University of Warsaw

Faculty of Mathematics, Informatics and Mechanics

Krzysztof Kamil Gogolewski

Student no. 291538

Matrix methods in transcriptomic and metabolomic data analysis

PhD's dissertation

in COMPUTER SCIENCE

Supervisor:

Prof. Anna Gambin

Institute of Informatics, University of Warsaw

February 2019

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of PhD of Computer Science.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Matrix methods in transcriptomic and metabolomic data analysis

In this dissertation we walk through various approaches of modelling and analysis of transcriptomic and metabolomic data. The dissertation opens with an introduction of the current state of the art in the context of high-throughput data along with genetic background description. We look through current technologies that are used for obtaining transcriptomic data as well as computational methods and tools for their analysis including various methods for decomposition of transcriptomic signal and integration with metabolomic knowledge. Throughout the main three chapters of this dissertation we discuss specific experimental settings and data for which adequate methods for transcriptomic and metabolomic data analysis are derived and applied. Each of these chapters presents a different computational method for inference of biological knowledge, and is supported with a case-study based on real life experimental data. Finally, results from joint work with Baylor Collage of Medicine concerning the role of FOXF1 gene in lungs disease are closing the dissertation.

Streszczenie

Metody macierzowe w analizie danych transkryptomicznych i metabolomicznych

W niniejszej rozprawie omawianych jest kilka podejść do modelowania i analizy danych transkryptomicznych. Pracę otwiera krótki wstęp do obecnego stanu wiedzy dotyczącego wysokoprzepustowych danych wraz z ogólnym wprowadzeniem do genetyki. Omówione zostają obecnie używane technologie do gromadzenia danych transkryptomicznych, jak również obliczeniowe metody ich modelowania i analizy, w szczególności dotyczących dekompozycji sygnału transkryptomicznego oraz jego integracji z wiedzą metabolomiczną. W ramach trzech głównych rozdziałów rozprawy dyskutowane są specyficzne scenariusze i dane eksperymentalne, do których zostają opracowane i zastosowane odpowiednie metody analizy danych transkryptomicznych. Każdy z rozdziałów prezentuje pewną metodę obliczeniową służącą pozyskiwaniu wiedzy biologicznej oraz jej zastosowanie w konkretnym studium przypadku używającym danych eksperymentalnych. Ostatecznie, wyniki pochodzące ze współpracy z Baylor Collage of Medicine dotyczące roli genu FOXF1 w rozwoju chorób płuc zamykają zasadniczą część rozprawy.

Keywords

Matrix Decomposition, Data Analysis, Systems Biology, Feature Selection, Transcriptomics, Metabolomics

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

I.6. Simulation and Modelling
J.3. Life and Medical Sciences

Tytuł pracy w języku polskim

Metody macierzowe w analizie danych transkryptomicznych i metabolomicznych

Contents

List of Figures

List of Tables

1 Introduction

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

— Maria Skłodowska-Curie

The rapid progress of biotechnological methods naturally entails the necessity to introduce new, sophisticated bioinformatical methods intended for analysis of fast-growing amounts of biomedical and molecular data. These data are mostly obtained using high-throughput technologies (Reuter *et al.*, 2015), which can be briefly described as automatic screening procedures for conduction of experiments such that large scale repetition becomes feasible. In the field of bioinformatics and systems biology, these technologies aim to perform an exact screening of cell biology and determine its molecular picture by the means of: genome structure (genomics), transcriptome composition (transcriptomics), structure and functions of proteins (proteomics), activity of chemical processes involving metabolites (metabolomics), and others (see Figure 1.1). In the brackets were given fields of study in biology dealing with the analysis of a specific type of data. They all constitute the family of *-omics* sciences that all together aim to characterize and quantify available data in the context of structure, function, and dynamics of a cell, an organism or a population (Hasin *et al.*, 2017).

From the above distinguished types of fields, in particular, we want to focus on transcriptomics (Lowe *et al.*, 2017) and metabolomics (Riekeberg and Powers, 2017) and the

corresponding types of data. It is worth noting, that these types were already used number of times as a source of knowledge in the research conducted in various fields of studies, e.g. general biology (Watson *et al.*, 2015; Shin *et al.*, 2018), personalized medicine (Li *et al.*, 2016), cancer diagnosis (Ren *et al.*, 2016; Yang *et al.*, 2017) or classification (Borgan *et al.*, 2010) and understanding aetiology of other diseases (Stempler *et al.*, 2014; Borrageiro *et al.*, 2018).

In this dissertation, we are about to present and describe novel computational frameworks, methods and algorithms dedicated to analysis of transcriptomic and metabolomic data. Additionally, each of these methodological ideas is accompanied by results from different biomedical case-studies carried on different types of data obtained using various high-throughput technologies. Simultaneously, we hope to prove that its scientific value constitutes a significant contribution to the interdisciplinary world of bioinformatics by both development of computational methods and algorithms that help to investigate biomedical data, as well as better understanding of molecular biology processes.

Nonetheless, before we start a thorough presentation of the results contained in this work, let us prepare a certain terminology that will allow us to better comprehend the interdisciplinary language of bioinformatics. For this purpose, we will briefly recall some concepts in the field of cell genetics and biology, that are used for precise description and understanding of the problems we are facing in the further course of the work.

1.1. Genetic Background

All known living organisms are defined by *genomes*, that contain all the genetic information that is required for their growth, development, functioning, and reproduction. This genetic information is encoded by monomeric units called *nucleotides*, that are organic molecules composed of: a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group: *cytosine* [C], *guanine* [G], *adenine* [A], *thymine* [T] (or alternatively *uracil* [U]). To simplify the notation, we usually designate a nucleotide molecule by its phosphate group (e.g. by C we mean 2′-deoxycytidine 5′-triphosphate nucleotide). Nucleotides bind together into a chain polymer structure by covalent bonds between the sugar of one nucleotide and the phosphate of the next one. A polynucleotide chain composed of A, C, G, T nucleotides forms a *DNA* (deoxyribonucleic acid) molecule. If a polynucleotide chain is composed of A, C, G, U (i.e. uracil instead of thymine) it constitutes a *RNA* (ribonucleic acid) molecule.

In living organisms, RNA naturally is found in a form of a single-stranded structure, that may fold onto itself. Alternatively, DNA usually exists as two polynucleotide chains coiled around each other and bound together by hydrogen bounds between two nucleotides

