrogens: developmental effects, molecular mechanisms, and clinical implications. J Mol Med. 1997;75:198–207. doi:10.1007/s001090050104.
43. Kabasenche WP, Skinner MK. DDT, epigenetic harm, and transgenerational environmental justice. Environ Heal. 2014;13:62. doi:10.1186/1476-069X-13-62.
44. Anirudhan TS, Sreekumari SS. Adsorptive removal of heavy metal ions from industrial effluents using activated carbon derived from waste coconut buttons. J Environ Sci. 2011;23:1989–98. doi:10.1016/S1001-0742(10)60515-3.
45. Mishra S, Bharagava RN, More N, Yadav A, Zainith S, Mani S, et al. Heavy Metal Contamination: An Alarming Threat to Environment and Human Health. In: Environmental Biotechnology: For Sustainable Future. Singapore: Springer Singapore; 2019. p. 103–25. doi:10.1007/978-981-10-7284-0_5.
46. Järup L. Hazards of heavy metal contamination. Br Med Bull. 2003;68:167–82. doi:10.1093/bmb/ldg032.
47. Inorganic and organic lead compounds. IARC Monogr Eval Carcinog Risks Hum. 2006;87:1–471.
48. IARC. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. IARC Monogr Eval Carcinog Risks Hum. 2012;100F February:225–248.
49. Wani AL, Ara A, Usmani JA. Lead toxicity: a review. Interdiscip Toxicol. 2015;8:55–64. doi:10.1515/intox-2015-0009.
50. Sokol RZ, Berman N. The effect of age of exposure on lead-induced testicular toxicity. Toxicology. 1991;69:269–78. doi:10.1016/0300-483X(91)90186-5.
51. Silbergeld E. Facilitative mechanisms of lead as a carcinogen. Mutat Res Mol Mech Mutagen. 2003;533:121–33. doi:10.1016/j.mrfmmm.2003.07.010.
52. Eid A, Bihaqi SW, Renehan WE, Zawia NH. Developmental lead exposure and lifespan alterations in epigenetic regulators and their correspondence to biomarkers of Alzheimer's disease. Alzheimer's Dement Diagnosis, Assess Dis Monit. 2016;2:123–31. doi:10.1016/j.dadm.2016.02.002.
53. IARC. IARC monographs on the evaluation of carcinogenic risks to humans, volume 100c: Cadmium and cadmium compounds. IARC Monogr. 2012;1993:121–45. https://monographs.iarc.fr/ENG/Monographs/vol100C/mono100C-8.pdf.
54. Nawrot T, Plusquin M, Hogervorst J, Roels HA, Celis H, Thijs L, et al. Environmental exposure to cadmium and risk of cancer: a prospective population-based study. Lancet Oncol. 2006;7:119–26. doi:10.1016/S1470-2045(06)70545-9.
55. ILYASOVA D, SCHWARTZ G. Cadmium and renal cancer. Toxicol Appl Pharmacol. 2005;207:179–86. doi:10.1016/j.taap.2004.12.005.
56. Song J kun, Luo H, Yin X hai, Huang G lei, Luo S yang, Lin D ren, et al. Association between cadmium exposure and renal cancer risk: a meta-analysis of observational studies. Sci Rep. 2016;5:17976. doi:10.1038/srep17976.

57. Joseph P. Mechanisms of cadmium carcinogenesis☆. Toxicol Appl Pharmacol. 2009;238:272–9. doi:10.1016/j.taap.2009.01.011.
58. Waalkes M. Cadmium carcinogenesis. Mutat Res Mol Mech Mutagen. 2003;533:107–20. doi:10.1016/j.mrfmmm.2003.07.011.
59. Sanders A, Smeester L, Rojas D, DeBussycher T, Wu M, Wright F, et al. Cadmium exposure and the epigenome: Exposure-associated patterns of DNA methylation in leukocytes from mother-baby pairs. Epigenetics. 2014;9:212–21. doi:10.4161/epi.26798.
60. Desaulniers D, Xiao G, Lian H, Feng Y-L, Zhu J, Nakai J, et al. Effects of Mixtures of Polychlorinated Biphenyls, Methylmercury, and Organochlorine Pesticides on Hepatic DNA Methylation in Prepubertal Female Sprague-Dawley Rats. Int J Toxicol. 2009;28:294–307. doi:10.1177/1091581809337918.
61. Bell MR, Hart BG, Gore AC. Two-hit exposure to polychlorinated biphenyls at gestational and juvenile life stages: 2. Sex-specific neuromolecular effects in the brain. Mol Cell Endocrinol. 2016;420:125–37. doi:10.1016/j.mce.2015.11.024.
62. Aluru N, Karchner SI, Krick KS, Zhu W, Liu J. Role of DNA methylation in altered gene expression patterns in adult zebrafish (Danio rerio) exposed to 3, 3', 4, 4', 5-pentachlorobiphenyl (PCB 126). Environ Epigenetics. 2018;4. doi:10.1093/eep/dvy005.
63. Georgiadis P, Hebels DG, Valavanis I, Liampa I, Bergdahl IA, Johansson A, et al. Omics for prediction of environmental health effects: Blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. Sci Rep. 2016;6:20544. doi:10.1038/srep20544.
64. Valavanis I, Sifakis EG, Georgiadis P, Kyrtopoulos S, Chatziioannou AA. A Composite Framework for the Statistical Analysis of Epidemiological DNA Methylation Data with the Infinium Human Methylation 450K BeadChip. IEEE J Biomed Heal Informatics. 2014;18:817–23. doi:10.1109/JBHI.2014.2298351.
65. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587. doi:10.1186/1471-2105-11-587.
66. Kelly RS, Kiviranta H, Bergdahl IA, Palli D, Johansson A-S, Botsivali M, et al. Prediagnostic plasma concentrations of organochlorines and risk of B-cell non-Hodgkin lymphoma in envirogenomarkers: a nested case-control study. Environ Heal. 2017;16:9. doi:10.1186/s12940-017-0214-8.
67. BARANY E, BERGDAHL IA, SCHÜTZ A, SKERFVING S, OSKARSSON A. Inductively Coupled Plasma Mass Spectrometry for Direct Multi-element Analysis of Diluted Human Blood and Serum. J Anal At Spectrom. 1997;12:1005–9. doi:10.1039/A700904F.
68. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. Comput Stat Data Anal. 2007;52:155–73. doi:10.1016/j.csda.2006.11.006.
69. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007;23:1495–502. doi:10.1093/bioinformatics/btm134.
70. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13:966–7. doi:10.1038/nmeth.4077.
71. CTD. Curated chemical-disease data were retrieved from the Comparative Toxicogenomics Database (CTD), North Carolina State University, Raleigh, NC and Mount Desert Island Biological Laboratory, Salisbury Cove, Maine. World Wide Web. 31-08-2017. 2017. http://ctdbase.org/.
72. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13. doi:10.1093/nar/gky1131.
73. Souza TM, Rieswijk L, Beucken T van den, Kleinjans J, Jennen D. Persistent transcriptional responses show the involvement of feed-forward control in a repeated dose toxicity study. Toxicology. 2017;375:58–63. doi:10.1016/j.tox.2016.10.009.
74. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008:P10008. doi:10.1088/1742-5468/2008/10/P10008.
75. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2012;41:D377–86. doi:10.1093/nar/gks1118.
76. Fuller-Pace F V. DEAD box RNA helicase functions in cancer. RNA Biol. 2013;10:121–32. doi:10.4161/rna.23312.
77. Jiang P, Hou Z, Bolin JM, Thomson JA, Stewart R. RNA-Seq of Human Neural Progenitor Cells Exposed to Lead (Pb) Reveals Transcriptome Dynamics, Splicing Alterations and Disease Risk Associations. Toxicol Sci. 2017;159:251–65. doi:10.1093/toxsci/kfx129.

78. Sonneborn D, Park H-Y, Petrik J, Kocan A, Palkovicova L, Trnovec T, et al. Prenatal polychlorinated biphenyl exposures in eastern Slovakia modify effects of social factors on birthweight. Paediatr Perinat Epidemiol. 2008;22:202–13. doi:10.1111/j.1365-3016.2008.00929.x.
79. Jeng HA. Exposure to Endocrine Disrupting Chemicals and Male Reproductive Health. Front Public Heal. 2014;2. doi:10.3389/fpubh.2014.00055.
80. Organ SL, Tsao M-S. An overview of the c-MET signaling pathway. Ther Adv Med Oncol. 2011;3 1_suppl:S7–19. doi:10.1177/1758834011422556.
81. Kim B, Jung N, Lee S, Sohng JK, Jung HJ. Apigenin Inhibits Cancer Stem Cell-Like Phenotypes in Human Glioblastoma Cells via Suppression of c-Met Signaling. Phyther Res. 2016;30:1833–40. doi:10.1002/ptr.5689.
82. Zhang Y, Xia M, Jin K, Wang S, Wei H, Fan C, et al. Function of the c-Met receptor tyrosine kinase in carcinogenesis and associated therapeutic opportunities. Mol Cancer. 2018;17:45. doi:10.1186/s12943-018-0796-y.
83. Simon T, Bromberg JS. Regulation of the Immune System by Laminins. Trends Immunol. 2017;38:858–71. doi:10.1016/j.it.2017.06.002.
84. Garg M, Braunstein G, Koeffler HP. LAMC2 as a therapeutic target for cancers. Expert Opin Ther Targets. 2014;18:979–82. doi:10.1517/14728222.2014.934814.
85. Gu B, Chen P-L. Expression of PCNA-binding domain of CtIP, a motif required for CtIP localization at DNA replication foci, causes DNA damage and activation of DNA damage checkpoint. Cell Cycle. 2009;8:1409–20. doi:10.4161/cc.8.9.8322.
86. Alvarez-Fernández M, Medema RH. Novel functions of FoxM1: from molecular mechanisms to cancer therapy. Front Oncol. 2013;3. doi:10.3389/fonc.2013.00030.
87. Gao Y-F, Zhu T, Mao X-Y, Mao C-X, Li L, Yin J-Y, et al. Silencing of Forkhead box D1 inhibits proliferation and migration in glioma cells. Oncol Rep. 2017;37:1196–202. doi:10.3892/or.2017.5344.
88. Gurnari C, Falconi G, De Bellis E, Voso MT, Fabiani E. The Role of Forkhead Box Proteins in Acute Myeloid Leukemia. Cancers (Basel). 2019;11:865. doi:10.3390/cancers11060865.
89. Hershey JWB, Sonenberg N, Mathews MB. Principles of Translational Control: An Overview. Cold Spring Harb Perspect Biol. 2012;4:a011528–a011528. doi:10.1101/cshperspect.a011528.
90. Sonenberg N, Hinnebusch AG. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell. 2009;136:731–45. doi:10.1016/j.cell.2009.01.042.
91. Zhou L, Jiang Y, Luo Q, Li L, Jia L. Neddylation: a novel modulator of the tumor microenvironment. Mol Cancer. 2019;18:77. doi:10.1186/s12943-019-0979-1.
92. Chen Y, Neve RL, Liu H. Neddylation dysfunction in Alzheimer's disease. J Cell Mol Med. 2012;16:2583–91. doi:10.1111/j.1582-4934.2012.01604.x.
93. Anasa VV, Ravanan P, Talwar P. Multifaceted roles of ASB proteins and its pathological significance. Front Biol (Beijing). 2018;13:376–88. doi:10.1007/s11515-018-1506-2.
94. Yoo J, Lee S-H, Lym K-I, Park SY, Yang S-H, Yoo C-Y, et al. Immunohistochemical Expression of DCUN1D1 in Non-small Cell Lung Carcinoma: Its Relation to Brain Metastasis. Cancer Res Treat. 2012;44:57–62. doi:10.4143/crt.2012.44.1.57.
95. Scimeca M, Orlandi A, Terrenato I, Bischetti S, Bonanno E. Assessment of metal contaminants in non-small cell lung cancer by EDX microanalysis. Eur J Histochem. 2014;58. doi:10.4081/ejh.2014.2403.
96. Tosto G, Fu H, Vardarajan BN, Lee JH, Cheng R, Reyes-Dumeyer D, et al. F-box/LRR-repeat protein 7 is genetically associated with Alzheimer's disease. Ann Clin Transl Neurol. 2015;2:810–20. doi:10.1002/acn3.223.
97. Enjuanes A, Fernàndez V, Hernández L, Navarro A, Beà S, Pinyol M, et al. Identification of Methylated Genes Associated with Aggressive Clinicopathological Features in Mantle Cell Lymphoma. PLoS One. 2011;6:e19736. doi:10.1371/journal.pone.0019736.
98. Roeleveld N, Bretveld R. The impact of pesticides on male fertility. Curr Opin Obstet Gynecol. 2008;20:229–33. doi:10.1097/GCO.0b013e3282fcc334.
99. Cui X, Jing X, Wu X, Yan M, Li Q, Shen Y, et al. DNA methylation in spermatogenesis and male infertility. Exp Ther Med. 2016;12:1973–9. doi:10.3892/etm.2016.3569.
100. Houten SM, Watanabe M, Auwerx J. Endocrine functions of bile acids. EMBO J. 2006;25:1419–25. doi:10.1038/sj.emboj.76010

Chapter 5

Integration of omics layers with SNP risk allele scores to identify arsenic-related exposure effects

T.J.M. Kuijpers[1]
M.Y.A. Rehman[2]
J. Krauskopf[1]
R. Naseem Malik[2]
J.C.S. Kleinjans[1]
D.G.J. Jennen[1]

[1] Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands
[2] Environmental Health Laboratory, Department of Environmental Sciences, Quaid-i-Azam University, Islamabad, Pakistan

Abstract

Introduction: Arsenic contamination of drinking water is a worldwide problem and is a significant threat to public health. Research suggests that there are different transcriptomic and epigenetic alterations due to arsenic exposure, but the interaction between the two layers is not fully understood. Furthermore, an individual's susceptibility due to DNA polymorphisms, such as single nucleotide polymorphisms, can influence the response to arsenic exposure per individual. In this study, we aim by integrating the transcriptomic and epigenomic omics layers from a Pakistani cohort to identify exposure-related profiles for subgroups with different susceptibility to arsenic exposure.

Method: Clusters with distinct transcriptome-epigenome profiles will be derived by performing multi-layer Nonnegative Matrix Factorization (NMF). Because of the rather small sample size (n=57), we will investigate whether data filtering and data processing tools could increase the arsenic patterns in our data. Therefore, an a priori filtering method based on the features derived from two linear mixed models as well as using the M-values for DNA methylation.

Results and conclusion: The M-value for DNA methylation did result in more detected subgroups within our data, whereas the use of a semi-supervised multi-layer approach did not give more information in comparison to the standard multi-layer NMF approach. When M-values are used in the omics integration strategy, we could identify genomic patterns that play a role in signaling pathways, mRNA translation processes, Golgi transport, metabolic transport, and innate and adaptive immune processes. Moreover, their association with arsenic exposure and their relationship with cardiovascular disease and diabetes makes these communities highly relevant.

Supplementary data available at: https://github.com/TJMKuijpers/PhDThesis

Introduction

Exposure to toxic chemicals, such as arsenic, poses a significant threat to public health [1]. The exposure to elevated levels of inorganic arsenic occurs mainly through the consumption of groundwater-derived drinking water containing high levels of inorganic arsenic but also through the consumption of food irrigated with high arsenic water resources (WHO) [2, 3]. Although arsenic contamination is a worldwide problem, it is a particular burden in areas of India, Argentina, Chile, Pakistan, and Bangladesh because of the high arsenic concentration in drinking water [4, 5]. Drinking arsenic-contaminated water over a long period is toxic and does affect various organs. Arsenic exposure is associated with multiple cancers including skin, lung, bladder, liver, and kidney [6, 7]. Furthermore, it has been linked to cardiovascular diseases [8, 9], respiratory diseases [10], and impaired neurodevelopment [11–13].