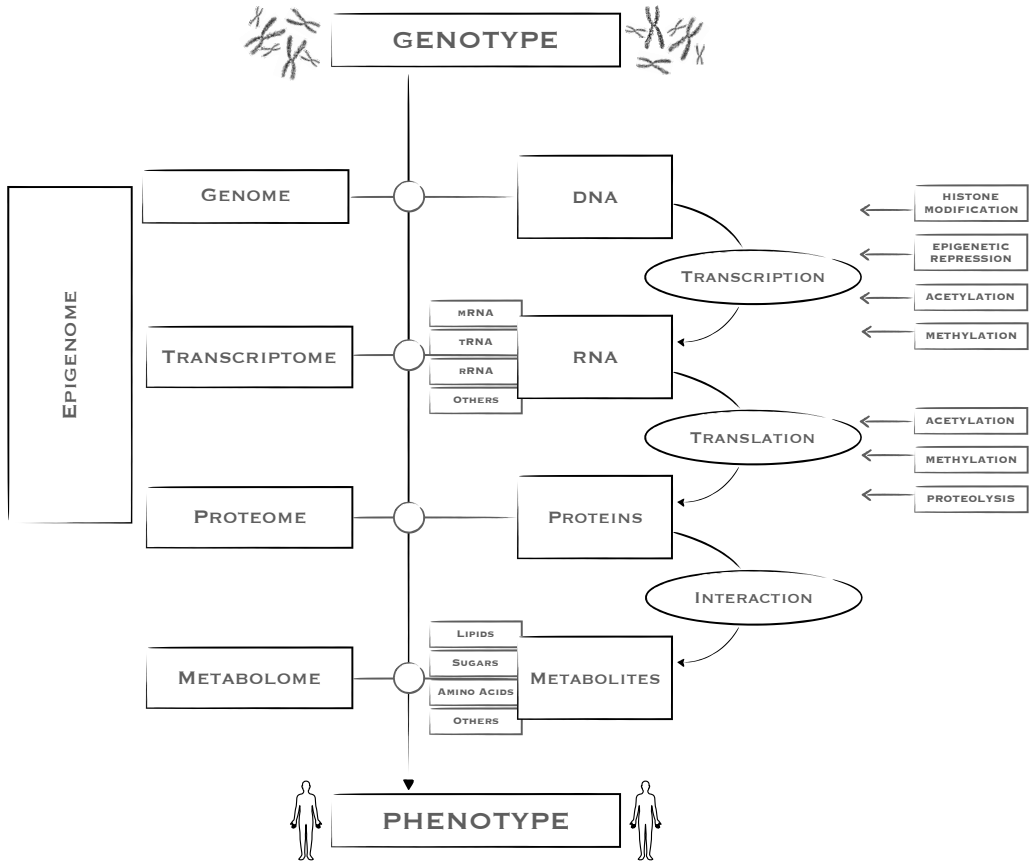


Figure 1.1: Genetic information flow in the human cell. The figure outlines the subsequent stages during the genotype–phenotype mapping. At each step a molecular information is processed leading from genetic code contained in an individual's genotype, through various genetic and epigenetic processes, to observable traits of the individual.

following base pairing (complementarity) rules (A together with T and C with G) forming a double helix. Such form of DNA is a chemical description of what constitutes whole genome sequences.

When it comes to its organisation, in eukaryotes, genome is composed of one or more linear DNA *chromosomes*. Chromosomes are composed of two chromatin fibers, each made of nucleosomes, that are parts of DNA strand (around 200 nucleotides) attached and wrapped around eight core histones and interconnected by sections of linker DNA (around 50 nucleotides). Thanks to packaging proteins such complex genome structure is properly placed and organized in the cell's nucleus.

Within genomes we can distinguish segments of nucleotides that contain biological information and code for a molecule that has a function (i.e. protein or functional non-coding RNA). Such sequence is called *gene*. In order for a protein to be produced from a gene a bunch of processes has to occur. The first step is the *transcription* process, which is coordinated by specific enzymes and proteins and results in copying a content of a single gene onto a RNA molecule. If the RNA is a coding molecule, it carries a genetic information about polypeptide sequence and is called *mRNA* (messanger RNA), otherwise it is a non-coding or

functional RNA. In case of mRNA molecules, they are subjected to *translation* process, that results in production of a polypeptide chain. Then, a polypeptide chain is folded into its characteristic and functional three-dimensional structure. Finally, if necessary, the protein is transferred to the place of demand and is used according to its purpose by participation in various biochemical reactions within intra- or extra-cellular space.

Here, hopefully one can notice, that there are many levels at which molecular activity can be measured and quantified. Specifically, the set of all mRNA (or RNA in general according to the experiment setting) molecules in a cell or population of cells constitutes its *transcriptome*. By analogy *proteome* is a set of all proteins and *metabolome* a set of all metabolites within a scope of interest (e.g. cell, tissue, organism).

1.2. Genotype-Phenotype Gap and Motivation

Naturally, the above description could have been much more detailed, however, for the purpose of this dissertation it should be sufficient for proper understanding of all problems that will be formulated further on. Nonetheless, there is also an additional reason why it is good to get familiar with the biological context and nomenclature that are inseparable from the bioinformatical background of this work. Namely, to get to know and understand the value of computational and algorithmic potential of computer science in the field of modern molecular biology.

In (Alberch, 1991), for the first time the author suggests the existence of a mapping between *genotype* (gr. γένος - descendant, offspring, race; τύπος - mark, reflection, impression), i.e. an organism's full hereditary information, and *phenotype* (gr. φαίνω - appear, shine), i.e. an organism's actual observed properties, such as morphology, development, or behaviour. Later on, others also discussed the role of genotype in the observable phenotype of an individual (Pigliucci, 2010; Gjuvsland *et al.*, 2013) and the problem is still open. A simple, everyday life example of how this relationship is inscrutable are monozygous (i.e. identical), who share the same genotype, yet they never share the same phenotype, because even their fingerprints are never fully identical. How is that possible? In this work, we will consequently address questions about how different molecular layers related to a genotype, e.g. transcriptome or metabolome, influence the observable phenotypic properties such as forms of activity in cell populations, cell type or cancer morphological subtypes.

Therefore, apart from the theoretical aspect of presenting novel algorithms and computational methods, the common goal, from the biological point of view, of all scientific projects described later in the dissertation, is to deepen, even to a small extend, our understanding of the genotype-phenotype relationship. In the following sections, we briefly describe methods that were used and obtained results in various attempts to understand

this relationship.

1.3. Main Results

As already mentioned, all of the results presented in this dissertation are an interdisciplinary blend, meaning that we propose new computational methods, that can be applied to specific biological problems. Therefore, each of the next four chapters formulates specific biological questions that we try to answer in an unprecedented way. In each case, we provide a description of a solution that we found and an accompanying case-study that we have carried based on real biomedical data. In short, the results in the dissertation are organized as follows.

Decomposition of Transcriptomic Signal

In Chapter 2 we focus on the problem of transcriptomic signal decomposition for averaged measurements of transcriptional activity in homogeneous populations of cells. Generally, in most studies that focus on tracking transcriptional changes caused by specific experimental conditions, researchers use samples that are composed of many cells. In such experimental setting, the information referring to high and low activity of genes is evaluated analysing the behaviour of relatively large population of cells by averaging its properties. However, even assuming perfect sample homogeneity with respect to cell type, different sub-populations of cells can exhibit diverse transcriptomic profiles, as they may follow different regulatory/signalling pathways. The purpose of the study presented in this chapter is to provide a novel methodological scheme to account for possible internal, functional heterogeneity in homogeneous cell lines, including cancer ones.

We propose a computational workflow to infer the proportion between sub-populations of cells that are expected to manifest various functional behaviour in a given sample. We assume, that the observed transcriptomic activity $A_{i,j}$ of a specific gene i and a sample j can be modelled as a sum of activities of gene i in each of k sub-populations $W_{i,l}$, where $1 \leq l \leq k$, scaled by the proportion $H_{l,j}$ of corresponding sub-population with respect to the whole sample:

$$A_{i,j} = \sum_{l=1}^{k} W_{i,l} H_{l,j} + E_{i,j}$$

where $E_{i,j}$ is the approximation error. Building on this premise, we incorporate two types of non-negative matrix factorization algorithms by (Erkkila *et al.*, 2010) and (Brunet *et al.*, 2004) along with biological expert knowledge and come up with a method called MPH (Molecular Process Heterogeneity), that allows to estimate: (i) transcriptomic profile of a specific molecular process carried by a cell; (ii) proportion of molecular activities ongoing

in homogeneous populations of cells. Additionally, the method can help in discovering potential biomarkers of a molecular process, which is a challenging task in modern medical diagnostics.

The MPH method was validated using data from transcriptome measurements of homogeneous neuroblastoma cell line performed with RNA microarrays technology. The experiment aimed to examine cell viability in two experimental conditions: treatment with C2–ceramide and C2–seramide accompanied by PJ34 PARP inhibitor. An advantage of the presented methodology is the fact, that it can be easily applied to RNA-seq or single-cell RNA-seq data. Moreover, it complements standard tools to indicate most important networks from transcriptomic data and in particular could be useful in the analysis of cancer cell lines affected by biologically active compounds or drugs.

The results presented in this chapter are already published in the paper (Gogolewski *et al.*, 2017) along with an additional case-study of Ovarian Cancer cells.

Latent Structure and Noise Reduction

In the next chapter, we are interested in extracting the systematic patterns from transcriptomic data in an unsupervised manner. Naturally, there are numerous different approaches to accomplish this task. In our work, we started from the robust PCA (RPCA) algorithm suggested by (Candès *et al.*, 2011).

Formally, RPCA algorithm aims to decompose the input matrix A, into low-rank matrix L and sparse matrix S components by minimizing the following optimization problem:

$$\min_{L,S} ||L||_* + \lambda_1 ||S||_1, \text{ where } A = L + S$$

where $||A||_*$ is the nuclear norm of matrix A and $||A||_1$ is the first norm of a vectorized matrix A. Originally, the algorithm was applied to the image analysis from video surveillance frames and aimed to distinguish fixed and varying parts of the images. Since in transcriptomic data we can also think of common and changing patterns, we decided to adapt the original algorithm to biomedical data, thus two extensions of RPCA algorithm were developed.

First, we present a truncated version of the mentioned RPCA (tRPCA) algorithm. Our extension draws from the truncated version of SVD algorithm, which means that compared to the original $A = U\Sigma V^*$ decomposition, we calculate only the desired number (k) of column vectors of U and row vectors of V^* corresponding to the k largest singular values of Σ. This extension makes the algorithm faster and more memory efficient than the original RPCA algorithm. Additionally, since this version of the algorithm imposes an additional constraint on the maximum rank of the L matrix of resulting decomposition, it enables possibility to specify the expected dimensionality of the background, fixed part of the data.

Next, since biomedical data are by nature influenced by the biological fluctuations of various nature the supposed, fixed part of low-rank L matrix will inevitably carry an unwanted stochastic effect. For this reason, we introduce another extension of the `tRPCA` algorithm by an additional reduction of dense noise using the L_2 regularization (`tRPCAL2`). Unlike both `RPCA` and `tRPCA` that only consider a low-rank L and sparse S matrices, the proposed algorithm extracts a noise E matrix inherent in modern genomic data. In such setting, the optimisation problem is extended to the following form:

$$A = L + S + E$$
$$\min_{L,S,E} ||L||_* + \lambda_1 ||S||_1 + \lambda_2 ||E||_F$$

where $||A||_F$ is the Frobenius norm of matrix A. In order to solve this minimization problem we extend the Alternating Direction procedure and introduce an additional threshold operator $\mathcal{E}$, that coordinates the content of E matrix.

In the last part of Chapter 3 we demonstrate the usefulness of the proposed theory. Our methods are applied to the peripheral blood mononuclear cell (PBMC) single cell RNA-seq data. In particular, we report how the clustering of a low-rank L matrix showcases better classification of unlabelled single cells. Moreover, possibility of cell subtypes detection and co-expression patterns detection are also presented. Finally, we point out how the proposed variants are well-suited for high-dimensional and noisy data that are routinely generated in genomics.

The major part of the results presented in this chapter are already published in the proceedings of the 14$^{\text{th}}$ International Symposium on Bioinformatics Research and Applications (ISBRA) 2018 (Gogolewski *et al.*, 2018b). Additionally, after invitation to publish the extended article, it is currently under review in the Journal of Computational Biology.

Overrepresented Structure of Metabolic Networks

It is known, that the huge amount of biological data gained from many different sources open up new possibilities in data integration. Nonetheless, process of integration should be carried with caution regarding the broadly understood problems of overfitting and overlearning. In Chapter 4 we present two types of results.

On one hand, we show how integration of metabolic and transcriptomic knowledge may lead to statistical redundancy, which results in artefact discoveries, that as a matter of fact can be supported by literature. This result has a meta-scientific value, because it shows how scientific reports can introduce knowledge imbalance in mathematical models describing biological phenomena. To show-case this problem, we present an example of integration of RNA-seq data with the genome-scale model of human metabolism RECON 2.2 (Swainston *et al.*, 2016) and describe the outcome of this study. The obtained results

suggest existence of metabolomic biomarkers that can clearly distinguish various groups of cancerous patients by specific metabolic activity pattern. Moreover, we are able to support our results by the literature. However, as a counter-example to all discoveries, we also present an analysis carried on random dataset. The obtained result proved that all observations should be treated as artefacts resulting from the information embedded in the used model.

Therefore, to overcome this problem, later on in the chapter we suggest a possible bias reduction technique, which we further on validate using the Renal Cell Carcinoma (RCC) TCGA dataset. Thanks to the bias reduction we were able to to provide a reliable metabolic description of known morphological RCC subtypes. Furthermore, we suggest a possible existence of poor-prognosis cluster of patients, which is characterized by common low activity of drug transporting enzymes important in chemotherapy.

We hope that the work presented in this chapter would be a warning sign, which reminds other researchers that the integration models should be carefully prepared. Otherwise, it should not be surprising, that if assumptions of a model are based on the current literature reports, it is not surprising that their outcomes can be supported by the same literature.

The main results described in this chapter were presented at the 2018 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* and are published in the conference proceedings (Gogolewski *et al.*, 2018a). Additionally, extension of the work to single-cell RNA-seq data was also presented at the *Workshop on Computational Advances for Single-Cell Omics Data Analysis* during the 2018 *IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS)* (Gogolewski and Gambin, 2018).

Interdisciplinary Cooperation for Biomedicine

In the final Chapter 5, we briefly discuss the results from bioinformatical analysis, mostly focused on feature selection in the context of various data from experiments aiming to understand the role of FOXF1 gene in the lungs development as well as origins of the alveolar capillary dysplasia with misalignment of pulmonary veins (ACDMPV) lungs developmental disorder. Additionally, we describe the results from the bioinformatical analysis concerning the effect of FOXF1 knock-in and up-regulation of its transcriptional activity. Finally, we present the comparative analysis of minisatelite distribution in different species and its potential role as a source of DNA replication errors.

All the work from this chapter is a result of interdisciplinary collaboration with the Baylor Collage of Medicine and are part of already published articles: (Dharmadhikari *et al.*, 2014, 2016).

Scientific Collaboration

It should be noted, that in each work that was conducted and presented within this dissertation different co-workers and scientific groups were involved.

The group of prof. Strosznjader from the Mossakowski Medical Research Centre, Polish Academy of Sciences conducted all the lab experiments and provided data from microarray experiments for the case-study presented in Chapter 2. Similarly, prof. Lesyng and his group supported the biological part of this study by interpretation and comments on the results of computational analysis. The group of prof. Stankiewicz from Baylor Collage of Medicine coordinated the research for the purpose of which all the bioinformatical analyses that are presented in Chapter 5 were conducted. Finally, in the cooperation with the group of Le Rouzic, PhD we were working on a side project related to the evolution of mobile genetic elements.

Additionally, individual researchers were involved in conducting the studies presented in this work and published in corresponding articles presented at the end of this chapter. The work presented in (Gogolewski *et al.*, 2017) was supported by biologists W. Wronowska and A. Lech who participated in biological interpretation and description of all results. M. Sykulski helped in the design and implementation of the algorithms in (Gogolewski *et al.*, 2018b), while N.C. Chung introduced the usage of scRNA-seq data. Next, M. Kostecki re-implemented the algorithm used in (Gogolewski *et al.*, 2018a) and helped in the structural analysis of RECON 2.2 model. Finally, prof. A. Gambin mentored, inspired and coordinated all the work presented in the whole dissertation and the published articles based on the dissertation's content.

Here, let us also note, that all figures used in the dissertation come from author's own articles or are his own unpublished graphics.

List of publications of major results from the thesis

Gogolewski K., Wronowska W., Lech A., Lesyng B., Gambin A. (2017) Inferring Molecular Processes Heterogeneity from Transcriptional Data. *Biomedical Research International*, 2017:6961786, doi: 10.1155/2017/6961786.

Gogolewski K., Kostecki M., Gambin A. (2018a) Renal cell carcinoma classification: a case study of pitfalls associated with metabolic landscape analysis. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 96–101, doi: 10.1109/-BIBM.2018.8621194.

Gogolewski K., Sykulski M., Chung N.C., Gambin A. (2018b) Truncated Robust Principal Component Analysis and Noise Reduction for Single Cell RNA-seq Data. In: Zhang F., Cai Z., Skums P., Zhang S. (eds) Bioinformatics Research and Applications. ISBRA 2018. *Lecture Notes in Computer Science*, vol 10847. Springer, Cham, doi: 10.1007/978-3-319-94968-0_32

List of other publications

Dharmadhikari A.V., Gambin T., Szafranski P., Cao W., Probst F.J., Jin W., Fang P., Gogolewski K., Gambin A., George-Abraham J.K., Golla S., Boidein F., Duban-Bedu B., Delobel B., Andrieux J., Becker K., Holinski-Feder E., Cheung S.W., Stankiewicz P. (2014) Molecular and clinical analyses of 16q24.1 duplications involving FOXF1 identify an evolutionarily unstable large minisatellite. *BMC Medical Genetics*, 15:128.