Due to the role of arsenic in toxicity and cancer development, it is key to understand the molecular mechanisms associated with arsenic-induced toxicity. The carcinogenic capacity of arsenic is linked to its biotransformation [14]. Multiple mechanisms have been proposed that relate arsenic toxicity to genotoxicity, in particular induction of oxidative stress [15], DNA repair and ligase inhibition, signal transduction, and chromosomal aberrations. These mechanisms are complex and not fully understood because arsenic metabolism involves five metabolites, which can induce toxic effects [16].
On a molecular level, arsenic exposure is associated with changes in both DNA methylation and gene expression. Rehman et al [17, 18] identified transcriptomic alterations concerning arsenic exposure and found a relation between the gene perturbations and nonalcoholic fatty liver disease, insulin resistance, and cancer-related pathways. Andrew et al [19] showed that high arsenic exposure is associated with an overrepresentation of genes involved in immune function, defense response, cell growth, apoptosis, regulation of cell cycle, and T-cell receptor signaling pathways. Arsenic exposure leads to a downregulation of tumor suppressor genes and an increase of pro-inflammatory mitogen-activated protein kinase pathways leading to a tumor-promoting microenvironment [20].

Arsenic does not only directly affect gene expression; it may also alternate gene expression via the perturbation of epigenetic control mechanism. Recent findings implied different histone methylation patterns in men and women as a result of arsenic exposure [21]. More importantly, the detoxification of arsenic requires the use of S-Adenosyl methionine (SAM) as a methyl donor, and consequently, arsenic-related epigenetic effects mainly derive from the depletion of the cellular methyl pool [22]. This depletion of the methyl group could be the reason why various studies show hypomethylation patterns in association with arsenic exposure [23]. Multiple studies show increased expression of oncogenes due to DNA hypomethylation of the promoter region induced by arsenic [24–26]. Moreover, specific promoter hypermethylation is observed after arsenic exposure,

affecting CpG islands of tumor suppressor genes including TP53, P16, and P21 [26]. Breda et al showed the consequence of TP53 hypomethylation due to arsenic exposure in vitro, which leads to altered expression of transcriptional targets of TP53 [27].

Because of the different epigenetic and transcriptomic alterations induced by arsenic exposure, the molecular mechanism induced by arsenic exposure is quite complex to understand. Furthermore, an individual's DNA sequence, with different polymorphisms of which a single nucleotide polymorphism is the most common type, is hypothesized to be a major cause of inter-individual variations in arsenic susceptibility. Since biotransformation of arsenic consists of reduction, oxidation, and methylation steps [24], polymorphisms affecting genes encoding important reductases and methyltransferases associated with arsenic metabolism are of relevance. These polymorphisms might influence arsenic metabolism and therefore have a downstream effect on the arsenic exposure-related events. Different polymorphisms are linked to arsenic exposure and impact human health [17]. Also, polymorphisms affecting genes involved in either oxidative stress or DNA damage repair pathways have been shown to impact the risk of arsenic-induced cancer [28].

We hypothesize, taking into account polymorphisms, DNA methylation, and gene expression, it is possible to identify exposure-related profiles for subgroups with different susceptibility to arsenic exposure, independently of the arsenic dose. Therefore, omics data from a Pakistani cohort study by Rehman et al [17,18] are used to study arsenic exposure. In their initial study, Rehman et al applied two Linear mixed models (LMMs) to identify genes and DNA regions associated with arsenic exposure. These LMMs can explain the variation in either gene expression or DNA methylation by arsenic exposure while considering cofounding effects such as gender, age, exposure, BMI, polymorphisms, lymphocyte count, and village. On the transcriptome, arsenic affects pathways such as insulin resistance- and NAFLD-related pathways. A separate LMM analysis on the epigenome identified DNA regions affected by arsenic exposure that relate to muscle contraction, cardiovascular diseases, and cell development. These differences in pathways show that arsenic exposure might induce a variety of alterations. However, this study placed each subject into an exposure group (low, medium, high) to determine exposure-related genes and DNA regions. Therefore, it is needed to investigate the power of an unsupervised strategy to identify arsenic exposure-related profiles that do depend not on predefined exposure groups but a subject's susceptibility to arsenic exposure.

To test this hypothesis, we aim to perform a combined omics integration by performing multi-layer Nonnegative Matrix Factorization (multi-layer NMF) to derive clusters with distinct genomic profiles. For each cluster, the SNP risk sore [29] is used to determine whether the obtained groups show a tendency towards a certain SNP or a combination of SNPs. Since the Pakistani cohort is quite small (n=57),

we will investigate different ways to increase the arsenic exposure-related signals. First, features are selected based on a priori knowledge from two single omics studies by Rehman et al [17,18], to perform a semi-supervised multi-layer NMF approach. Second, a test is performed to assess whether the DNA methylation M-value is a better choice in case of small sample sizes. Although the gold standard is the β-value, current insights suggest that M-values can store more statistical power [30]. Moreover, recent findings suggest that in large sample size studies there is no difference in predictive power when the M-value or β-value is chosen, but the M-value might perform better in small study sizes [31]. Investigating the added value of a semi-supervised multi NMF method or the use of the DNA methylation M-value will help us to understand whether omics integration is applicable for small cohort studies with inter-individual variations to arsenic exposure.

Method

Study design and data processing

The study design is described in detail by Rehman et al [17]. Total urinary arsenic levels were used as the main biomarker for exposure and based on their levels, subjects were stratified into low (0-50 microgram/gram creatine), medium (51-100 microgram/gram creatine), and high (<101 microgram/gram creatine). Subjects were excluded if they are smokers or belong to the age group of 11-15 years old. This resulted in a sample population of 57 subjects, for which gene expression and DNA methylation are measured.

RNA samples from individual samples were hybridized on Agilent SurePrint G3 Human Gene Expression 8 × 60K arrays. Quantile normalization and data processing were performed using ArrayQC (https://github.com/BiGCAT-UM/arrayQC_Module/), a quality control pipeline in R. This pipeline is used to flag bad spots, controls, and spots with too low intensity and normalized the data with local background correction [17].

DNA samples are hybridized from Nimblegen 2.1M Deluxe Promotors arrays. Log2 ratios of the intensities were computed (ratio of MeDIP signal/Input signal) and centered around zero. Methylated DNA regions are identified via the Probe Sliding Window-ANOVA algorithm by applying a sliding window of 750 base pairs. We selected peaks in the methylated DNA regions if a region contained at least 8 consecutive probes. NimbleScan v2.6 software HOMER was used to map those peaks to regions of the human genome (HG19) [32].

A variance filter has been applied to reduce the number of redundant measured probes in both the gene expression and DNA methylation data. A median variance threshold is used to remove all low variant probes in both the transcriptomic and epigenetic data. This variance filtered data will be used for the omics integration with the multi-layer NMF approach and further referenced as the input data.

Polymorphisms and risk allele score

Single Nucleotide Polymorphisms (SNPs) have been measured for 7 genes: GSTT1, GSTM1, DNMT1, EGFR, MTHFR, ERCC2, and As3MT. As described in the original paper by Rehman et al [17], these SNPs are selected based on their involvement in arsenic metabolism, DNA methylation, Folic Acid metabolism, and cancer risk. Furthermore, they are non-synonymous SNPs, meaning they are in protein-coding regions and have different alleles that encode for different amino acids. Homozygous carriers with 2 alleles were coded as 0, heterozygous carriers 1, and homozygous carriers with an increased health risk are coded 2.

DNA methylation: β-values and M-values

We used DNA methylation arrays to determine the methylation status of a specific DNA region by measuring the intensity of the unmethylated and methylation probe variant of the DNA region. The methylation status of a DNA region can be expressed by either the β-value (Equation 1) or the M-value (Equation 2), in which IM and IU are the signal intensity of the measured methylated and unmethylated probe. The β-value follows a β-distribution and is finite within the range from 0 (hypomethylated) to 1 (hypermethylated). Because of this β-distribution, it is statistically different from the most common infinite scale in expression studies. Du et al showed that the variance of the β-value is not constant and varies with the β-value [30]. Therefore, they proposed to calculate the methylation levels as the log2 ratios of the intensities of the methylated probe versus the unmethylated probe: the M-value. As one can see, the M-value is the logistic version of the β-value. Du et al showed that, although the β-value allows a more intuitive biological interpretation, the M-value is more statistically valid for the differential analysis of methylation levels.

$$\beta_i = \frac{\max\left(y_{i,meth}, 0\right)}{\max\left(y_{i,unmeth}, 0\right) + \max\left(y_{i,meth}, 0\right) + \alpha} \quad (1)$$

$$M_i = log_2\left(\frac{\max\left(y_{i,meth}, 0\right) + \alpha}{\max\left(y_{i,unmeth}, 0\right) + \alpha}\right) \quad (2)$$

Multi-layer Nonnegative Matrix Factorization

Multi-layer NMF is applied to reduce the features in the omics data sets while taking into account the role of each omics layer during the clustering of the subjects. The original data matrices X_i are estimated by the product W_iH (Equation 3). The matrices W_i and H are updated by their update rules (Equations 4 and 5 respectively) while the Kullback – Leibler divergence is minimized (Equation 6). In the end, n matrices W are obtained that store the latent features and one coefficient matrix H that stores the clustering coefficients. To analyze the difference in methylation and gene expression profile of each cluster, each matrix W_i is scored by using the method proposed by Kim et al [16]. For each cluster, the entities are selected as features, if those entities have a high probability of explaining a cluster.

$$\sum_{i=1}^{n} X_i \approx \sum_{i=1}^{n} W_i H \qquad (3)$$

$$W_{w+1} = W * \frac{\frac{X_i}{W_i H} H^T}{\sum H} \qquad (4)$$

$$H_{H+1} = H * \frac{\sum \frac{X_i}{W_{iH}}^T}{\sum W_i^T} \quad (5)$$

$$KL\ divergence = \sum \left(X_i * \log \left(\frac{X_i}{W_i H} - X_i + W_i H \right) \right) \quad (6)$$

To estimate the number of k clusters in our data, 50 simulations are performed to calculate the silhouette score. The most optimal value for k is used to calculate the final multi-layer NMF solution with 150 simulations to correct for randomness.

Semi-supervised NMF: selection features based on linear mixed models
A semi-supervised multi-layer NMF approach is created by removing all gene and CpG probes unrelated to arsenic exposure. We selected the significant genes and DNA regions as predicted by the linear mixed models and removed all non-significant probes from the transcriptome and epigenome data. Here, two simulations are performed: i) FDR-selected genes and p-value selected DNA regions to balance the two input matrices, and ii) FDR selected genes and DNA regions with a cutoff < 0.05.

Gen*omic interaction network*
To study the relationship between the transcriptome and epigenome, we constructed a genomic interaction network from the obtained feature genes and DNA regions. This network will contain the feature genes and DNA regions, to understand how different exposures change DNA methylation or gene expression levels. A feature gene will be represented in the network as a node and node interactions are derived from OmniPath [33] (for gene-gene interactions), the Comparative Toxicogenomics Database (CTD) [34] for compound – gene interactions), STRINGdb [35], transcription catalog from Souza et al [36] (transcriptional activation or inhibition of genes), and CpG – gene interactions.
The Louvain method [37] was used to detect communities in the genomic interaction network. Communities are groups of nodes, which share a high degree of interactions with each other compared to nodes in other communities. Initially, a node is placed inside its own community and the algorithm will merge small communities of this will increase the modularity of a partition of the network. Low modularity (-0.5) indicates non-modular clustering, whereas maximum modularity (1) indicates a fully modular clustering.

Results

We have investigated the role of semi-supervised multi-layer NMF and the two distributions for DNA methylation. First, our standard transcriptome – epigenome pipeline is applied by using the multi-layer NMF approach with β-values for DNA methylation and log2(intensity) for gene expression. This will give us a baseline simulation used to evaluate the performance of two further simulations: 1) semi-supervised multi-layer NMF approach with β-values and 2) (semi-supervised) multi-layer NMF approach with the M-value for DNA methylation. For each simulation, multiple case studies are defined to evaluate one single parameter (Table 5.1). The case studies differ in selection criteria for the DNA methylation filter. Hence, the number of selected DNA regions would drastically reduce in the case of the FDR selection (case studies 2 and 5), but by applying a p-value filter on the epigenome (case studies 1 and 4), the number of probes in the transcriptomic and epigenetic input data are in balance.

Table *5.1* Overview of the simulations with their filter options and settings

Simulation	***Case study***	***Gene Expression filter***	***DNA methylation filter***
Multi-layer NMF with β-values	*Baseline*	-	-
Semi-supervised multi-layer NMF with β-values	*1*	LMM features FDR ≤ 0.05	LMM features p-value ≤ 0.05
	2	LMM features FDR ≤ 0.05	LMM features FDR ≤ 0.05
Multi-layer NMF with M-values	*3*	-	-
Semi-supervised Multi-layer NMF with M-values	*4*	LMM features FDR ≤ 0.05	LMM features p-value ≤ 0.05
	5	LMM features FDR ≤ 0.05	LMM features FDR ≤ 0.05

Baseline simulation: unsupervised omics integration without prior SNP information

For the baseline simulation, DNA methylation and gene expression data of the 57 subjects are integrated without adding the risk allele score for the seven SNPs. This baseline simulation is used to examine if the multi-layer NMF approach is capable of finding exposure-related profiles. Our results highlighted two main groups: one small group of 13 samples and a second larger group of 44 samples (Figure 5.1A, clusters outlined with green). Each cluster contains subjects exposed to different levels of arsenic exposure (low, medium, and high) (Figure 5.1B).

To identify the driving force behind the clustering of the subjects, we examined the genomic patterns for each cluster. Here, it became clear that there is a strong

difference in methylation patterns between cluster 1 and cluster 2, with a very strong trend towards hypermethylation in cluster 1. This might imply a stronger impact of arsenic exposure on the epigenome and could correlate to some of the SNPs that can alter DNA methylation. Although the obtained genomic profiles show differences, it is important to gain insights into their role in arsenic exposure. Therefore, the arsenic-gene interactions from CTD are used to highlight curated gene-arsenic interactions [34].

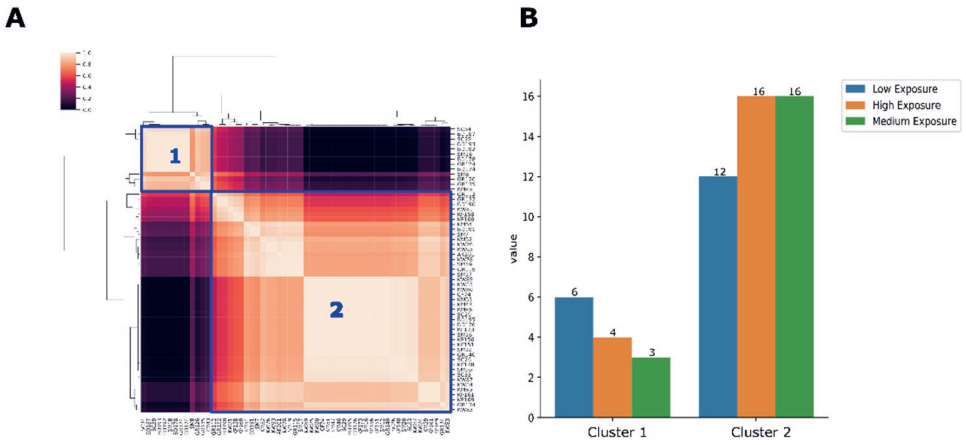


Figure 5.1 A: Consensus map of the cluster results for the 150 NMF simulations. These results show two clusters, in which cluster 1 has 13 members and cluster 2 has 44 members. **B**: Distribution of exposure levels within the two clusters. Here we see that both clusters contain a mixture of exposure levels, which could indicate that their susceptibility to arsenic exposure is placing them together in one cluster.