Gogolewski K., Startek M., Gambin A., Le Rouzic A. (2016) Modelling the proliferation of transposable elements in populations under environmental stress. *ArXiv e-prints, Quantitative Biology - Populations and Evolution*, 92-08, J.3, arXiv:1611.04812.

Dharmadhikari A.V., Sun J.J., Gogolewski K., Carofino B.L., Ustiyan V., Hill M., Majewski T., Szafranski P., Justice M.J., Ray R.S., Dickinson M.E., Kalinichenko V.V., Gambin A., Stankiewicz P. (2016) Lethal lung hypoplasia and vascular defects in mice with conditional Foxf1 overexpression. *Biology Open*, 5, 11:1595-1606.

Gogolewski K., Gambin A. (2018) PCA-like Methods for the Integration of Single Cell RNA-seq Data with Metabolic Networks. In: *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences* (ICCABS). Springer, doi: 10.1109/ICCABS.2018.8542112.

Acknowledgements

Although in the scientific sense this part of the dissertation does not bring anything crucial, it creates the space where I would like to express my gratitude to all people that in one way or another were helping me while I was working on a content that constitutes this dissertation.

First and foremost, I want to express my immense gratitude to my supervisor prof. Anna Gambin, who not only coordinated my research throughout the whole PhD studies, but above all was a guide on the paths of science, interdisciplinary cooperation and also life. It was you who helped me to courageously discover what was undiscovered, new and sometimes forgotten.

Next, I want to thank my dear collaborators, co-authors of my academic papers, people who inspired my work, suggested improvements and were always ready to help: Weronika Wronowska, Agnieszka Lech, Maciej Sykulski, Neo Christopher Chung, Marcin Kostecki, Bogdan Lesyng, Paweł Stankiewicz, and again, Anna Gambin.

Furthermore, I would like to express my gratitude to those who were constantly supporting me during my studies. My dearest parents Bogda and Andrzej, my sister Edyta, my niece Kasia and all the other family.

Finally, for all those who were not mentioned by name. Dear Friend, if you are reading this you were probably involved in this big project of mine. Please, take my: 'Thank you!' and be aware, that even by your presence you supported my work!

Undoubtedly, PhD studies would be much harder without financial support and scholarships. Therefore, I would like to also thank the institutions that supported me with funding:

- the National Science Centre for entrusting me with the PRELUDIUM NCN grant no. 2016/21/N/ST6/01507;
- the Institute of Informatics, University of Warsaw and the Warsaw Center for Mathematical Sciences for the scholarships for PhD students.

Additionally, let me mention other National Science Centre grants that I worked in and received financial support from: no. 2012/06/M/ST6/00438, no. 2014/12/W/ST5/00592 and no. 2011/01/B/NZ2/00864.

2

Non-negative Matrix Factorization in Subcellular Heterogeneity

"Peace is not unity in similarity but unity in diversity, in the comparison and conciliation of differences."

— Mikhail Gorbachev

Heterogeneity (gr. ἕτερος - other, another, different; γένος - generation, genealogy) is a general property, that can be considered on many levels, in many different contexts. In particular, heterogeneity is a property that describes many phenomena of molecular biology or medicine. Various mutations at the same loci causing the same phenotype effect are called locus heterogeneity. Consequently, different alleles within the same gene causing similar phenotype response define allelic heterogeneity. Whilst, heterogeneous medical condition in medicine are medical conditions that have several etiologies (e.g. diabetes). In other words, heterogeneity is a diversity within a population of objects, samples or phenomena, in the context of some specific property, or group of properties that is used to distinguish some sub-populations. The measure of heterogeneity can be quantitative or qualitative, however most of the time the aim would be to describe the proportion of sub-populations within population in question. The natural opposite of heterogeneity is homogeneity (gr. ὅμοιος - similar), thus similarity,

or even identity, according to the given characteristic of a population. For the purpose of this chapter, we focus on a population of cells.

2.1. Levels of Heterogeneity in Populations of Cells

Naturally, a population of cells has many levels on which heterogeneity can be observed. The fundamental and most intuitive diversity among group of cells may be determined by their cell type. As an example we can think of a blood sample, but any other tissue can be considered. The most general distinction of blood cells distinguishes: erythrocytes, leukocytes, thrombocytes. However, we can think of more specific divisions, e.g. related to specialization of a cell in the context of signalling cascades of the immune system. These may be grouped into following sub-populations: neutrophils, eosinophils, basophils, lymphocytes, monocytes; within which we can distinguish further divisions.

What is worth noting at this point, is the meaning of an outcome of the gene expression measurement performed on a cell population composed of various cell types. Naturally, we can think of gene expression of a blood sample, but in fact it is an average expression of a heterogeneous population, because each cell follows an average pattern of gene activity related to its cell type. Therefore, the outcome of a gene expression level measurement may be considered as a weighted average of activity of each sub-population defined by its cell type.

So far, in the literature, a couple of methods were proposed to deal with the problem of mixed cell types in biological samples, i.e. tissues. Mostly they are based on the expression matrix decomposition and yield the already mentioned information:

- proportions of different cell types in a given sample;
- expression profiles specific for each detected cell type.

As an example, in (Repsilber *et al.*, 2010) authors introduce the method based on the least squares non-negative matrix factorization for discovery of cell-specific marker genes with noisy signals because of varying cell-type proportions in a sample. The state of the art in computational methods for determination of sample cellular content and cell-specific expression profiles is summarized in (Shen-Orr and Gaujoux, 2013).

Nonetheless, in this chapter, we take another step into the complexity of the cell nature. Let us consider a population of cells that is homogeneous in terms of their common cell type. Now, the unity at the cellular level guarantees that the gene expression measurement is cell type independent or, in other words, cell type specific. However, we already know that throughout its life a cell undergoes specific molecular processes, that are responsible for: growth, intra- and extracellular signalization, metabolism, fate decisions, etc. It is

the activity of molecular processes that may incorporate activity of various groups genes, which will influence the observed transcriptomic profile of the population. We expect the sub-cellular, functional activity to be the main source of the heterogeneity within a homogeneous population of cells with respect to their cell type.

However, if we want to look for the characteristics of molecular processes it is reasonable to use additional knowledge about gene regulatory networks or signalling cascades and pathways. In case of the reconstruction of regulatory networks from mRNA expression data a plethora of methods was proposed, but none of them brought a spectacular success, which proves the complexity of the problem. In particular, (Zhang *et al.*, 2012) presented a method considering the path consistency algorithm based on the conditional mutual information. Also some improvements of the standard path consistency algorithms have already been proposed, such as the elimination of the gene ordering problem (Aghdam *et al.*, 2015). The other approach proposed by (Dojer *et al.*, 2006) successfully applies dynamic Bayesian networks for the gene regulatory network inference based on the perturbed gene expression data.

On the other hand, the results and discoveries collected from numerous studies led to the systematization of our understanding of cell communication methods. So far, there are several knowledge bases summarizing scientific reports in this field. For this reason, we decided to explore the already existing knowledge on the regulome and the signalome to provide an insight into the heterogeneity of molecular processes in a type heterogeneous cell populations.

Therefore, the methodology proposed in this chapter complements the above mentioned ideas for cell type heterogeneity detection by inference about gene regulatory networks activity. Its main objective is to examine the functional heterogeneity of a given cell population sample, assumed to be homogeneous, through the quantification of the intensity of molecular processes occurring in it. Since, the whole idea behind this approach is based on the non-negative factorization of gene expression matrix, let us first briefly describe this family of algorithms.

2.2. Non-Negative Matrix Factorization

Historically, the first approach to decomposition of a matrix into non-negative factors appeared as positive matrix factorization (Paatero and Tapper, 1994) and was popularized as non-negative matrix factorization (NMF) by (Lee and Seung, 1999). In general, NMF techniques can be described as an unsupervised learning paradigm involving the approximative decomposition of a non-negative matrix $A \in \mathbb{R}_+^{m,n}$ into a product of two non-negative matrices, $W \in \mathbb{R}_+^{m,k}$ and $H \in \mathbb{R}_+^{k,n}$. Since, in case of matrices multiplication the dimensions of

the factors may be significantly lower than the product, the value of k may, and is, expected to be less than m and n. This simple observation constitutes the basis and main objective of NMF, to generate factors with significantly reduced dimensions compared to the original data matrix. Additionally, thanks to the non-negativity constraint, the approach obtains the parts-based representation of input data, which corresponds to the general premise, that perception of the whole is based on perception of its parts.

When it comes to specific realizations of this paradigm, there are few ways and restrictions in which NMF can be approached, thus we will shortly comment on them to have an overview of possible realization of this paradigm.

In most cases, NMF is formulated as an approximative, minimization problem: Given an input matrix A and cost function $\mathcal{F}$ find non-negative matrices W and H that minimize the function $\mathcal{F}(V, WH)$. The cost function $\mathcal{F}$ is usually set to the squared error or the Kullback–Leibler divergence for positive matrices and the minimization problem is usually solved using iterative algorithms. There are also more restricted versions, e.g. convex NMF, which restricts the columns of W to convex combinations of the A vectors, which results in improved data representation of W and higher sparsity of H (Thurau *et al.*, 2011). Moreover, constraint may require columns of H to sum up to 1, which can be easily interpreted as a proportion or probability matrix. On the other hand, there are also Bayesian approaches based on a normal likelihood and exponential priors, and derive an efficient Gibbs sampler to approximate the posterior density of the NMF factors (Schmidt *et al.*, 2009).

Even though not all possibilities were described, one can see, that the spectrum of possible approaches to NMF algorithms is broad. However, in this study we will take advantage of approximative NMF, trying to represent the input data A in the form of a product of W and H matrices, both of given low rank k along with a residual. This rank constraint is meant to detect an inner, expected structure of our data. Additionally, it allows to cluster the columns of A into k clusters, using the entries of H as cluster indicators.

The NMF was already used in biomedical data analysis and in particular for analysis of transcriptomic data. Various methods were successfully applied to mRNA microarray data for classification and identification of signatures. In (Brunet *et al.*, 2004) NMF was used to cluster cancer-related microarray data by sample morphology in three different case studies. The algorithm based on non-smooth NMF was proposed by (Carmona-Saez *et al.*, 2006) to identify coherent substructures composed by sets of genes mainly expressed in samples related to the same tumour type. Aiming for estimation of cell-type proportions of heterogeneous tissue samples (Erkkila *et al.*, 2010) proposed Bayesian approach to NMF. (Jia *et al.*, 2015) used discriminant NMF in order to perform gene ranking on RNA-seq data. Also in the context of single-cell transcriptomic data NMF was applied to study the problem of identifying cell-type sub-populations by analysis of their transcriptomic profiles (Shao

and Hofer, 2017), as well as, thorough joint analysis of single-cell RNA-seq and single-cell ATAC-seq data (Duren *et al.*, 2018).

2.3. Aproximation of Molecular Processes Activity

As we can see, NMF has already been used in various aspects of transcriptomic data analysis. Here, we propose an approach, in which we sequentionally put together two already mentioned NMF algorithms that previously were used in the analysis of the proportion of cell types in a heterogeneous population of cells, i.e. tissue. Namely, `DSection` suggested by (Erkkila *et al.*, 2010) and `ssKL` from (Brunet *et al.*, 2004).

Both of these algorithms aimed to find the best approximation of a gene expression matrix $A \in \mathbb{R}_+^{m,n}$ with m genes and n samples as a product of two matrices: $W \in \mathbb{R}_+^{m,k}$ and $H \in \mathbb{R}_+^{k,n}$ of a given rank k. This factorization can be simply formulated as an matrix equation:

$$A = WH + E.$$

Originally, the rank k was related to the number of expected cell types in the n analysed samples. However, by analogy, we reformulate this assumptions, so that they infer the knowledge about the number of molecular processes conducted in the populations of cells composing n type-homogeneous cells. Therefore, a column $W_j = (W_{1,j}, W_{2,j}, ..., W_{m,j})$ of W is expected to describe a process-specific transcriptomic activity pattern, thus $W_{i,j}$ is an expression value of i-th gene for j-th process characteristic profile. Consequently, a column $H_j = (H_{1,j}, H_{2,j}, ..., H_{k,j})$ of H is interpreted as vector of contributions (proportions, weights) of each process to cells population representing the sample, thus $H_{i,j}$ is a linear combination coefficient with which i-th profile contributes to the observed expression of j-th sample. Finally, $E \in \mathbb{R}^{m,n}$ is a residual matrix. In the above setting expression $A_{i,j}$ of i-th gene in j-th sample can be formulated as linear combination of gene activities from each profile:

$$A_{i,j} = \sum_{l=1}^{k} W_{i,l} H_{l,j} + E_{i,j}$$

Let us now shortly describe both NMF algorithms that we make use of in the context of proposed workflow that infers heterogeneity of molecular processes. Both algorithms were originally formulated and dedicated to analyse heterogeneous tissue samples. However, for the purpose of further description of the workflow, they are expressed in terms of molecular processes conducted by cells.

Bayesian NMF

`DSection` algorithm (Erkkila *et al.*, 2010) is an unsupervised approach to the matrix decomposition problem. The only knowledge, that is expected as an input, is the initial, *a priori* proportion of cell molecular activities within analysed sample. The algorithm estimates both cell-specific transcriptome and cell proportions using the Markov Chain Monte Carlo approach.

In the context of our problem, assume expression data matrix $A \in \mathbb{R}^{m,n}$ is given, with n cell-type homogeneous samples, described with expression level measurements of m genes. Each sample is assumed to be a population of cells, out of which each cell is involved in one of the k expected molecular processes, constituting k sub-populations that are characterized by the conducted process. For each sample j the *a priori* knowledge $\mathbf{p}_j = (p_{1,j}, ..., p_{k,j})$ about proportions of cells with specific function is assumed to be given. Then the expression $A_{i,j}$ of the i-th gene in j-th sample is modelled as:

$$A_{i,j} = \mathbf{W}_i \mathbf{p}_j^T = \sum_{l=1}^{k} W_{i,l} p_{l,j} + E_{i,j}$$

where $\mathbf{W}_i = (W_{i,1}, ..., W_{i,k})$ describes the pure expression pattern of i-th gene in each of k cell sub-populations (i.e. representing unique molecular activity) and $\epsilon_{i,j} \sim \mathcal{N}\left(0, \frac{1}{\lambda_i}\right)$ is a normally distributed noise reflecting replication noise with variance $\frac{1}{\lambda_i}$, for some constant λ_i. In other words, $A_{i,j}$ is sum of weighted activities of gene i in all possible molecular processes, plus a normally distributed noise term. In this setting, the likelihood of $A_{i,j}$ can be described as:

$$A_{i,j} | \mathbf{p}_j, \mathbf{W}_i, \lambda_i \sim \mathcal{N}\left(\sum_{l=1}^{k} W_{i,l} p_{l,j}, \frac{1}{\lambda_i}\right)$$

Note that the replication variance, $\frac{1}{\lambda_i}$, is modelled as heteroscedastic across genes and homoscedastic across conducted molecular processes. Finally, assuming i.i.d. measurements of $A_{i,j}$ (which naturally may not always be the case) the joint data ($\mathcal{D} = \sum_{i,j} A_{i,j}$) likelihood in factorized form can be formulated as:

$$f(\mathcal{D}|\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n} f(y_{i,j} | \mathbf{p}_j, \mathbf{W}_i, \lambda_i)$$

where θ represents all model parameters, i.e. $\theta = \bigcup_{i,j} (\mathbf{p_j}, \mathbf{W_i}, \lambda_i)$. Note that the prior p corresponds to the matrix H from the general description of NMF approach introduced in the beginning of this subsection.

More detailed description of further prior specifications, sampler construction and posterior sampling, i.e. probability of acceptance of one step in constructed mixed Gibbs and Metropolis-Hastings sampler can be found in the Methods section and Supplementary Material of (Erkkila *et al.*, 2010). In our approach, the main task of `DSection` algorithm is to

determine the similarities between samples and probable clusters and suggest genes that distinguish each cluster by their activity pattern.