On the transcriptome, the selected genes by multi-layer NMF play various roles in the VEGF signaling pathway, MAP kinase pathway, signaling by receptor Tyrosine kinases and Valine, Leucine, and isoleucine degradation (Supplementary table 5.1). From this set of 603 genes, 84 genes are known to interact with arsenic (CTD [34]), including *FMOD, CCL2, GFM1, GATA4, SLC39A12, ENPP3, HERC2, C20orf151, WNT7B,* and *AJAP1*. Most of these genes are protein-coding genes and perturbation of these genes could have a downstream effect such as the WNT signaling pathway (*WTN7B*), IL-17 signaling pathway (*CCL2* and *FOSB*), inflammation (*CCL2*) [38], cardiovascular disease (*FMOD* and *GATA4*) [39, 40], DNA repair (*HERC2*) [41], and diabetes (*CCL2* and *GATA4*) [42, 43]. On the epigenome, 13172 DNA regions stratify cluster 1 and cluster 2. Of these DNA regions, 1916 DNA regions map against the currently known arsenic - gene interactions and could be important for arsenic exposure. The genes associated with these DNA regions are involved in transcriptional regulation as well as other vital cellular processes [44–46]. Finally, a combined effect is present on both the

transcriptome and epigenome for a set of genes known to interact with arsenic. A hypomethylation and consequently a higher gene expression of *FMOD, CCL2, GFM1, GATA4, ENPP3*, and *FOSB* in cluster 1, whereas the opposite is observed in cluster 2.

To study the relation between SNP risk allele scores and the cluster results, we investigated the distribution of all risk allele scores within a cluster (Figure 5.2). Here, it becomes apparent that no clear pattern arose within one cluster. This indicates that the obtained clusters, although they contain exposure-related features, cannot be used to investigate the role of SNPs and the subject's susceptibility to arsenic exposure. Exploratory analysis of the meta data showed no association with skin disease, digestive tract disease, respiratory disease, or cardiovascular disease. Furthermore, no link exists between Body Mass Index (BMI), social-economic status, residential proximity, and the two clusters.

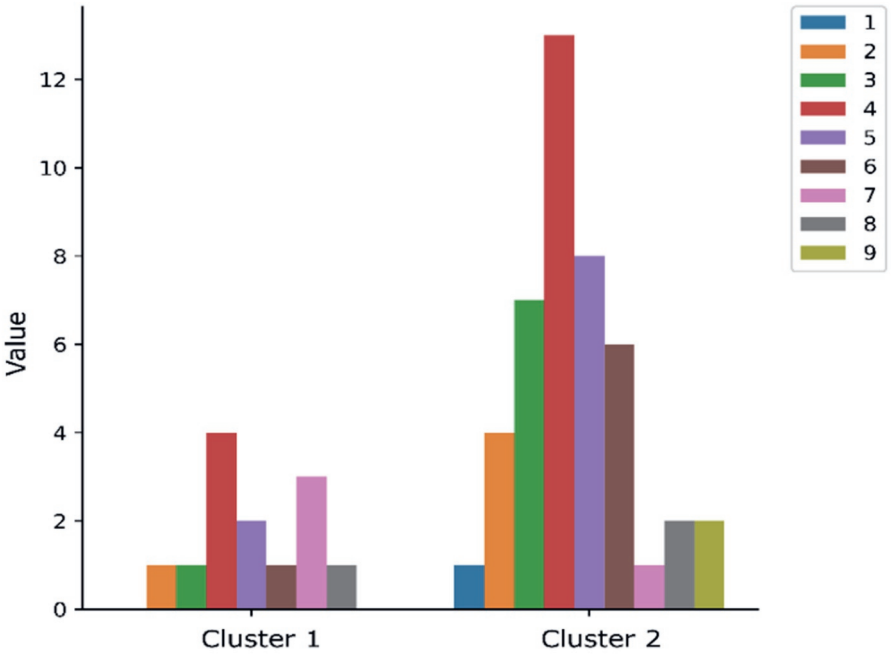


Figure 5.2 SNP risk allele distribution: for every cluster member we calculated the total risk allele score by taking the sum of each individual SNP risk allele score. The total risk allele scores (y-axis) are visualized to gain a better understanding of the risk allele score.

Since our results show two clusters with no relation to arsenic exposure or SNP risk allele score, we performed a semi-supervised multi-layer NMF approach to investigate whether this approach is more helpful for clustering subjects based on their susceptibility to arsenic exposure. Therefore, a new input data set is derived from the transcripts and DNA regions identified with two linear mixed models

(individual model for the transcriptome and epigenome) as an input for the multi-layer NMF method.

Semi-supervised multi-layer NMF: improving cluster prediction to identify arsenic exposure effect

In the previous analysis, the results showed that unsupervised multi-layer NMF can extract different DNA methylation and gene expression profiles. The obtained clusters show a mixture of exposure groups, which can be due to different susceptibility to arsenic exposure or to the fact that the profiles do not relate to arsenic exposure. We hypothesize that by using previously identified single-layer arsenic exposure signals, we improve our integration approach because irrelevant signals are removed. Moreover, by integrating SNP-associated genes and DNA regions the clusters might show susceptibility-related signals to arsenic exposure.
To integrate these confounding variables, the output generated by two Linear Mixed Models (LMMs) [17] are used as input for the multi-layer NMF. Two case studies were defined (Table 5.1, case studies 1 and 2) and their output has been compared against the baseline simulation. Filtering the input data with the LMM features does improve the cluster stability only for case study 1 (Gene FDR, DNA regions p-value) compared to the baseline simulation (Figure 5.3). Although the silhouette score increased, it does not reach the threshold ≥ 0.7 and only the solution for k=2 is stable. In case study 2, with the most stringent selection criteria, it becomes clear that the FDR selection of both transcripts and DNA regions results in a decrease in the stability of the different clusters (Figure 5.3, case study 2). A more stringent selection did not improve the clustering and possibly removed patterns containing explanatory features.

We hypothesized that one of the advantages of performing a semi-supervised multi-layer NMF approach is that the features should relate to arsenic exposure. The features should reflect profiles that relate to the impact of arsenic exposure for the obtained clusters. However, it does not improve the cluster stability and does not lead to more clusters identified. This could be because the signal related to arsenic exposure in both omics layers is not that strong, or because the sample size is too small. Therefore, an additional simulation will be performed with the M-value for DNA methylation.

Comparison between M-value and β-value and cluster estimation

Multi-layer NMF depends on the patterns in the input data to identify the number of clusters hidden in the data. Therefore, it is of importance to upgrade the “visibility” of these patterns in our data sets. As mentioned previously, the M-value is suggested to generate higher statistical power, and therefore we expect the transition of the β-value to the M-value to increase the identified clusters.

Since M-values are presented by a dual bell-shaped curve, centered around zero (i.e. a negative M-value means hypomethylation, a positive M-value means

hypermethylation), it is necessary to transform these onto a nonnegative scale. To keep the important information on hypo- and hypermethylation, the DNA methylation matrix is separated into two matrices: one matrix DNA_{hyper} ($\mathbb{R}^+$) and one matrix DNA_{hypo} ($\mathbb{R}^-$). By calculating the absolute values for the DNA_{hypo} matrix, important information on hypo- and hypermethylation is saved. This approach also results in two sparse matrices, which could improve the multi-layer NMF method.

For each case study (Table 5.1, multi-layer NMF with M-values), the multi-layer NMF method for a different number of k clusters and calculated the silhouette score (Figure 5.3). These results highlighted the influence the effect of integrating the M-value for DNA methylation values as well as applying a semi-supervised multi-layer NMF approach. Overall, an increase in the silhouette score for the different k clusters is observed compared to the baseline simulation, indicating we can distinguish more subgroups in our data.

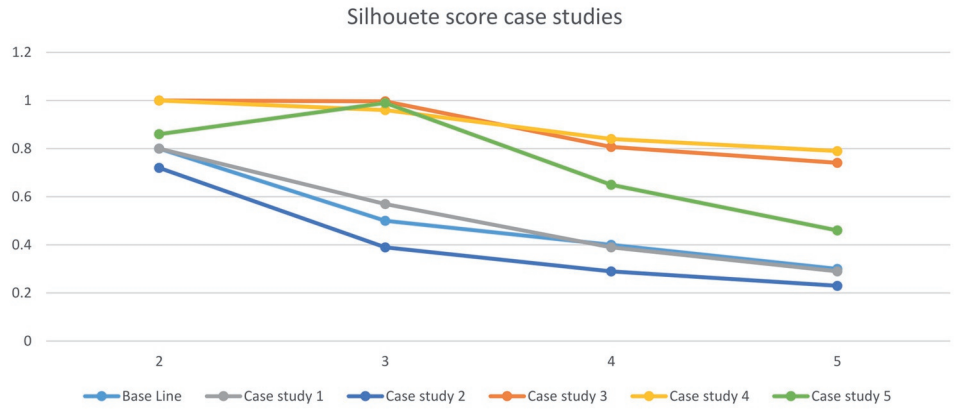


Figure 5.3 Silhouette score over a range k from 2 to 5 for all case studies**.** An optimal silhouette score for k=2 is reached for all three simulations. Baseline simulation consists of the variance filtered genes and DNA regions. Case study 1 consists of the FDR selected genes and p-value selected DNA regions (β-value). Case study 2 consists of the FDR-selected genes and DNA regions (β-value). Case study 3 has the same filter as the baseline simulation but with M-values for DNA methylation. Case study 4 consists of the FDR selected genes and p-value selected DNA regions (M-value). Case study 5 consists of the FDR selected genes and DNA regions (M-value).

Here, it becomes clear that the FDR selection criteria for both the transcriptome and epigenome lead to less stable identified clusters (case study 5). This is in line with case study 2 and might highlight again the risk of removing valuable features from the data that drive the clustering of the cohort data. Therefore, the M-value could be a better distribution in the case of the Pakistani cohort since the multi-layer NMF method could find more patterns hidden in the data. These results also indicate that we can apply an unsupervised multi-layer NMF approach because the overall performance of case studies 3 and 4 are similar.

Omics integration to study the effect of chronic arsenic exposure on gene expression and DNA methylation

The previous results indicate that by using the M-value for DNA methylation and log2(intensities) for gene expression values, more subgroups can be identified within our population that are affected differently by arsenic exposure. Therefore, we carried out one final simulation where the input data sets is defined by the DNA methylation as M-values and gene expression as log2(intensities). A stable solution is found for k=6, for which the silhouette score lies above the threshold of 0.7. The results contained six clusters with their associated explanatory features (Figure 5.4A), with a mixture of exposure levels within each cluster (Figure 5.4B). In each cluster, there is a mixture of sex, age, and social-economic status and thus those factors do not drive the clustering. The patterns within the transcriptome and epigenome, and potentially the SNPs, are the driving power behind the clusters.

To deepen our understanding of the genomic profiles of these clusters, we extracted the features on the epigenome and transcriptome based on the multi-layer NMF results. Here, the top 1000 transcripts and DNA regions are selected to investigate the biological drivers behind the clustering.

The top 1000 transcripts contain 155 genes known to interact with arsenic and thus provide some validation in the identified patterns. Pathways associated with these genes are immune system, Cytokine-cytokine receptor interaction, signal transduction, signaling pathways regulating pluripotency of stem cells, PI3K-Akt signaling pathway, signaling by Rho GTPases, innate immune system, and Rho GTPase effector (Supplementary table 5.2).

Alterations in the DNA methylation patterns occur not only on multiple chromosomes but also in different DNA regions. Hyper- and hypomethylation of the different DNA regions happen both at the primary transcription and the transcription start sites, thus affect gene expression in different ways. Pathway analysis showed that the methylated genes are related to potential interesting processes (Supplementary table 5.3). These pathways represent different signaling pathways including Cytokine – Cytokine receptor interaction, chemokine signaling pathway, NF-kappa B signaling pathway, Toll-like receptor signaling pathway, and the mTOR signaling pathway. Note that the pathways related to the immune system and the cytokine-cytokine receptor interaction are both found in the pathways associated with the feature genes and feature DNA regions. However, there is only a small overlap in the epigenetic and transcriptomic features: *AKT1*, *MRPS17*, and *PRKAR1B* for the hypomethylated genes and *CD86*, and *GBP5* for the hypermethylated genes.

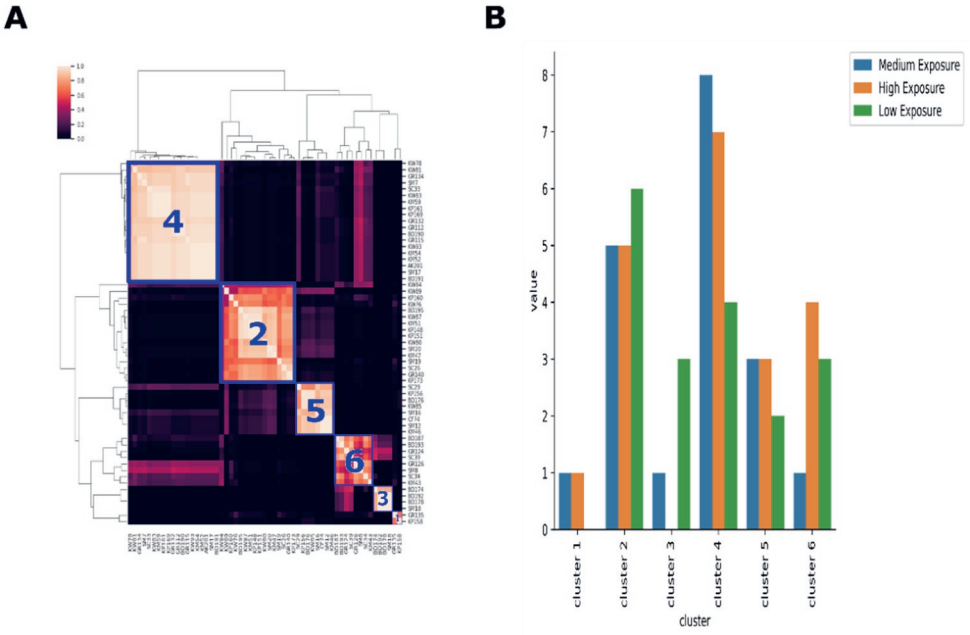


Figure 5.4 A: Consensus map for the six clusters obtained by integrating the epigenome and transcriptome. A green border highlights the clusters. **B:** The exposure groups per cluster. For each cluster, the exposure group of a member is retrieved and summed to gain an exposure group profile per cluster.

To study the interactions between methylated DNA regions and genes, we constructed a genomic interaction network. In this network, different communities are present containing feature genes and DNA regions related to signal processes, immune system processes, as well as metabolic-related processes (Figure 5.5). This information provides us with some guidance in identifying potential interesting genes and methylated DNA regions. Here, the most interesting processes are selected concerning arsenic exposure. Signaling pathways are important processes that regulate a variety of cellular processes and the disruption of those pathways can have adverse effects. The metabolic processes are of interest because arsenic exposure is associated with diabetes and cardiovascular diseases and the underlying metabolic alterations in cholesterol and vitamin and cofactors metabolism can lead to the development of those diseases.

When we analyze the SNP distribution (Figure 5.6), no significant difference is observed in the SNPs profiles. For most of the clusters, the presence of homozygote (+), homozygote (-) or heterozygote scores for the seven SNPs are equal over the clusters.

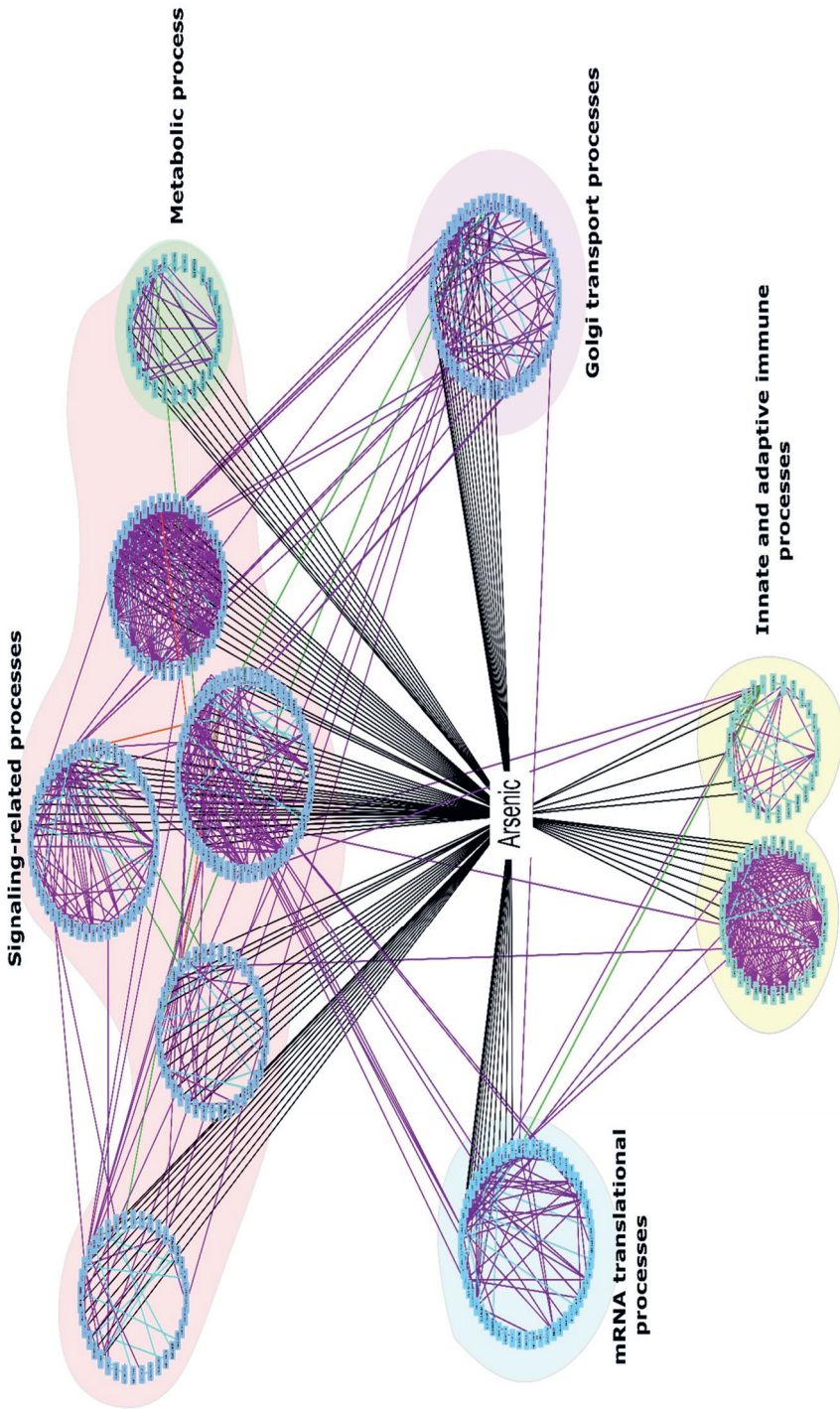


Figure 5.5 Communities grouped on their associated processes: signaling pathways include Cytokine-Cytokine receptor interaction, cytokine signaling, mTOR signaling, Interferon alpha/beta/gamma signaling, NOD-like signaling, IFNA signaling, JAK-STAT signaling, Chemokine signaling, FGFR3 and FGFR1 signaling. Metabolic processes (green) include cholesterol metabolism, metabolism of vitamins and cofactors, and insulin processes. Innate and adaptive immune system (blue) include neutrophil degranulation, Class 1 MHC mediated antigen processing, and ubiquitination of proteins. Golgi transport processes (yellow) and mRNA transformation processes (light blue) contain subgroups of the family process.

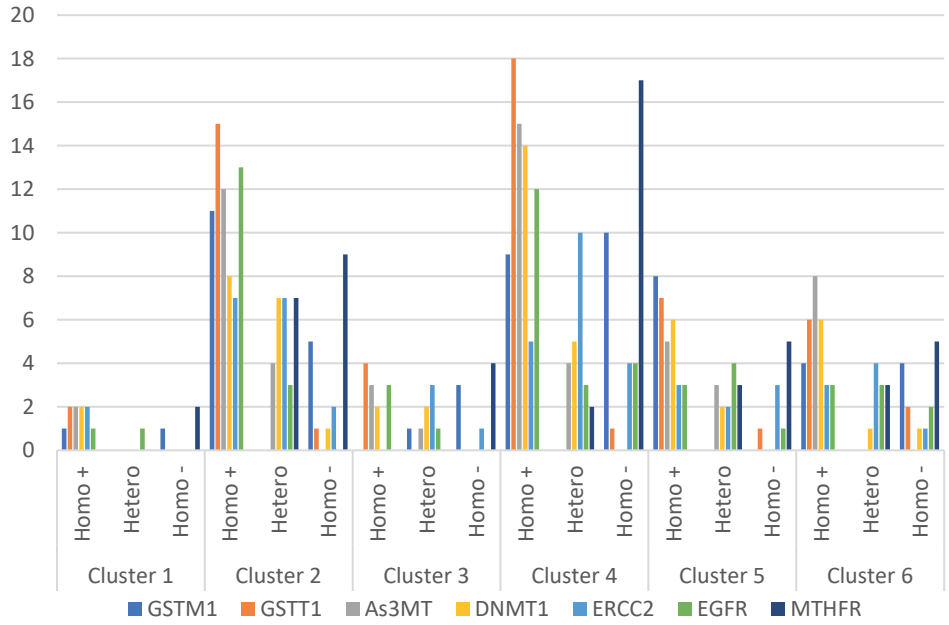


Figure 5.6 SNP distribution for the six clusters**:** for each cluster, we visualized the counts per allele type (homozygote positive, heterozygote, and homozygote negative). Note there is a strong mixture of the different allele types per cluster.

Discussion

It is known that to detect a small effect size, a large sample size is needed. Increasing the sample size is not always possible, especially if the data is already collected but also because field studies to gather material are time-consuming and costly. Therefore, increasing the signal with a computational approach such as identifying significant entities related to exposure before applying an omics integration technique would be one way to overcome the small effect size and high sample size. Here, we investigated different ways to improve omics integration in a small sample size cohort by adding extra selection steps before applying the multi-layer NMF approach.

To study the effect of data reduction by an additional filtering step, multiple case studies are defined. In the case studies, we tested if an a priori data selection based on LMMs is useful to increase the arsenic-exposure patterns in our data set and whether the β-value or M-value is a more appropriate measure for DNA methylation in our workflow. These different case studies provided us with valuable information: i) a priori feature filtering with too stringent criteria does not improve the multi-layer NMF approach, and ii) performing an omics integration strategy with M-values for DNA methylation resulted in more clusters detected. The sparsity in the M-value matrices might be of importance because the silhouette scores do not improve if the M-values are shifted into the positive range (data not shown). The power of the

M-value might thus be stored within the positive and negative values of the M-value distribution.

The semi-supervised multi-layer NMF approach did not provide us with any additional information compared to the baseline simulation. The maximum number of clusters detected is equal for both unsupervised and semi-supervised multi-layer NMF. Moreover, our results showed the effect of a too stringent a priori selection. Hence, the silhouette score decreased, and we are not capable of extracting more information from the data. The translation of DNA methylation values from a β-value to an M-value did improve the predictive power of the multi-layer NMF method. The integration of DNA methylation M-values in combination with gene expression values did result in more detected subgroups in our data. Not only did the number of clusters increase, an increase in the silhouette score and thus in cluster stability is observed. Furthermore, the use of the M-value does seem to improve the omics integration in such a way that it is not necessary to perform a semi-supervised multi-layer NMF.

The integration of the DNA methylation and gene expression data, with M-values for methylation, resulted in six clusters. These six clusters show different genomic profiles that stratify the subjects into the six clusters. Here, the top 1000 ranked features per omics layer are selected to understand which features are driving the clustering. The pathways associated with the gene and DNA regions relate to different signaling pathways. This is in agreement with known alterations in cellular signal transduction by arsenic exposure [47]. The alteration of cellular responses to hormones and growth factors and the inability to transduce signals might result in long-lasting pathological effects [48].

Various communities are present within our genomic interaction network related to signaling pathways, Golgi transport processes, mRNA transformation processes, metabolic processes, and innate and adaptive immune system. Moreover, some of the communities, and thus the pathways, are connected via protein-protein interactions. The interaction between the Golgi transport processes and signaling pathways might indicate that changes in signaling pathways, due to arsenic exposure, will also affect secretion by the Golgi apparatus. This link has already been proposed in the development of many diseases, including neurological disorders [49] and cancer cell metastasis [50] and disruption due to arsenic exposure might be associated with those diseases.

In the past decade, studies linked arsenic exposure to inflammatory cytokines in urothelial cell models and urine [51], the thymus [52], and human lymphocytes [53]. In the genomic interaction network, eight genes are identified to play a role in cytokine – cytokine receptor interactions and might be vital in the response to arsenic exposure. *CXCL9* is an important inflammatory gene and shares many protein-protein interactions with *CXCR2, CXCL3, CXCR1, CCL4,* and *CCR1*.

These interactions are all between chemokines and connect chemokine subfamilies CXC (chemotactic for neutrophils) and CC chemokines (chemotactic for monocytes and some lymphocytes) [54]. DNA methylation can induce an imbalance or dysregulation of the cytokine-cytokine receptor interactions and several methylated regions are found related to the CXC and CC chemokine family (*CCL1, CCL3L1, CCR1, CCL4, CCL18, CXCR1, CCL4L1, CCL4L2, CCL3L3, CXCR6*, and *CXCR2*). Furthermore, different interferon alpha units (*IFNA4, IFNA7, IFNA10*, and *IFNA16*), another class of cytokines, show different methylation patterns and could indicate a role for methylated interferon-alpha units in this pathway with relation to arsenic exposure. The interferon-alpha subunits play a vital role in both the innate and adaptive immune response [55].

Previous studies [17, 56–58] related arsenic exposure with multiple signaling pathways. In a Bangladesh cohort [56], arsenic exposure resulted in methylation of the NFkB signaling pathway and inflammatory responses. NFkB is an essential transcription modulator of chemokines *CXCL3*, *CXCL9*, and *CCL4* [59]. Research suggests a potential role for interferon-gamma and alpha genes [60] to interact with the NFkB signaling pathway. This could be of interest since we identified multiple interferons in our genomic interaction network. The central role for the NFkB signaling pathway, chemokines, and interferon genes could be an important mechanism underlying the adverse effects of arsenic exposure.

Disruption of metabolic pathways can lead to the development of diseases, including arsenic-related diseases cardiovascular disease, and diabetes. In vitro experiments show that arsenic inhibits key regulators of lipid homeostasis and as a result disrupts cholesterol clearance [61]. Here, the genes *APOE, APOA2, SCD1*, and *LRP12* play a potentially pivotal role in the six clusters (Figure 5.4A). APOE is a protein involved in lipid transport, atherosclerosis but also associated with neurodegenerative disorders [62]. It can modulate various cellular functions, including the functions of macrophages, suppress T-cell proliferation, inhibit the proliferation of smooth muscle cells but can also interact with cytokines [62]. SCD1, a stearoyl CoA desaturase 1, catalyzes the production of fatty acids, and alterations in SCD1 can lead to various effects on cellular function [63]. High SCD1 expression correlates with obesity and insulin resistance, whereas low expression is protective against those diseases [63, 64]. SCD1 is involved in the regulation of inflammation and stress in various cell types including macrophages, endothelial cells, and myocytes [63]. Furthermore, the complete loss of SCD1 leads to atherosclerosis, dermatitis, and intestinal colitis [65]. These interactions could connect alteration in metabolic processes with immune-related processes.

Our results give new insights into the perturbations of biological processes due to arsenic exposure. However, our results differ from the supervised single omics results with respect to the obtained groups. Whereas the single omics LMM strategy is a supervised method that identifies features related to low, medium, and

high arsenic exposure, our method is an unsupervised strategy that finds clusters based on the transcriptomic and epigenetic patterns. Since similar expression patterns are observed within a cluster, we might have to reconsider using the exposure levels as a label and take susceptibility into account. It might also be that arsenic exposure could introduce a dose-related and dose-independent effect. Therefore, a comparison is made to calculate the overlap in genes and DNA regions from the LMM and multi-layer NMF results. The LMM and multi-layer NMF features share 81 transcripts, as well as 310 unique transcripts (LMM output) and 122 unique transcripts (multi-layer NMF) that are known to have a relation with arsenic (CTD arsenic - gene interactions [66]). On the epigenome, 66 DNA regions overlap between the LMM and multi-layer NMF output, whereas there are 300 unique DNA regions (LMM output) and 173 unique DNA regions (multi-layer NMF) for which the corresponding gene shared interaction with arsenic (CTD arsenic – gene interactions [66]). These results indicate that the two approaches are complementary and can both help us to unravel arsenic-related disease onset.

There are some limitations to the current study. To start, the sample size is rather small and therefore we cannot perform a GWAS study to measure all possible polymorphisms. In our case, the risk allele score is only calculated for seven SNPs since other SNPs are not measured. Although the investigated polymorphisms have been selected based on their role in arsenic metabolism, they do not directly relate to the other biological processes affect by arsenic metabolism. Furthermore, the small sample size makes it harder to detect the arsenic-related patterns in our data set, especially since subjects could have a different response to arsenic exposure. To overcome this problem, redundant features are removed and DNA methylation values are expressed by the M-value. The M-value showed to be more discriminative between groups, and as a result, more different patterns could be identified within our data. These groups show changes in methylation and gene expression related to key signaling processes, immune response, and metabolic processes that might explain the adverse health effects of arsenic exposure.

Conclusion

When dealing with a small data set to apply omics integration, the use of the DNA methylation M-value can improve the number of detectable clusters. By using the M-value instead of the β-value, more cluster are identified.. This integration of the transcriptome and epigenome lead to a genomic interaction network with communities related to several important processes including signaling pathways, mRNA translation processes, Golgi transport, metabolic transport, and innate and adaptive immune processes. Moreover, their association with arsenic exposure as well as their relationship with cardiovascular disease and diabetes makes these communities highly relevant. These different genomic profiles do not correlate to exposure level and might be driven by susceptibility to arsenic exposure. However, we are unable to detect a possible relation between the SNP risk allele scores and the integrated omics data.