Iterative NMF

On the other hand, the `ssKL` algorithm draws from an approach that is based on iteratively computed approximation of a desired decomposition. In our case the algorithm is supported by a list of marker genes characteristic for each expected molecular function that are meant to be active in the samples of interest and regulate its observed transcriptome. Specifically, the factorization $A \sim WH$, given a desired rank k of W and H, thus expected number of cell sub-populations, starts with a random initialization of W and H matrices which are then iteratively, updated to minimize a divergence functional:

$$\mathcal{F}(A, W, H) = \sum_{i,j} A_{i,j} \log\left(\frac{A_{i,j}}{(WH)_{i,j}}\right) - A_{i,j} + (WH)_{i,j}$$

The approximation in the algorithm is performed in two steps using the coupled divergence equations:

$$H_{i,j} \leftarrow H_{i,j} \frac{\sum_t W_{t,i} \frac{X_{t,j}}{(WH)_{t,j}}}{\sum_k W_{k,i}}$$

$$W_{t,i} \leftarrow W_{t,i} \frac{\sum_j H_{i,j} \frac{X_{t,j}}{(WH)_{t,j}}}{\sum_l H_{i,l}}$$

It should be emphasized that even though the algorithm does not need to necessarily converge to the same solution on each run, see (Brunet *et al.*, 2004) for details, in our case-study the method turned out to provide consistent and robust results throughout repetitive runs.

As it was already stated, both of the above algorithms originally referred to the identification of cell-type specific transcriptomic profiles composing the investigated sample. However, we will show that these methods, can be successfully used along with signalomic knowledge to determine the process activity-related proportions and markers in cell-type homogeneous samples that are characterized by the heterogeneity of undergoing molecular processes.

2.4. Composition of the Workflow

Finally, let us formulate the main result of this chapter. Our main goal, as it was already stated, is to infer some knowledge about the heterogeneity of molecular processes in samples that are homogeneous at the level of cell type and described in terms of their transcriptomic signal. This inference is achieved by estimation of the proportion in which different sub-populations bring their transcriptomic activity to the activity observed in the analysed type-homogeneous sample (i.e population of cells forming this sample).

We assume that there exist at least two different molecular processes, each manifested by different transcriptional activity of a specific cell sub-population conducting a given process. Specifically, the proportion should be understood as a qualitative contribution of each sub-population into the transcriptomic signal observed in the data retrieved from the whole sample.

Along with the proportions, we determine transcription patterns specific for each molecular activity. Again, it should be mentioned that values assigned to the expression of each gene in the patterns are not strictly levels of transcription, in the sense of values gathered, e.g. from microarray assays, but rather correspond to the trends observed in detected sub-populations.

Therefore, to quantify the composition of the sample, we adapted the computational framework called `CellMix` designed originally for the decomposition of a gene expression matrix from heterogeneous samples (Gaujoux and Seoighe, 2013). The framework decomposes an expression matrix into components representing the description of different tissues, that were mixed in the sample. Here, since we assume the homogeneity at the level of cell type (as we will study samples obtained from a cell line), the decomposition is expected to reveal the heterogeneity at the level of molecular processes. Figure 2.1 presents the outline of the `MPH` (molecular processes heterogeneity) method.

The procedure starts with the routine processing of the raw data from the microarray experiments resulting in normalized, filtered (i.e. quality controlled) expressions for each experimental scenario. Then, for the further analysis we select only these genes that differentiate the experimental conditions in the considered study in a statistically significant manner.

The next step is the already introduced `DSection` algorithm (i.e. an unsupervised stage of matrix decomposition) (Erkkila *et al.*, 2010). Since It requires the starting proportions as an *a priori* knowledge for Bayesian model there are two options to provide it: (i) use an expert knowledge about functional activity composition of samples, (ii) generate random proportions table. This step provides an information about specific gene profiles detected during the decomposition of the input expression matrix A. Intuitively, one can think of this step as a form of clustering procedure - the algorithm points out samples described by similar transcription patterns and consequently by potentially similar contributions of functional activities, given the presumed proportion table. From these transcriptomic pattern profiles statistically significant genes are selected, then annotated and validated using DAVID tool (Huang *et al.*, 2009). Finally, these genes constitute a marker list for each profile describing specific functions and/or pathways currently active in the population of cells under the study.

These lists of marker genes are then used as an input to the second decomposition phase

that is based on the ssKL algorithm. These marker genes are meant to provide a transcriptomic pattern that should be followed in case of specific molecular activities. The outcome of this step is twofold:

- a matrix of estimated proportions of functionally homogeneous cell sub-populations in the analysed samples (H),
- a pattern of gene expression per molecular activity defining each sub-populations within sample (W).

Results of this step are validated using the dedicated *in vitro* wet-lab assay (in our case study the MTT assay is used to evaluate the level of proliferation). Finally, marker genes from the previous step were assessed whether they characterize specific functions.

The presented MPH workflow was used to verify a hypothesis that was addressed in the conducted experiment. With this experiment, the aim was to computationally quantify the influence of PJ34 PARP inhibitor on the viability of the C2–ceramide treated cells. The biological hypothesis claims that PJ34 poses a cytoprotective character.

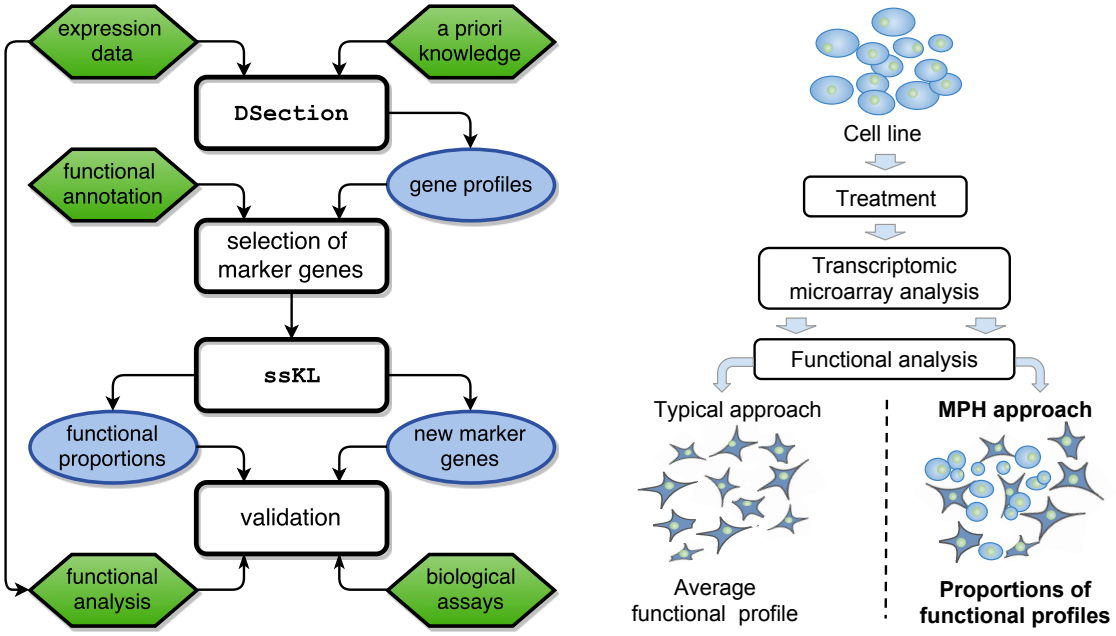
Finally, thanks to our approach that determines function-specific transcriptomic profiles, we provide the most important markers characterizing concrete cellular functions. Here, it should be emphasized that the knowledge about gene regulation in cancerous cells is still quite vague and selection of novel biomarkers for specific cellular processes might improve our understanding of cancer development mechanisms and is the strongest aspect of the proposed worflow.

2.5. Validation: Case-Study of Neuroblastoma Cell Line

In order to verify the proposed methodology we have used the experimental data that investigated the activity of cells in samples derived from neuroblastoma cell line. Various experimental conditions were examined to determine the effect of C2–ceramide and changes resulting from co-treatment with the PARP1 inhibitor: PJ34. First, we provide some details of the wet experiments.