References

1. Naujokas MF, Anderson B, Ahsan H, Aposhian HV, Graziano JH, Thompson C, et al. The Broad Scope of Health Effects from Chronic Arsenic Exposure: Update on a Worldwide Public Health Problem. Environ Health Perspect. 2013;121:295–302. doi:10.1289/ehp.1205875.
2. Scientific Opinion on Arsenic in Food. EFSA J. 2009;7:1351. doi:10.2903/j.efsa.2009.1351.
3. Xue J, Zartarian V, Wang S-W, Liu S V., Georgopoulos P. Probabilistic Modeling of Dietary Arsenic Exposure and Dose and Evaluation with 2003–2004 NHANES Data. Environ Health Perspect. 2010;118:345–50. doi:10.1289/ehp.0901205.
4. Tapio S, Grosche B. Arsenic in the aetiology of cancer. Mutat Res Mutat Res. 2006;612:215–46. doi:10.1016/j.mrrev.2006.02.001.
5. Argos M, Ahsan H, Graziano JH. Arsenic and human health: epidemiologic progress and public health implications. Rev Environ Health. 2012;27. doi:10.1515/reveh-2012-0021.
6. Smith AH, Hopenhayn-Rich C, Bates MN, Goeden HM, Hertz-Picciotto I, Duggan HM, et al. Cancer risks from arsenic in drinking water. Environ Health Perspect. 1992;97:259–67. doi:10.1289/ehp.9297259.
7. Chen C-J, Chen C, Wu M-M, Kuo T-L. Cancer potential in liver, lung, bladder and kidney due to ingested inorganic arsenic in drinking water. Br J Cancer. 1992;66:888–92. doi:10.1038/bjc.1992.380.
8. Chen Y, Graziano JH, Parvez F, Liu M, Slavkovich V, Kalra T, et al. Arsenic exposure from drinking water and mortality from cardiovascular disease in Bangladesh: prospective cohort study. BMJ. 2011;342 may05 2:d2431–d2431. doi:10.1136/bmj.d2431.
9. Yuan Y, Marshall G, Ferreccio C, Steinmaus C, Selvin S, Liaw J, et al. Acute Myocardial Infarction Mortality in Comparison with Lung and Bladder Cancer Mortality in Arsenic-exposed Region II of Chile from 1950 to 2000. Am J Epidemiol. 2007;166:1381–91. doi:10.1093/aje/kwm238.
10. Sanchez TR, Perzanowski M, Graziano JH. Inorganic arsenic and respiratory health, from early life exposure to sex-specific effects: A systematic review. Environ Res. 2016;147:537–55. doi:10.1016/j.envres.2016.02.009.
11. Tsai S-Y, Chou H-Y, The H-W, Chen C-M, Chen C-J. The Effects of Chronic Arsenic Exposure from Drinking Water on the Neurobehavioral Development in Adolescence. Neurotoxicology. 2003;24:747–53. doi:10.1016/S0161-813X(03)00029-9.
12. Calderón J, Navarro ME, Jimenez-Capdeville ME, Santos-Diaz MA, Golden A, Rodriguez-Leyva I, et al. Exposure to Arsenic and Lead and Neuropsychological Development in Mexican Children. Environ Res. 2001;85:69–76. doi:10.1006/enrs.2000.4106.
13. Gerr F, Letz R, Ryan PB, Green RC. Neurological effects of environmental exposure to arsenic in dust and soil among humans. Neurotoxicology. 2000;21:475–87. http://www.ncbi.nlm.nih.gov/pubmed/11022857.
14. Ebert F, Weiss A, Bültemeyer M, Hamann I, Hartwig A, Schwerdtle T. Arsenicals affect base excision repair by several mechanisms. Mutat Res Mol Mech Mutagen. 2011;715:32–41. doi:10.1016/j.mrfmmm.2011.07.004.
15. Shi H, Shi X, Liu KJ. Oxidative mechanism of arsenic toxicity and carcinogenesis. Mol Cell Biochem. 2004;255 1/2:67–78. doi:10.1023/B:MCBI.0000007262.26044.e8.
16. Valenzuela OL, Borja-Aburto VH, Garcia-Vargas GG, Cruz-Gonzalez MB, Garcia-Montalvo EA, Calderon-Aranda ES, et al. Urinary Trivalent Methylated Arsenic Species in a Population Chronically Exposed to Inorganic Arsenic. Environ Health Perspect. 2005;113:250–4. doi:10.1289/ehp.7519.
17. Rehman MYA, van Herwijnen M, Krauskopf J, Farooqi A, Kleinjans JCS, Malik RN, et al. Transcriptome responses in blood reveal distinct biological pathways associated with arsenic exposure through drinking water in rural settings of Punjab, Pakistan. Environ Int. 2020;135:105403. doi:10.1016/j.envint.2019.105403.
18. Rehman MYA, van Herwijnen M, Krauskopf J, Jennen DGJ, Briedé JJ, Malik RN, et al. Integrating SNPs-based genetic risk factor with blood epigenomic response of differentially Arsenic-exposed rural subjects reveals new disease-associated pathways.
19. Andrew AS, Jewell DA, Mason RA, Whitfield ML, Moore JH, Karagas MR. Drinking-Water Arsenic Exposure Modulates Gene Expression in Human Lymphocytes from a U.S. Population. Environ Health Perspect. 2008;116:524–31. doi:10.1289/ehp.10861.
20. Hunt KM, Srivastava RK, Elmets CA, Athar M. The mechanistic basis of arsenicosis: Pathogenesis of skin cancer. Cancer Lett. 2014;354:211–9. doi:10.1016/j.canlet.2014.08.016.
21. Chervona Y, Hall MN, Arita A, Wu F, Sun H, Tseng H-C, et al. Associations between Arsenic Exposure and Global Posttranslational Histone Modifications among Adults in Bangladesh. Cancer Epidemiol Biomarkers Prev. 2012;21:2252–60. doi:10.1158/1055-9965.EPI-12-0833.

22. Simeonova PP, Luster MI. Mechanisms of arsenic carcinogenicity: genetic or epigenetic mechanisms? J Environ Pathol Toxicol Oncol. 2000;19:281–6. http://www.ncbi.nlm.nih.gov/pubmed/10983894.
23. Jensen TJ, Wozniak RJ, Eblin KE, Wnek SM, Gandolfi AJ, Futscher BW. Epigenetic mediated transcriptional activation of WNT5A participates in arsenical-associated malignant transformation. Toxicol Appl Pharmacol. 2009;235:39–46. doi:10.1016/j.taap.2008.10.013.
24. Hubaux R, Becker-Santos DD, Enfield KS, Rowbotham D, Lam S, Lam WL, et al. Molecular features in arsenic-induced lung tumors. Mol Cancer. 2013;12:20. doi:10.1186/1476-4598-12-20.
25. Miao Z, Wu L, Lu M, Meng X, Gao B, Qiao X, et al. Analysis of the transcriptional regulation of cancer-related genes by aberrant DNA methylation of the cis-regulation sites in the promoter region during hepatocyte carcinogenesis caused by arsenic. Oncotarget. 2015;6:21493–506. doi:10.18632/oncotarget.4085.
26. Ren X, McHale CM, Skibola CF, Smith AH, Smith MT, Zhang L. An Emerging Role for Epigenetic Dysregulation in Arsenic Toxicity and Carcinogenesis. Environ Health Perspect. 2011;119:11–9. doi:10.1289/ehp.1002114.
27. Breda SGJ Van, Claessen SMH, Lo K, Herwijnen M Van, Gaj S, Kok TMCM De, et al. Epigenetic mechanisms underlying arsenic - associated lung carcinogenesis. Arch Toxicol. 2015;:1959–69. doi:10.1007/s00204-014-1351-2.
28. Breton C V., Zhou W, Kile ML, Houseman EA, Quamruzzaman Q, Rahman M, et al. Susceptibility to arsenic-induced skin lesions from polymorphisms in base excision repair genes. Carcinogenesis. 2007;28:1520–5. doi:10.1093/carcin/bgm063.
29. Ketelslegers HB, Godschalk RW, Gottschalk RW, Knaapen AM, Koppen G, Schoeters G, et al. Prevalence of at-risk genotypes for genotoxic effects decreases with age in a randomly selected population in Flanders: a cross sectional study. Environ Heal. 2011;10:85. doi:10.1186/1476-069X-10-85.
30. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587. doi:10.1186/1471-2105-11-587.
31. Zhuang JJ, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinformatics. 2012;13:59. doi:10.1186/1471-2105-13-59.
32. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89. doi:10.1016/j.molcel.2010.05.004.
33. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13:966–7. doi:10.1038/nmeth.4077.
34. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 2019;47:D948–54. doi:10.1093/nar/gky868.
35. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13. doi:10.1093/nar/gky1131.
36. Souza TM, Rieswijk L, Beucken T van den, Kleinjans J, Jennen D. Persistent transcriptional responses show the involvement of feed-forward control in a repeated dose toxicity study. Toxicology. 2017;375:58–63. doi:10.1016/j.tox.2016.10.009.
37. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008:P10008. doi:10.1088/1742-5468/2008/10/P10008.
38. Dangleben NL, Skibola CF, Smith MT. Arsenic immunotoxicity: a review. Environ Heal. 2013;12:73. doi:10.1186/1476-069X-12-73.
39. Andenæs K, Lunde IG, Mohammadzadeh N, Dahl CP, Aronsen JM, Strand ME, et al. The extracellular matrix proteoglycan fibromodulin is upregulated in clinical and experimental heart failure and affects cardiac remodeling. PLoS One. 2018;13:e0201422. doi:10.1371/journal.pone.0201422.
40. Sharma A, Masri J, Jo OD, Bernath A, Martin J, Funk A, et al. Protein Kinase C Regulates Internal Initiation of Translation of the GATA-4 mRNA following Vasopressin-induced Hypertrophy of Cardiac Myocytes. J Biol Chem. 2007;282:9505–16. doi:10.1074/jbc.M608874200.
41. Morice-Picard F, Benard G, Rezvani HR, Lasseaux E, Simon D, Moutton S, et al. Complete loss of function of the ubiquitin ligase HERC2 causes a severe neurodevelopmental phenotype. Eur J Hum Genet. 2017;25:52–8. doi:10.1038/ejhg.2016.139.
42. Xu H, Wang Z, Sun Z, Ni Y, Zheng L. GATA4 protects against hyperglycemia-induced endothelial dysfunction by regulating NOX4 transcription. Mol Med Rep. 2017. doi:10.3892/mmr.2017.8062.

43. Czemplik M, Kulma A, Wang YF, Szopa J. Therapeutic Strategies of Plant-derived Compounds for Diabetes Via Regulation of Monocyte Chemoattractant Protein-1. Curr Med Chem. 2017;24. doi:10.2174/0929867324666170303162935.
44. Hannenhalli S, Kaestner KH. The evolution of Fox genes and their role in development and disease. Nat Rev Genet. 2009;10:233–40. doi:10.1038/nrg2523.
45. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, et al. Zinc-finger proteins in health and disease. Cell Death Discov. 2017;3:17071. doi:10.1038/cddiscovery.2017.71.
46. Kamachi Y, Kondoh H. Sox proteins: regulators of cell fate specification and differentiation. Development. 2013;140:4129–44. doi:10.1242/dev.091793.
47. Druwe IL, Vaillancourt RR. Influence of arsenate and arsenite on signal transduction pathways: an update. Arch Toxicol. 2010;84:585–96. doi:10.1007/s00204-010-0554-4.
48. Nicotera P, Dybing E. Alterations in Cell Signaling and Cytotoxicity. In: Pharmacological Sciences: Perspectives for Research and Therapy in the Late 1990s. Basel: Birkhäuser Basel; 1995. p. 447–52. doi:10.1007/978-3-0348-7218-8_45.
49. Caracci MO, Fuentealba LM, Marzolo M-P. Golgi Complex Dynamics and Its Implication in Prevalent Neurological Disorders. Front Cell Dev Biol. 2019;7. doi:10.3389/fcell.2019.00075.
50. Millarte V, Farhan H. The Golgi in Cell Migration: Regulation by Signal Transduction and Its Implications for Cancer Cell Metastasis. Sci World J. 2012;2012:1–11. doi:10.1100/2012/498278.
51. Liu S, Sun Q, Wang F, Zhang L, Song Y, Xi S, et al. Arsenic Induced Overexpression of Inflammatory Cytokines Based on the Human Urothelial Cell Model in Vitro and Urinary Secretion of Individuals Chronically Exposed to Arsenic. Chem Res Toxicol. 2014;27:1934–42. doi:10.1021/tx5002783.
52. Gera R, Singh V, Mitra S, Sharma AK, Singh A, Dasgupta A, et al. Arsenic exposure impels CD4 commitment in thymus and suppress T cell cytokine secretion by increasing regulatory T cells. Sci Rep. 2017;7:7140. doi:10.1038/s41598-017-07271-z.
53. Zarei MH, Pourahmad J, Nassireslami E. Toxicity of arsenic on isolated human lymphocytes: The key role of cytokines and intracellular calcium enhancement in arsenic-induced cell death. Main Gr Met Chem. 2019;42:125–34. doi:10.1515/mgmc-2019-0014.
54. Palomino DCT, Marti LC. Chemokines and immunity. Einstein (São Paulo). 2015;13:469–73. doi:10.1590/S1679-45082015RB3438.
55. TOUGH DF. Type I Interferon as a Link Between Innate and Adaptive Immunity through Dendritic Cell Stimulation. Leuk Lymphoma. 2004;45:257–64. doi:10.1080/1042819031000149368.
56. Demanelis K, Argos M, Tong L, Shinkle J, Sabarinathan M, Rakibuz-Zaman M, et al. Association of Arsenic Exposure with Whole Blood DNA Methylation: An Epigenome-Wide Study of Bangladeshi Adults. Environ Health Perspect. 2019;127:057011. doi:10.1289/EHP3849.
57. GHOSH P, BANERJEE M, GIRI A, RAY K. Toxicogenomics of arsenic: Classical ideas and recent advances. Mutat Res Mutat Res. 2008;659:293–301. doi:10.1016/j.mrrev.2008.06.003.
58. Bustaffa E, Stoccoro A, Bianchi F, Migliore L. Genotoxic and epigenetic mechanisms in arsenic carcinogenicity. Arch Toxicol. 2014;88:1043–67. doi:10.1007/s00204-014-1233-7.
59. Richmond A. NF-κB, chemokine gene transcription and tumour growth. Nat Rev Immunol. 2002;2:664–74. doi:10.1038/nri887.
60. Pfeffer LM. The Role of Nuclear Factor κB in the Interferon Response. J Interf Cytokine Res. 2011;31:553–9. doi:10.1089/jir.2011.0028.
61. Rui L. Energy Metabolism in the Liver. In: Comprehensive Physiology. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2014. p. 177–97. doi:10.1002/cphy.c130024.
62. Zhang H, Wu L-M, Wu J. Cross-Talk between Apolipoprotein E and Cytokines. Mediators Inflamm. 2011;2011:1–10. doi:10.1155/2011/949072.
63. Liu X, Strable MS, Ntambi JM. Stearoyl CoA Desaturase 1: Role in Cellular Inflammation and Stress. Adv Nutr. 2011;2:15–22. doi:10.3945/an.110.000125.
64. García-Serrano S, Moreno-Santos I, Garrido-Sánchez L, Gutierrez-Repiso C, García-Almeida JM, García-Arnés J, et al. Stearoyl-CoA Desaturase-1 Is Associated with Insulin Resistance in Morbidly Obese Subjects. Mol Med. 2011;17:273–80. doi:10.2119/molmed.2010.00078.
65. Brown JM, Chung S, Sawyer JK, Degirolamo C, Alger HM, Nguyen T, et al. Inhibition of Stearoyl-Coenzyme A Desaturase 1 Dissociates Insulin Resistance and Obesity From Atherosclerosis. Circulation. 2008;118:1467–75. doi:10.1161/CIRCULATIONAHA.108.793182.
66. CTD. Curated chemical-disease data were retrieved from the Comparative Toxicogenomics Database (CTD), North Carolina State University, Raleigh, NC and Mount Desert Island Biological Laboratory, Salisbury Cove, Maine. World Wide Web. 31-08-2017. 2017. http://ctdbase.org/

Chapter 6
Summary and General Discussion

To study and understand the mechanisms and characteristics of exposure-induced toxicity or cancer development, a robust analysis utilizing computational methods is crucial to understand the induced perturbations. As discussed earlier in this thesis, a single-omics data set will only capture information of one type of biomolecules and thus represents a small subset of the biological cascade. Therefore, we need to integrate multiple layers of information to provide a systemic understanding of exposure-induced toxicity or cancer development. It is intuitive to integrate omics datasets: various omics measurements derived from one biological system draw the complete picture of that system. Modern molecular research has revealed important insights, not only within one layer of information but also between multiple integrated omics layers by studying the interactions. However, human genomes are complex, and incorporating different layers of biological data to predict phenotypes is not straightforward. This is highlighted by the fact that omics integration is already an ongoing line of research since the nineties [1,2]. Therefore, we explored different strategies to integrate omics data sources as well as to integrate known information to increase our understanding of these complex systems.

There are different strategies to integrate multi-omics data, including knowledge-to-knowledge integration, data-to-knowledge integration, and data-to-data integration. Each integration approach has its advantages and disadvantages. Knowledge-to-knowledge integration integrates knowledge across different omics layers to gain a better understanding of a system but does not integrate new data. Data-to-knowledge integration does integrate new data but does integrate this data on an existing model. This approach relies existing mathematical models, which is are always available in the field of toxicogenomics. A data-to-data approach does integrate the different omics layers without a priori knowledge and is driven by the patterns in the data. We hypothesized that a data-to-data-driven approach will give novel insights in a biological system compared to a knowledge-based approach since our knowledge of biology is mostly hidden in our data. However, as we will discuss later, we will combine the data-to-data-driven approach with a knowledge integration step because existing knowledge could enhance our understanding of the observed results.

To perform a data-to-data approach, there are different classes of algorithms one can select: matrix factorization, correlation-based, Bayesian methods, network-based methods, multiple kernel learning, and multi-step analysis. Because of the huge amount of options, selecting the best algorithm can be a challenge. One of the most important choices is to choose either an unsupervised or a supervised data integration method. An unsupervised data integration method will learn the patterns from the data, whereas a supervised data integration method relies on a label, as guidance for finding patterns in the data. We have used an unsupervised data integration method since we did not want to rely on a label to learn structures within our data, but also because strong labels representing a clear phenotypic

endpoint are not always available in exposure studies. This leaves us with three unsupervised learning classes: matrix factorization, correlation-based and Bayesian methods. To integrate our omics data sets, we want to perform a simultaneous omics integration approach that reduces the dimension of the data (thus removes irrelevant features) and stratifies the subjects into clusters. Therefore, we have looked at matrix factorization techniques and selected the nonnegative matrix factorization (NMF) method as our mathematical model for omics integration. NMF has been proven to not only extract patterns from single omics layers [3–5] but also from multiple omics layers with an integrated NMF approach [6,7]. One major advantage of NMF is that it can handle imbalanced data sets. This is important because, with the current omics techniques, we can measure tens of thousands of probes, but it is unfeasible to measure the same number of samples. Furthermore, by setting constraints on our matrix calculations, we can combine omics data sets that have a different underlying data distribution. As one of the constraints for our model, we have set the Kullback-Leibler divergence as an error function instead of the classical error function. The Kullback-Leibler divergence (KL divergence) is based on log-differences, whereas the standard error function minimizes the Euclidean distance, which is based on square-differences [8]. The KL divergence measures the distance between distributions and decreases when the distance between our predicted distribution and real data is decreasing. The KL divergence is shown to perform well in biological research [9,10] and the translation to a multi-layer divergence is beneficial for the current NMF tool.

The multi-omics NMF method will provide us with a list of features, in our case, transcriptomic features and epigenetic features. After extracting patterns from the data, the next challenge is to gain relevant biological information from those patterns. Here, we have applied a combination of a data-to-data approach (multi-layer NMF) and the data-to-knowledge approach by constructing genomic interaction networks. We build these genomic interaction networks from the features identified by the multi-layer NMF method (data-to-data), and curated information from different databases (knowledge-to-data). A genomic interaction network is a great way to analyze genomic relationships but also to visualize these important interactions. These networks represent snapshots of a cellular system and capture transitions between different states due to perturbations induced by gene mutations, disease development, or environmental factors including compound exposure.

In **chapter 2.1**, we developed a network visualization tool to visualize dynamic changes in biological networks. Here, we focused on integration dynamic changes on the two building blocks of a network: nodes and edges. In a dynamic biological system, we observe changes in expression of biomolecules. We visualized this change in expression on a node by applying a color transition based on the expression value. This way, we can study the effect of an increase or decrease in

expression values over time or state. However, these changes are not the only important ones happening in a cellular system. Interactions between biological entities can appear or disappear and this gives important information as well. Therefore, we have also integrated a visualization to capture dynamic edges in our networks. Both phenomena are important since the high expression of one node could lead to a downstream effect in the network, whereas a disappearing interaction might influence other gene-gene or protein-protein interaction. Furthermore, we have integrated another data-to-knowledge approach in which we link node information (genes) to different databases such as Genecards and NCBI but also pathway and disease information. By highlighting this information on a node, the user can directly link alterations of that node to pathways or diseases. Knowledge integration is important since the interpretation of a network heavily relies on the interpretation by the user [11]. By allowing the user to study dynamic effects, highlight connected nodes, and providing information about each node, we help to improve the clarity and completeness of the network. One of the major challenges in network visualization is the high number of edges and/or nodes. These very dense networks, also called hairball networks, are difficult to analyze because they contain too much information. DynOVis tried to overcome this problem by visualizing the network in 3D. However, this seems not a solution to the hairball problem since it adds complexity to the analysis of the network. Whereas we can see the full network in 2D, our field of view is limited in 3D which makes it harder to study patterns in a network. DynOVis could benefit from additional modules with novel network representation algorithms. Furthermore, a stand-alone version like the program Cytoscape could improve the overall performance.

In **chapter 2.2**, we developed an API framework to construct genomic interaction networks, in which we used the transcriptomic and epigenetic features as the building blocks for these networks. These genomic interaction networks store information about the relationships within an omics layer (relation between genes) but also the relation across omics layers (DNA methylation and genes). This approach allowed us to analyze the downstream effect of hypo- or hypermethylated genes. For example, a hypermethylated transcription factor cannot transcriptionally activate its target, and thus transcriptional silencing could be observed. The interactions within the genomic interact networks are important since we use them to gain a better understanding of our multi-omics model. Therefore, we have defined a set of rules to filter out trivial interactions. We only add interactions if they are curated or have a confidence score above a threshold (threshold = 0.7). A high confidence score reflects the high probability of an interaction happening in our cells and thus, if they are perturbed, can explain a phenotypic endpoint. This selection criterion does reduce the number of noncausal interactions and increases causal relationships between the different entities.

In this thesis, we proposed an omics integration workflow that applies an NMF method to extract relevant omics patterns, which we further integrate with our

genomic interaction networks. In our work, we show that it is beneficial to combine data-to-data with existing knowledge to increase our understanding of a biological system. First, we tested this workflow on cancer in vitro data sets to identify if our workflow can simultaneously detect transcriptomic and epigenetic patterns that translate into a genomic interaction network. We used two in vitro cancer cell line data sets because these data sets are widely applied to test new methods.

First, we applied our omics integration workflow to identify transcriptomic and epigenetic patterns in the NCI-60 human tumor cell line data (**chapter 3.1**). The combination of the transcriptome and epigenome resulted in four clusters build with different genomic profiles. One of those clusters showed a strong profile for melanoma cell lines, in which we identified melanoma-enriched genes related to drug resistance. These findings are of interest because melanoma is the most aggressive form of skin cancer with a poor prognosis due to resistance to conventional chemotherapy [12]. In our genomic interaction network, we identified a central role for *MITF*, a transcription factor amplified in 20% of melanoma cases that results in a reduced 5-year survival [13]. *MITF* is a member of the *IRF4-ABCB5-MITF* axis in our network, thus showing a possible relation between the three genes that all are associated with drug resistance in melanoma.

The results described in **chapter 3.1** showed that with a multi-omics integration approach in combination with network analysis and knowledge integration, we could extract new information that increased our understanding of cancer cell lines. This case study showed that our proposed omics integration network can help us to gain a better understanding of the role between the transcriptome and epigenome. However, this data set is rather small and only contains nine tumor types. Therefore, it was of interest to identify if we could extract genomic profiles in data sets that contain more samples but also more different tumor types (**chapter 3.2**). The 2019 Cancer Cell Line Encyclopedia (CCLE) contains many samples derived from various tumor types. In this study, we were capable of separating clusters based on their DNA methylation and gene expression profiles across a wide range of cell lines derived from multiple human cancer types, and we have identified potential candidate genes that characterize cancer cell lines of the type for lymphoid and hematopoietic neoplasms.

In the lymphoid and hematopoietic neoplasms cancer cell lines, we identified multiple methylated DNA promoter regions that affect transcriptional activation of *EGFR*, which might affect the protein-protein interactions with the oncogenes *FGR* and *PTK2*. The DNA hypermethylation of *EGFR* could be of interest since this gene is linked to drug resistance [14]. DNA hypermethylated of *YAP1* is a second important characteristic because hypermethylation did lead to low gene expression and consequently no transcriptional activation of *JAG1*. Although *YAP1* has tumor-suppressive characteristics, it is relevant to highlight that reversing the methylation status of *YAP1* may lead to *YAP1* transcriptionally activating the oncogene *JAG1*.

Finally, we could identify a potential dysregulation of the *CD28-CD86-CTLA4* axis in the different lymphoid neoplasm cancer cell lines. Our current insights show that, although assumed different, B and T cell lymphomas potentially share similar genomic alterations. These key alterations are important to understand the development and progression of lymphoid and hematopoietic neoplasms.

The work in **chapter 3** showed that we can retrieve valuable insights in the genomic profiles for the two in vitro cancer cell line data sets. Our unsupervised omics integration approach does extract relevant patterns on the transcriptome and epigenome describing different clusters within a data set. In **chapter 4** and **chapter 5**, we will use in vivo data to investigate the possible effect of compound exposure on the transcriptome and epigenome.

In **chapter 4**, we investigated the role of persistent organic pollutants and heavy metals on the transcriptome and epigenome of subjects from two population studies: the Northern Sweden Health & Disease Study (NSHDS) cohort and the EPIC-Italy cohort. Persistent organic pollutants (POPs) and heavy metals are toxic compounds with established adverse effects on human health. We applied a multi-layer nonnegative matrix factorization method to integrate gene expression and DNA methylation profiles with blood levels of POPs including PCBs, DDE, and HCB, as well as of lead and cadmium to study the adverse health effects. We could stratify the subjects into three clusters based on their omics profiles (combined cohort: EPIC-Italy and NSHDS cohorts) and found distinct exposure profiles for DDE and lead. It became clear that POPs and heavy metal exposure are involved in various biological processes including signaling processes, immune-related processes, DNA repair, mRNA translation, and mRNA transcription. Due to a possible cohort-specific effect, we performed two additional simulations, one for each cohort. The simulation on the NSHDS cohort revealed a potential sex-specific exposure-response of male gamete formation in the Swedish males, as well as a sex-specific exposure-response of cytochrome P450 genes and bile acid synthesis in the Swedish females. In the EPIC-Italy cohort, we identified methylation differences for sex-specific CpG sites associated with lead exposure, which might point towards a sex-specific exposure-related effect. These CpG sites are associated with mantle cell lymphoma and thus could connect lead exposure with the onset of this disease.

The data from the NSHDS and EPIC-Italy cohorts in **chapter 4** showed why an unsupervised learning method is important. Data for gene expression and DNA methylation is measured in blood taken from healthy individuals. In this case, there are no clear disease/healthy labels since all subjects are considered healthy at the start of the study. Thus, we needed to apply a computational method that can ‘blindly’ learn patterns from the data to understand exposure-related health effects. We hypothesized that we need a large data set to capture exposure-related patterns and therefore we combined the NSHDS and EPIC-Italy cohorts. The

combination of these two cohorts did introduce a potential problem: cohort bias in the data. Therefore, we have applied an additional filtering step before we integrate the omics layers. Filtering methods rely on the characteristics of the variables and are a powerful tool to eliminate irrelevant, redundant, or constant features. This is important in most biological studies because there are many features we can measure, and not all of them relate to the phenotype we would like to investigate. To remove the cohort-specific bias, we performed another filtering step: Analysis Of Variance (ANOVA) univariate test. ANOVA assumes a linear relationship between the variables and the outcome (i.e. gene expression and cohort) and is well-suited for continuous variables with a binary target. We applied an ANOVA filtering to remove only those genes stratifying the subjects based on cohort and not on any of the other variables (For example exposure or sex).

This study provides insights into exposure-related changes in the transcriptome and epigenome. By combining the EPIC and NSHDS cohorts, we were able to identify communities in our interaction network that relate to POP and lead exposure that could explain the development of certain diseases. In general, we could not identify clear epigenetic alterations that directly associate with observed changes in the transcriptome. However, the combination of the epigenetic and transcriptomic alterations provided us with genomic profiles that relate to the exposure profiles.

We know that epigenetic mechanisms are not the only events that regulate transcription. It is now established that noncoding regulatory variants play a central role in the development of diseases [15]. Regulatory elements are non-coding DNA sequences that have a critical role in controlling gene expression [16] and therefore could be the knowledge gap in the connection between CpG methylation and gene expression. Genetic variants cannot only change gene expression but also the susceptibility of a subject to exposure. Because of their high prevalence in the general population, genetic variants that determine susceptibility to environmental exposures may contribute to the development of exposure-related diseases [17]. Genome-wide association studies (GWAS) have shown a polygenetic architecture within common disorders and have enabled researchers to identify genetic variants associated with diseases. This genetic architecture consists of the genetic phenotypic basis, including the epigenome, transcriptome, and its genetic variants. These genetic variants could have been of importance in chapter 4 since susceptibility to lead and POP toxicity depends on an individual's variability determined by genetic variation due to polymorphisms.

In **chapter 5**, we studied the role of certain polymorphisms in the transcriptomic and epigenetic response to arsenic exposure. These polymorphisms, in this case, single nucleotide polymorphisms (SNP), could explain the difference is susceptibility to arsenic exposure. We used our multi-layer NMF approach to integrate the transcriptomic and epigenetic data to identify clusters with different

risk allele scores for each SNP. We defined these risk allele scores by considering the polymorphisms for genes related to arsenic metabolism.