Cell Culture Experiments

The human neuroblastoma cell line SH-SY5Y was obtained from American Type Culture Collection (ATCC); Cells were maintained at 37°C in a humidified incubator containing 5% CO_2, cultured in MEM/F-12 Ham Nutrient Mixtures (1:1) (Biowest) supplemented with 15% heat-inactivated FBS (Cytogen), 1% penicillin/streptomycin and 2mM glutamine (Sigma-Aldrich). Prior to treatment, the cells were cultivated in low serum medium (2% FBS).



(a) The MPH algorithm

(b) Experimental design

Figure 2.1: An overview of the concept presented in this chapter: (a) an insight into the workflow of the MPH method, (b) the general use case of our algorithm.

SH-SY5Y cells were cultivated for 3, 6 and 24 hours in the following experimental settings supplied with: (i) $25\mu M$ C2–ceramide d18:1/2:0 (Enzo Life Sciences); (ii) $25\mu M$ C2–ceramide d18:1/2:0 (Enzo Life Sciences) and $20\mu M$ PARP1 inhibitor PJ34 (Sigma-Aldrich); (iii) pure medium for control samples. The experiment was done in three replicates. After 3, 6 and 24 hours of cultivation cells were collected and RNA was purified using Affymetrix PrepEasy RNA spin KIT. The whole-transcript analysis were conducted using Affymetrix Human Gene 2.1 ST Array Plates. Gene lists were annotated and analysed using QIAGEN's Ingenuity Pathway Analysis tools[1].

After treating SH-SY5Y cells with different compounds, cell viability was evaluated using 2-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide assay (MTT; Thermo Fisher Scientific). Representative samples were collected from each experimental setting, MTT was added to all of the wells. The cells were incubated at 37°C for 2 h, followed by cell lysis and spectrophotometric measurement at 595 nm.

Estimation of Proliferative and Apoptotic Activity

Here, we assume that, the expression level for each gene is the result of an expression activity in two distinct cell sub-populations. In our case study, based on the wet experiment setting, one of them consists of actively dividing, proliferating cells and the second one consists of cells with activated apoptotic signalling pathways. Building on this premise,

[1] www.qiagen.com/ingenuity, IPA®, QIAGEN Redwood City

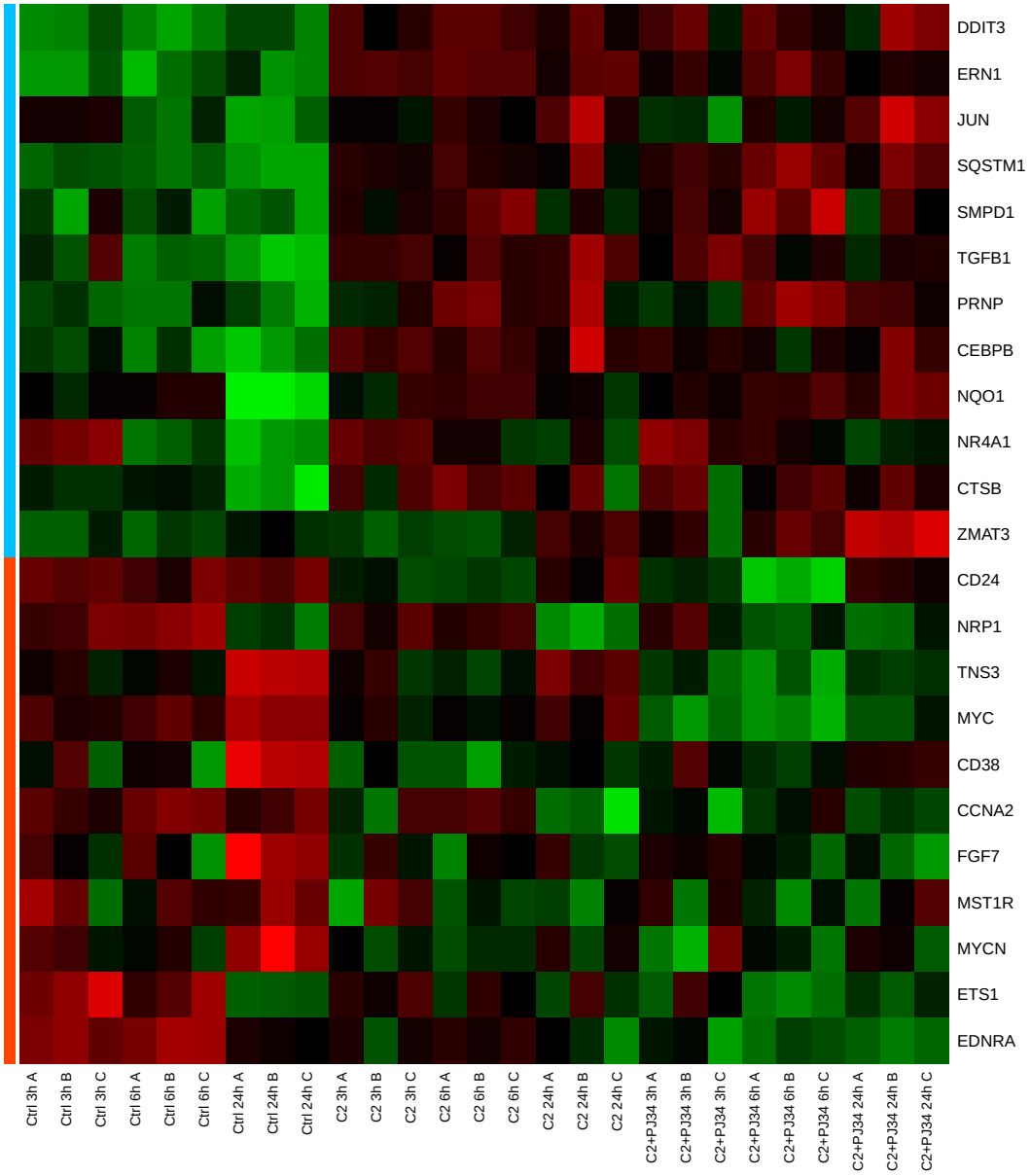


Figure 2.2: Expression patterns of marker genes for the presented case study of neuroblstoma. The panel describes activity of 12 genes related to the positive regulation of the cell death processes (depicted with a blue stripe) and 11 genes related to the positive regulation of proliferation mechanisms (red stripe). Each column corresposnds to specific experimental condition (time point and treatment), which is described by the column label.

the decomposition of the expression matrix yields the proportions of the characterized cells sub-populations.

Marker Genes Selection

In order to assure the computational efficiency of used algorithmic solutions, we first reduce the number of analysed genes. We use only genes that differentiate experimental conditions in a statistically significant way according to the two-sample T-test with p-value less than 0.05. Such filtering resulted in 3983 genes that were used in the further decomposition analysis.

According to the presented workflow, we first answered the question if samples profiles for such selected genes exhibit any specific inner structure. Application of `DSection` algorithm determined a decomposition of the expression matrix. Taking into consideration the character of our experimental data and provided assays, which track specific activity of cells, we set the *prior* knowledge as a uniform proportion of two functional sub-populations constituting the samples.

Annotation of the genes that differentiate well between two obtained profiles (i.e. top 500 genes with the highest fold change), was performed using DAVID tool (Huang *et al.*, 2009). It provides groups of marker genes characteristic for each profile based on the functional enrichment analysis. This classification was systematically verified using literature reports on each gene function. For the subsequent analysis only genes known to enhance particular process were selected.

In the analysed case study of neuroblastoma we end up with the marker genes for: (i) proliferation regulation including 11 genes: CD24, NRP1, TNS3, MYC, CD38, CCNA2, FGF7, MST1R, MYCN, ETS1, EDNRA; (ii) cell death related regulation including 12 genes: DDIT3, ERN1, JUN, SQSTM1, SMPD1, TGFB1, PRNP, CEBPB, NQO1, NR4A1, CTSB, ZMAT3.