During the integration of the omics data, we ran into problems with one common denominator: the sample size. The transcriptomics and epigenomics integrated did not result in cluster with a clear effect of arsenic exposure. As mentioned, we believed the sample size of this study could be a problem. To identify specific profiles on the transcriptome and epigenome, we need either a strong and distinguishable signal or an appropriate sample size to detect smaller signals. In **chapter 4**, we could reveal exposure-related effects within our data because of the large sample size. In **chapter 5**, the sample size might be too small and therefore our method identifies patterns not related to exposure. Here, it also becomes clear there is a potential difference between in vitro and in vivo data. In **chapter 3**, we could identify cancer-related profiles both in large (CCLE) and smaller (NCI60) cancer studies. One explanation is the different nature of the data: in vitro versus in vivo data. In vitro data, in most cases, has less sample variability that only occurs across and not within tissue types [18]. In vitro data contains a larger sample variability that can make it harder to identify phenotype-related alterations on the transcriptome and epigenome. This is even more problematic in data set with low sample sizes with smaller exposure-related effects. The identification and extraction of relevant patterns strongly depend on the signal-to-noise ratio (SNR). If the signal of the relevant markers, for example, CpG sites, is strong, we expect the multi-layer approach to identify the signal correctly. However, when the signal strength is low/moderate it could be possible that the relevant markers are lost within the noise

The easiest solution is to increase the sample size, but this will automatically lead to higher experimental costs and is not always feasible. We explored different possibilities to enhance arsenic-related exposure patterns without increasing the sample size. In **chapter 5**, we investigated two approaches which we believed would increase the exposure-related patterns: i) semi-supervised multi-layer NMF, and ii) M value for DNA methylation. First, we tried the semi-supervised multi-layer NMF approach, in which we selected CpG sites and gene probes from two studies performed by Rehman et al. We hypothesized that this selection step would increase the arsenic-related patterns and thus would give clusters with different exposure and risk allele profiles. When we investigated the obtained clusters, we could not identify any exposure-related effects. This might be because the FDR selection is based on a linear model with predesigned groups based on arsenic levels in urine. The combination of the omics layers might not reflect those predesigned groups but might highlight other arsenic-related effects.

With our second approach, we could identify more clusters by using M-values for DNA methylation instead of the commonly used β-values. Previous research suggested that the M-value stores more information and has a higher statistical

power [19,20]. This might be an important characteristic of the M-value, especially for studies with small sample sizes. Because the M-value can be positive or negative, the positivity constraint of NMF enforces us to split the M-value matrix into a matrix of positive M-values and a matrix of the absolute values of the negative M-values. The introduction of the two matrices introduces sparsity into our data, which is advantageous because sparsity can improve the interpretability of the patterns and thus dimension reduction and clustering. In cohort studies with a small sample size, it can therefore be beneficial to transform the β-values to M-values.

The integration of the gene expression (log2(intensity) and DNA methylation (M-values) resulted in six clusters. In the genomic interaction network, different communities are identified with transcripts and CpG sites explaining the different clusters. These communities are of interest because of their role in signaling pathways, mRNA translation processes, Golgi transport, metabolic transport, and innate and adaptive immune processes. One of the signaling pathways identified in our communities is associated with cytokines. Cytokines are important intracellular regulators and cell mobilizers involved in the innate and adaptive inflammatory response, cell growth, differentiation, angiogenesis, and homeostasis. Research identified a relation between arsenic exposure and inflammatory cytokines in urothelial cell models and urine [21], the thymus [22], and human lymphocytes [23]. Multiple interferon-alpha units (*IFNA4*, *IFNA7, IFNA10*, and *IFNA16*) show different methylation patterns and could indicate a relation between epigenetic aberrations of interferon-alpha units and arsenic exposure. The interferon-alpha subunits play a vital role in both the innate and adaptive immune response [24]. Finally, when we investigated if these clusters would have a different profile for the risk allele scores, we could not find a specific pattern.

The lack of risk allele score patterns could be because of the small number of SNPs (n=7) present in our data set. The power to discover disease-associated variants depends on the number of risk loci, their frequencies, and effect size [25]. Most of our subjects carry the homozygote allele for two important genes involved in arsenic metabolism, As3MT, and GSTT1. GSTT1 is found to contribute to the observed variability in arsenic metabolism and GSTT1 null individuals may be more susceptible to arsenic exposure [26]. If we could perform a GWAS study, we could identify more SNPs and potentially a higher number of SNPs associated with arsenic exposure. For example, GSTO1, GSTO2, and PNP polymorphisms are associated with developing arsenic-induced skin lesions [27], something we cannot study in our population because data on those 3 genetic polymorphisms are absent. Here, the main problem is the sample size of our study. If certain SNPs do not occur frequently in a population, a sub selection of this population would result in a data set where we cannot observe SNP-related patterns. Increasing the sample size could result in a larger frequency of SNPs, and potentially lead to the discovery of disease-associated variants.

Chapter 4 and **chapter 5** showed the importance of large omics data sets for data integration. If one of the layers is missing large chunks of information, omics integration by multi-layer NMF cannot improve our understanding of a biological system. During the design of a study, one should ask what they would like to achieve by integration omics layers, which omics layers are needed, and if it is feasible to obtain a large sample size.

Furthermore, **chapters 4** to **5** gave us insights into the performance of our integration strategy for population studies. In **chapter 4,** we combined two cohorts (EPIC-ITALY and NSHDS) because the criteria for subjects to take part in the study were similar. Moreover, by combining the two cohorts, we expected the statistical power to increase and potentially obtain patterns related to specific exposure profiles. Even more, it would be of interest to identify a common exposure-related effect in both cohorts. Previous research by Georgiadis et al [28] identified a weaker response in both the Italian cohort and the Swedish females on the epigenome. Espín-Pérez et al [29] identified a sex-specific effect on the transcriptome after combining the cohorts. Therefore, we hypothesized the complementary information on the transcriptome and epigenome could provide us with an integrated exposure profile. On the contrary, the model's outcome predicted three clusters with a strong cohort effect, even after data filtering to remove transcripts or CpG sites associated with either one of the clusters.

This could be because we introduced a new problem in **chapters 4** and **5** compared to the in vitro data in **Chapter 3**:confounding variables. These confounding variables could influence the observed exposure-outcome effect. NMF is in principle an unsupervised method and therefore additional measures have to be applied to correct for those confounding variables. In **chapter 4**, we added a pre-filtering step to remove confounding variables per omics layer. In the future, it might be worthwhile to investigate additional constraints on the multi-layer NMF method, especially on *W* and *H*, to correct for confounding variables in a multi-omics way. Here, the subject's metadata could be used in the objective function to account for confounding variables while minimizing the objective function. An alternative approach would be to use the subject's metadata to initialize *H,* by creating a semi-supervised multi-layer NMF approach.

The case studies provided us with insights in the type of data multi-layer NMF can handle, as well as the usefulness of the genomic interaction networks. Multi-layer NMF is capable of extracting relevant profiles in case of *in vitro* data. From the obtained profiles, genomic interaction networks can be built to further unravel cancer biology. When we went made the transition from *in vitro* data to *in vivo* data, it became clear that our omics integration approach can only identify exposure profiles related to the strongest component in the mixture of compounds. Multi-layer NMF lacks the sensitivity to identify weaker patterns related to compounds with a lower exposure concentration. This lower exposure to a compound can still be of

interest since it can alter cellular biology over time. Therefore improving the sensitivity of our integration approach is vital.

Limitations in linking the epigenome and transcriptome
In the previously described studies, we observed different methylation and gene expression patterns and in most cases these patterns gave us a better understanding of the cellular processes. However, we could not always link the alterations on the epigenome with the transcriptome. In our results described in **chapter 3**, we could identify the relation between DNA hypermethylation and lower gene expression. In **chapter 4** and **chapter 5**, we could not always identify a clear methylation effect on the transcriptome. One reason could be the fact that in vitro data is based on cell lines that show strong differences but also have a strong phenotypic endpoint. The *in vivo* data we used, from healthy subjects, is much more diverse, not only in general parameters as age, body mass index, and sex but also in their maternal DNA methylation as well as different exposure profiles. Furthermore, individual genetics can contribute to faster or slower development of a disease and therefore individuals are more sensitive to chemical exposure. During pre- and postnatal life, environmental factors (including dietary factors) in each tissue microenvironment interact with cellular genetic regulatory architecture and may modulate gene expression. When we consider all those effects, unsupervised learning methods can be too insensitive to extract chemical-induced patterns in different omics layers.

Another complexity is to understand whether CpG site or CpG island is the most important epigenetic factor. We know CpG-poor regions often exhibit hypermethylation while CpG-dense regions often exhibit hypomethylation. If we only study CpG islands, we look at the dense regions and thus neglect a part of the genome methylation. CpG islands are mainly present in the promoter region of a gene (approximately 70%) and promoter methylation leads to gene silencing. When we look at CpG sites, we can study the methylation changes in all regions of the genome and thus gain more novel information. CpG sites outside the promotor region, thus in the gene body, are shown to have vital effects on transcription. Gene-body DNA methylation positively correlates with transcription thus not linked to gene silencing [30,31]. Because one gene can have a large number of associated CpG sites, it is hard to assess whether the interaction between one CpG site and a gene is important. It is vital to experimentally validate if an alteration in one CpG site would be the driver of a molecular event.

Besides methylation in the gene body or promoter region leading to gene amplification or silencing, DNA methylation can also influence alternative splicing. In **chapter 5**, we did see that several CpG sites are differently methylated that are located on the primary transcript region. Whereas we measure every CpG site with the current techniques, the traditional microarrays are designed to measure the total level of expression of a gene without attempting to distinguish between

different splice forms [32]. To gain more knowledge of a biological system, we could also improve the experimental techniques we use to measure different omics layers. It is believed that RNA-sequencing is superior to micro-array techniques because it is a quantifiable technique. This makes it possible to accurately quantify changes in gene expression, gene discovery, differential use of promoters, and splice variants. The changes in gene expression are of interest but of even more interest are the differential use of promoters and splice variants. This could be the connection between the epigenome and the transcriptome that we are now missing and cannot study in full detail.

Conclusion and future research

Overall, the research conducted in this thesis provides novel insights into the relationship between DNA methylation and gene expression. By utilizing genomic interaction networks, we identified potential interesting biological processes altered by aberrations in DNA methylation and gene expression. The developed workflow to integrate omics layers, as well as the integration of existing knowledge, shows to be a promising approach to understand cancer-related or exposure-related features. This group of features, thus genes and/or DNA regions, altered in specific subsets of cancer cell lines or populations might be useful as potential markers to stratify cancer types or exposure-related adverse health effects. The main challenge, however, is to validate and integrate these findings into new studies to further understand their mechanistic and causal effect. Due to the complexity of a biological system, it is non-trivial to investigate the downstream effect of reversing hypo- or hypermethylation of one gene and other biological processes. The proposed interactions in our genomic interaction network might serve as guidance to study this downstream effect, but the validation of these interactions by in vitro experiments is of utter importance. The prediction of disease- or exposure-related patterns in combination with experimental validation could be the final step to increase our current knowledge of biological systems.

References

1. Chung SY, Wong L. Kleisli: a new tool for data integration in biology. Trends Biotechnol. 1999;17:351–5. doi:10.1016/S0167-7799(99)01342-6.
2. Benton D. Bioinformatics — principles and potential of a new multidisciplinary tool. Trends Biotechnol. 1996;14:261–72. doi:10.1016/0167-7799(96)10037-8.
3. Devarajan K, Ebrahimi N. Class Discovery via Nonnegative Matrix Factorization. Am J Math Manag Sci. 2008;28:457–67. doi:10.1080/01966324.2008.10737738.
4. Shao C, Höfer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. Bioinformatics. 2017;33:235–42. doi:10.1093/bioinformatics/btw607.
5. Pascual-Montano A, Carazo JM, Kochi K, Lehmann D, Pascual-Marqui RD. Nonsmooth nonnegative matrix factorization (nsNMF). IEEE Trans Pattern Anal Mach Intell. 2006;28:403–15. doi:10.1109/TPAMI.2006.60.
6. Fujita N, Mizuarai S, Murakami K, Nakai K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Sci Rep. 2018;8:9743. doi:10.1038/s41598-018-28066-w.
7. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS One. 2017;12.
8. Gaujoux R, Seoighe C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. Infect Genet Evol. 2012;12:913–21. doi:10.1016/j.meegid.2011.08.014.

9. Chen G, Yuan A, Cai T, Li C, Bentley AR, Zhou J, et al. Measuring gene–gene interaction using Kullback–Leibler divergence. Ann Hum Genet. 2019;83:405–17. doi:10.1111/ahg.12324.
10. Kasturi J, Acharya R, Ramanathan M. An information theoretic approach for analyzing temporal patterns of gene expression. Bioinformatics. 2003;19:449–58. doi:10.1093/bioinformatics/btg020.
11. Buescher JM, Driggers EM. Integration of omics: more than the sum of its parts. Cancer Metab. 2016;4:4. doi:10.1186/s40170-016-0143-y.
12. Grossman D, Altieri DC. Drug resistance in melanoma: mechanisms, apoptosis, and new potential therapeutic targets. Cancer Metastasis Rev. 2001;20:3–11. doi:10.1023/a:1013123532723.
13. Shtivelman E, Davies MA, Hwu P, Yang J, Lotem M, Oren M, et al. Pathways and therapeutic targets in melanoma. Oncotarget. 2014;5:1701–52. doi:10.18632/oncotarget.1892.
14. Jin J, Wang L, Tao Z, Zhang J, Lv F, Cao J, et al. PDGFD induces ibrutinib resistance of diffuse large B-cell lymphoma through activation of EGFR. Mol Med Rep. 2020. doi:10.3892/mmr.2020.11022.
15. Pai AA, Pritchard JK, Gilad Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS Genet. 2015;11:e1004857. doi:10.1371/journal.pgen.1004857.
16. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science (80-). 2012;337:1190–5. doi:10.1126/science.1222794.
17. Christiani DC, Mehta AJ, Yu C-L. Genetic susceptibility to occupational exposures. Occup Environ Med. 2008;65:430–6. doi:10.1136/oem.2007.033977.
18. Alemu EY, Carl JW, Corrada Bravo H, Hannenhalli S. Determinants of expression variability. Nucleic Acids Res. 2014;42:3503–14. doi:10.1093/nar/gkt1364.
19. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587. doi:10.1186/1471-2105-11-587.
20. Zhuang JJ, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinformatics. 2012;13:59. doi:10.1186/1471-2105-13-59.
21. Liu S, Sun Q, Wang F, Zhang L, Song Y, Xi S, et al. Arsenic Induced Overexpression of Inflammatory Cytokines Based on the Human Urothelial Cell Model in Vitro and Urinary Secretion of Individuals Chronically Exposed to Arsenic. Chem Res Toxicol. 2014;27:1934–42. doi:10.1021/tx5002783.
22. Gera R, Singh V, Mitra S, Sharma AK, Singh A, Dasgupta A, et al. Arsenic exposure impels CD4 commitment in thymus and suppress T cell cytokine secretion by increasing regulatory T cells. Sci Rep. 2017;7:7140. doi:10.1038/s41598-017-07271-z.
23. Zarei MH, Pourahmad J, Nassireslami E. Toxicity of arsenic on isolated human lymphocytes: The key role of cytokines and intracellular calcium enhancement in arsenic-induced cell death. Main Gr Met Chem. 2019;42:125–34. doi:10.1515/mgmc-2019-0014.
24. TOUGH DF. Type I Interferon as a Link Between Innate and Adaptive Immunity through Dendritic Cell Stimulation. Leuk Lymphoma. 2004;45:257–64. doi:10.1080/1042819031000149368.
25. Hong EP, Park JW. Sample Size and Statistical Power Calculation in Genetic Association Studies. Genomics Inform. 2012;10:117. doi:10.5808/GI.2012.10.2.117.
26. Kile ML, Houseman EA, Quamruzzaman Q, Rahman M, Mahiuddin G, Mostofa G, et al. Influence of GSTT1 Genetic Polymorphisms on Arsenic Metabolism. J Indian Soc Agric Stat. 2013;67:197–207. http://www.ncbi.nlm.nih.gov/pubmed/24511153.
27. Luo L, Li Y, Gao Y, Zhao L, Feng H, Wei W, et al. Association between arsenic metabolism gene polymorphisms and arsenic-induced skin lesions in individuals exposed to high-dose inorganic arsenic in northwest China. Sci Rep. 2018;8:413. doi:10.1038/s41598-017-18925-3.
28. Georgiadis P, Gavriil M, Rantakokko P, Ladoukakis E, Botsivali M, Kelly RS, et al. DNA methylation profiling implicates exposure to PCBs in the pathogenesis of B-cell chronic lymphocytic leukemia. Environ Int. 2019;126:24–36. doi:10.1016/j.envint.2019.01.068.
29. Espín-Pérez A, Hebels DGAJ, Kiviranta H, Rantakokko P, Georgiadis P, Botsivali M, et al. Identification of Sex-Specific Transcriptome Responses to Polychlorinated Biphenyls (PCBs). Sci Rep. 2019;9:746. doi:10.1038/s41598-018-37449-y.
30. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23:555–67. doi:10.1101/gr.147942.112.
31. Bender CM, Gonzalgo ML, Gonzales FA, Nguyen CT, Robertson KD, Jones PA. Roles of Cell Division and Gene Transcription in the Methylation of CpG Islands. Mol Cell Biol. 1999;19:6690–8. doi:10.1128/MCB.19.10.6690.
32. Lee C, Roy M. Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol. 2004;5:231. doi:10.1186/gb-2004-5-7-231.

Chapter 7
Impact paragraph

The primary goal of this thesis is to deepen our knowledge on the interaction between the transcriptome and epigenome. Although both layers have been well studied individually, their combined effect is not completely understood. In this thesis, we showed that the integration of the transcriptome and epigenome increased our understanding of mechanisms related to disease mechanisms and exposure-related effects.

Omics technologies provide exciting data to study cellular mechanisms in much more detail. The abundance of biological data made data integration approaches increasingly popular and provided us with many computation methods [1]. However, biological interpretation remains a challenge, and improved computational tools are needed. One way to increase our understanding of a system is to use a data-driven approach and extract the relevant biomolecules. However, these approaches do not make us of valuable pre-existing knowledge available in many databases. Finally, since a biological system is not static, we should explore the options to visualize dynamic changes. In this thesis, we combine a data-driven technique with a knowledge-driven approach to create genomic interaction networks. The existing knowledge can help us answer important questions: what is the role of a biomolecule, is it associated with a disease, and in which pathway is it involved? This information, in combination with network analysis methods, helps us to identify important genes and their epigenetic status. To study a dynamic system, we developed DynOVis [2] to highlight gene expression changes on a network, to highlight the most affected genes over time or dose. To integrate the two omics layers, we have developed the tool GINBuilder to study the relationship between the epigenome and transcriptome. Both tools have been made freely available to the scientific community.

The integration of the transcriptome and epigenome is not only important to study a disease mechanism but also, as our work shows, cellular characteristics. Cancer cell lines are important models for drug discovery and development. However, those studies are not always successful when translated to the clinic and failure rates in drug development are high [3]. Cancer cell lines can carry specific alterations that make them different from the type of cancer they try to mimic. Current research acknowledged the necessity to further look into the molecular characteristics of cancer cell lines [4]. In our research, we showed for two cancer cell line data sets that cell lines derived from the same tissue can show different transcriptomic and epigenomic profiles. We identified a potentially interesting event leading to a drug-resistant state of melanoma cells and proposed a marker for melanoma inhibition. In a subset of lymphoid neoplasms, we were able to identify epigenetic and transcriptomic changes that might contribute to drug resistance [5]. Moreover, we showed there is a potential dysregulation of an important signaling axis. These findings increase our understanding of those cancer cells and contribute to cancer research in general. Furthermore, they highlight the importance of understanding the characteristics of your cell model.

As previously mentioned, omics integration can help us to unravel exposure-related effects. This can increase our understanding of mechanisms leading to disease development but also provide us with potential biomarkers as early indicators of disease progression due to environmental exposure. Chronic exposure to persistent organic pollutants (POPs) and heavy metals have a major adverse health effect on humans [6, 7]. It is therefore vital to understand the relationship between exposure levels and transcriptomic and/or epigenetic changes. Identifying those exposure-related changes associated with disease-related processes could help us understand the underlying mechanisms and define possible biomarkers for early diagnosis. Our work emphasized important biological processes, including immune-related, signaling, and DNA repair processes affected by POPs and heavy metals. Furthermore, we identified a sex-specific effect, showing that it is vital to consider the relation between sex, exposure, and disease development. The biological signatures associated with exposure can be of interest to use as a potential marker or as a target to understand disease-related exposure.

To further improve our understanding of the relationship between toxic exposure on either the transcriptome or epigenome, we performed an exploratory study to identify the effect of Single Nucleotide Polymorphisms (SNPs) on this relationship. One of the challenges we faced was the small population size of this study. We solved this problem by expressing DNA methylation as M-values instead of the golden standard β-values. This increased the overall signals in the data and is a worthwhile consideration for researchers working with small sample sizes. Although we could not identify a strong correlation between exposure-related effects and SNPs, our results indicated that exposure levels might not be the only driver of the adverse health effects. Different processes could be linked to arsenic exposure and cardiovascular diseases and diabetes. Those processes do not correlate to exposure and might be driven by susceptibility to arsenic exposure. Therefore, we believe it is worthwhile to further investigate the role of SNPs. Our work indicates it is vital to consider the sample size needed to detect variations in SNPs during the design phase of a study.

Overall, we demonstrated that the integration of omics data with a combination of data-driven and knowledge-driven network integration can help us to further unravel the adverse health effects of a chemical compound as well as an understanding of important mechanisms in cancer cell lines. The framework can help to further elucidate disease and exposure-related effects on the transcriptome and epigenome in future studies. It can lead to the discovery of key components in disease development or early biomarker discovery.

References

1. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017;8. doi:10.3389/fgene.2017.00084.
2. Kuijpers TJM, Wolters JEJ, Kleinjans JCS, Jennen DGJ. DynOVis: a web tool to study dynamic perturbations for capturing dose-over-time effects in biological networks. BMC Bioinformatics. 2019;20:417. doi:10.1186/s12859-019-2995-y.
3. Hingorani AD, Kuan V, Finan C, Kruger FA, Gaulton A, Chopade S, et al. Improving the odds of drug development success through human genomics: modelling study. Sci Rep. 2019;9:18911. doi:10.1038/s41598-019-54849-w.
4. Jaeger S, Duran-Frigola M, Aloy P. Drug sensitivity in cancer cell lines is not tissue-specific. Mol Cancer. 2015;14:40. doi:10.1186/s12943-015-0312-6.
5. Kuijpers TJM, Kleinjans JCS, Jennen DGJ. From multi-omics integration towards novel genomic interaction networks to identify key cancer cell line characteristics. Sci Rep. 2021;11:10542. doi:10.1038/s41598-021-90047-3.
6. Carpenter DO. Health effects of persistent organic pollutants: the challenge for the Pacific Basin and for the world. Rev Environ Health. 2011;26. doi:10.1515/reveh.2011.009.
7. Jaishankar M, Tseten T, Anbalagan N, Mathew BB, Beeregowda KN. Toxicity, mechanism and health effects of some heavy metals. Interdiscip Toxicol. 2014;7:60–72. doi:10.2478/intox-2014-0009.

Addendum

Acknowledgements

Als eerste wil ik Prof. dr. Jos Kleinjans bedanken voor het aanbieden van mijn promotieplek bij TGX. Tijdens de jaren die wij samen hebben gewerkt was jij de factor die er voor zorgde dat het onderzoek in een concrete richting bleef gaan, zonder van links naar rechts te stuiteren. De vele vragen en discussies hebben mij geholpen in het verder uitwerken van de analyses, door ook de data vanuit een ander perspectief te bekijken.

Natuurlijk wil ik ook dr. Danyel Jennen bedanken voor de wekelijkse begeleiding, waar wij het niet alleen over werk konden hebben maar ook altijd wel even tijd was voor bijzaken. Jij gaf mij de ruimte om zelf ideeën uit te proberen en deze samen veder uit te werken. Ik zal ook zeker niet ons 'uitstapje' in de kunstwereld vergeten met onze deelname aan de Biology Art Design award (BAD, een perfect gekozen afkorting). Het gesprek over het kleuren van witte doeken met urine, zal mij altijd bij blijven als het meest vreemde gesprek dat ooit heeft plaats gevonden in het *in sillico* lab.

Ik wil ook prof.dr. Ilja Arts, prof. dr. ir. Natal Van Riel, prof. dr. Jan Aerts en dr. Marc Teunis bedanken voor het beoordelen van mijn thesis.

During the lockdown, it become clear that you cannot survive a PhD without some nice colleagues.

Rajinder, you introduced me to all the nice Indian dishes, the world of cricket, and the famous Raji-shots. It was always nice to visit you and Shweta and enjoy some comfort food after a hard day of work. **Jarno,** you were always so kind to answer any question and offer help when needed. **Evelyn,** the queen of gossip, I never met someone so direct as you, and you never hesitated to speak your mind. It took some time to get used to but most of the time your comments were just funny. **Manon**, you entered our office in its prime time of chaos. However, you took control over Juan like a mother figure, so we could work again. Hopefully you have the same power over your PhD as you had over Juan. **Juan,** you are one of the most energetic, enthusiastic, and strangest (in a good way) person I have ever met. You are always open for a good discussion about anything, work, food, or if we should turn on/off the heater. **Marta,** it was fun to learn about the Italian culture, and I must say, I'm still a bit jealous you had the courage to bring your own fancy high quality coffee to the office. **Nhan,** you are not only a master of the martial arts but also the master in making one-liners. I still remember you said I look like a hobo when I was getting ready for the GROW science day panel discussion. **Julian** thank you for the nice talks about work and normal stuff. It was an honor to run the EuroTox funrun with you at least one time. **Terezinha,** you are one of the brightest bioinformaticians around, thank you for all the nice and insightful discussions. **Marcha** and **Heloise,** I didn't have lab buddies (what's a lab right?) but luckily I had you as *train buddies* during my first year travelling between Maastricht and Eindhoven. It made traveling

much more fun. **Jian,** you introduced me to real tea (after telling me my tea smelled horrible) but you also gave plenty tips on how to survive the PhD. **Daniela** and **Julia,** you always brought some positive energy to the office. **Maria** and **Almudena**, thank you for all the conversations during lunch.

Ik wil ook alle andere collega's bedanken: **Twan**, **Theo**, **Jacco**, **Juma**, **Sacha**, **Duncan**, **Simone** en **Florian** voor hun vragen tijdens de donderdagochtend presentaties. **Marcel** en **Yannick** voor de vele gesprekken bij het koffieapparaat, maar ook voor de gezelligheid tijdens de groepsuitjes en de promotiefeestjes. **René** en **Rob** voor het regelen van de algemene zaken, maar vooral **Christa,** jij hebt er voor gezorgd dat elke trip naar het buitenland vlekkeloos is verlopen. **Tom**, **Kevin** en **Sean**, jullie hebben er voor gezorgd dat de server bleef draaien en waren altijd in de buurt voor vragen. I would also like to thank **Yasir Rehman** and **Soterios Kyrtopoulos** for the discussions and contribution to the work in this thesis.

Papa, **mama**, **Anique** en **Peter**, bedankt voor het meebrengen van de Brabantse gezelligheid tijdens de bezoekjes aan Maastricht. Zo werd de afstand tussen Den Bosch en Maastricht toch weer een stukje kleiner. Bedankt voor alle steun tijdens mijn PhD, maar natuurlijk ook tijdens alle jaren die daar aan vooraf gingen. **Frank, Gemma, Marjolein** en **Mark**, bedankt voor de interesse in mijn werk en de gezellige uitstapjes.

Tot slot wil ik **Denise** natuurlijk bedanken voor alle steun tijdens de afgelopen vier jaar. We hebben samen veel leuke dingen ondernomen die ervoor zorgde dat ik niet bezig was met mijn promotie. De lekkere gebakjes die jij in het weekend maakte zorgden er steeds weer voor dat ik met genoeg energie aan de week begon. Hopelijk gaan we nog veel leuke en nieuwe avonturen beleven.

Curriculum vitae

Tim Kuijpers was born on February 8th, 1988 in 's-Hertogenbosch, the Netherlands. He got his bachelor degree in the field of biomedical engineering at the University of Technology in Eindhoven. For his master degree, he joined the research group of computational biology at the department of biomedical engineering at the same institution. During his master program, he worked on the network-based analysis of human hepatic cells, to identify clinically relevant genotype-phenotype differences from liver biopsies in subgroups of nonalcoholic fatty liver disease patients. After finishing his master thesis, he visited the Mardinoglu lab at the Science for life laboratory in Stockholm, Sweden. Here he worked on the identification of liver cancer-specific FASN inhibitors through coexpression network analysis.

In November 2016, he started working as a Ph.D. student at the department of Toxicogenomics at Maastricht University. During his Ph.D., he focused on the interaction between the epigenome and transcriptome in cancer cell lines and population studies. Here, he applied omics integration techniques to extract relevant features to build genomic interaction networks.

In 2021, he joined the BioInterface Science group of Jan de Boer at the University of Technology, Eindhoven. Here, he is working on FAIR data infrastructure for the Research Center for Materials-Driven Regeneration consortium as well as the Biomaterial atlas. This atlas will be the first pilot, to investigate the combination of new data and knowledge integration from existing data for biomaterial discovery. He will also explore new opportunities to combine *in silico* approaches with biomaterial screening.

List of Publications

Published papers

- "From multi-omics integration towards novel genomic interaction networks to identify key cancer cell line characteristics", **Kuijpers T.J.M.**, J.C.S. Kleinjans, and Jennen D.G.J., Scientific reports, 2021
- "DynOVis: a web tool to study dynamic perturbations for capturing dose-over-time effects in biological networks", **Kuijpers T.J.M**., Wolters J.E.J., Kleinjans J.C.S. and Jennen D.G.J., BMC bioinformatics, 2019
- "Network analyses identify liver-specific targets for treating liver diseases", Sunjae Lee, Cheng Zhang, Zhengtao Liu, Martina Klevstig, Bani Mukhopadhyay, Mattias Bergentall, Resat Cinar, Marcus Ståhlman,Natasha Sikanic, Joshua K Park, Sumit Deshmukh, Azadeh M Harzand , **Tim Kuijpers**, Morten Grøtli, Simon J Elsässer, Brian D Piening, Michael Snyder, Ulf Smith, Jens Nielsen, Fredrik Bäckhed, George Kunos,

Mathias Uhlen, Jan Boren, Adil Mardinoglu, Molecular Systems Biology, 2017

Conference papers

- "Updating strategies for nonnegative matrix factorization to integrate cross omics layers", **T.J.M. Kuijpers**, J.C.S. Kleinjans, D.G.J. Jennen, Toxicology Letters, 2019
- "DynOVis: a new tool to visualize dynamic perturbations of biological networks after toxic exposure", **T.J.M. Kuijpers**, J.C.S. Kleinjans, D.G.J. Jennen, Toxicology Letters, 2018

List of conference presentations

Oral presentations

- *"Multi omics data integration to study the relevance of in vitro disease models through the creation of genomic interaction networks",* World Congress on Alternatives and Animal Use in the Life Sciences, 2021, online event
- "*Genomic interaction networks to investigate relationships between biological alternations on omics levels." BioSB conference 2020, Lunteren, The Netherlands*

Poster presentations

- Poster presentation Society of Toxicology 2020, Anaheim, USA (event cancelled, online poster)
- Poster presentation EuroTox 2019, Helsinki
- Poster presentation AstraZeneca Educational Reception 2019, Helsinki
- Poster presentation Dutch Bioinformatics & Systems Biology congress 2019
- Poster presentation EuroTox 2018, Brussel
- Demo presentation Dutch Bioinformatics & Systems Biology congress 2018