The expression patterns of all selected marker genes can be investigated in Figure 2.2, while their full names are given in the Table 2.1. Additionally, a bunch of tests was performed to detect how the outcome of the `MPH` method is dependent on the number of marker genes used in the `ssKL` step. It turned out that a set of 10–12 most differentiating genes between expected sub-populations is sufficient to keep the resulting estimation of their proportion in the sample at the robust level. Therefore, additional marker genes, characterized by lower fold change between sub-populations, were a secondary knowledge for the `MPH` workflow.

Cell death		Proliferation	
Symbol	Name	Symbol	Name
DDIT3	DNA Damage Inducible Transcript 3	CD24	CD24 Molecule
ERN1	Endoplasmic Reticulum To Nucleus Signaling 1	NRP1	Neuropilin 1
JUN	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit	TNS3	Tensin 3
SQSTM1	Sequestosome 1	MYC	V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
SMPD1	Sphingomyelin Phosphodiesterase 1	CD38	CD38 Molecule
TGFB1	Transforming Growth Factor Beta 1	CCNA2	Cyclin A2
PRNP	Prion Protein	FGF7	Fibroblast Growth Factor 7
CEBPB	CCAAT/Enhancer Binding Protein Beta	MST1R	Macrophage Stimulating 1 Receptor
NQO1	NAD(P)H Quinone Dehydrogenase 1	MYCN	V-Myc Avian Myelocytomatosis Viral Oncogene Neuroblastoma Derived Homolog
NR4A1	Nuclear Receptor Subfamily 4 Group A Member 1	ETS1	ETS Proto-Oncogene 1, Transcription Factor
CTSB	Cathepsin B	EDNRA	Endothelin Receptor Type A
ZMAT3	Zinc Finger Matrin-Type 3		

Table 2.1: Lists of functional marker genes that were used in the analysis of Neuroblastoma cells activity. The above two lists are composed of indicator genes known to be characteristic for the cell death and proliferation processes.

Estimation of Functional Heterogeneity

During the next phase, the obtained marker genes sets were used for `ssKL` method to predict the final proportion of different cell activities in the analysed samples. In both case studies we observed significant differences in fractions of cells exhibiting investigated behaviours (see Figure 2.3).

In order to validate presented methods and significance of their outcomes we have performed the log-likelihood ratio test. It was measured how much better the complex model with more degrees of freedom (assuming heterogeneous population) fits the data than the hypothesis assuming homogeneous population (H_0). Namely, assuming that X is a vector of gene expression levels our null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \; X \sim \mathcal{N}(\mu_1, \sigma_1^2) \\ H_1 &: \; X \sim p \cdot \mathcal{N}(\mu_1, \sigma_1^2) + (1-p) \cdot \mathcal{N}(\mu_2, \sigma_2^2) \end{aligned} \tag{2.1}$$

We used the Expectation-Maximization (EM) algorithm to estimate the best parameters and calculate the likelihood for both models.

Since there are only three measurements per experimental condition, statistical confirmation of our results is hard to provide (i.e. on three samples there is no basis for rejection of null hypothesis: sample represents uni-modal normal distribution). Nevertheless, in order to provide at least partial statistical justification of our results, using k-means algorithm, for each experimental condition we have clustered all genes that were used in the analysis.

6 h	Control			C2–cer			C2–cer + PJ34		
Proliferation	0.521	0.526	0.525	0.458	0.458	0.469	0.404	0.423	0.422
Cell death	0.479	0.474	0.475	0.542	0.542	0.531	0.596	0.577	0.578

Table 2.2: The proportion of cells in the population exhibiting specific molecular processes estimated by the `MPH` method.

Such clustering provided larger samples with common expression pattern that could have been analysed with EM algorithm. An effective size of the sample for EM algorithm was set to 100, determining the number of expected clusters in k-means equal to 40.

In each experimental condition for most of the clusters the likelihood ratio test rejected the H_0 in favour of the H_1. For C2–supplied experiment H_0 was rejected 37 (with mean p-value: $6.6 \cdot 10^{-4}$), 39 ($1.2 \cdot 10^{-3}$), 39 ($4.4 \cdot 10^{-4}$) times (out of 40 cases) for 3^{rd}, 6^{th} and 24^{th} hour of experiment, respectively. Similarly, for the above time points in PJ34+C2–supplied experiment H_0 was rejected in 36 ($3.8 \cdot 10^{-4}$), 39 ($3.4 \cdot 10^{-4}$), 38 ($8.0 \cdot 10^{-4}$) cases and for control samples in 38 ($8.2 \cdot 10^{-4}$), 39 ($1.0 \cdot 10^{-3}$), 39 ($1.5 \cdot 10^{-3}$).

In case of C2–supplied experiment the hypothesis of the H_0 model was rejected in favour of the M_1 model in 798 (3h of experiment), 824 (6h), 759 (24h) out of 1000 cases, with mean pvalues $4.244 \cdot 10^{-3}, 3.883 \cdot 10^{-3}, 5.720 \cdot 10^{-3}$ in respective time points.

Same verification procedure was performed for the control samples and PJ34+C2–supplied experiments, all resulting in 751 up to 811 rejected H_0 models in all time points, with mean pvalues from $(4.224 \cdot 10^{-3}, 4.851 \cdot 10^{-3})$ interval.

These results suggest that, indeed, in each experimental condition, it was reasonable to expect that the observed expression level is a composition of signals from different sources.

In most of the analysed experimental variants our results were consistent with the results of biological assays and the literature based predictions.

Concerning the neuroblastoma cell line we have shown that in the control environment the fraction of proliferating cells is higher than in both experimental conditions, which was expected and consistent with MTT assay. Moreover, we observe that the reaction to external treatment is the strongest in the 6^{th} hour time frame, and seems to stabilize around 24^{th} hour.

Interesting fact is related to our results concerning C2–ceramide and C2–ceramide+PJ34 experiments on the fraction of proliferating cells, since it varies from the results obtained using the MTT assay. Our computational method suggests that PARP inhibition along with C2–ceramide supplementation results in lower proliferation abilities than in experiment without the inhibition. This fact was not detected in our MTT assay probably due to its limitations, as well as in the analogous experiment presented in (Czubowicz and Strosznajder, 2014).

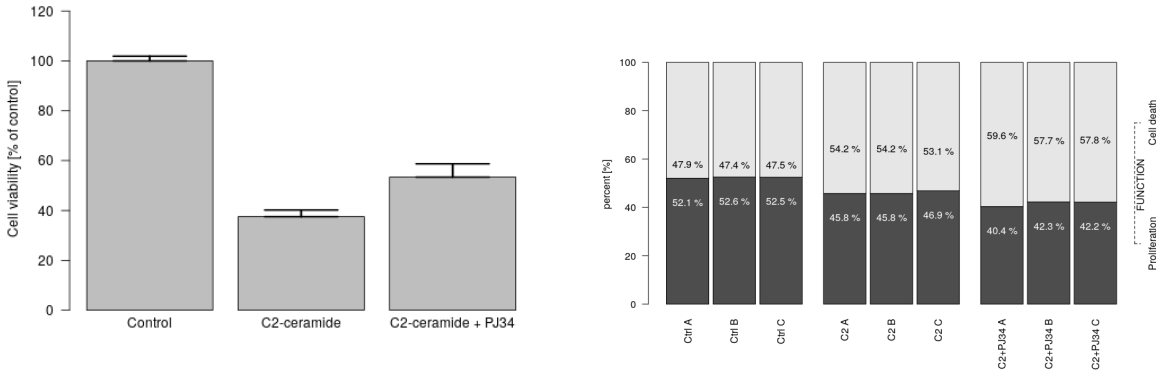


Figure 2.3: The comparison of the experimental results (left) with the theoretical estimation of functional proportions in sub-populations (right). The figure compares behaviour in the 6^{th} hour of the SH-SY5Y neuroblastoma experiment.

The predicted proportions for each sample are presented in Table 2.2. What is worth emphasizing is the behaviour of replicates for given time and setting which is stable (i.e. average standard deviation ≈ 0.01).

Functional Validation

The additional validation of our results was performed using the classical approach to the functional analysis of transcriptomic data. We intend to verify if profiles detected by application of the `MPH` method are consistent with general, average tendencies indicated by functional analysis tools. In order to identify molecular and cellular functions altered by treatments we performed core analysis using Ingenuity Pathway Analysis tool (`IPA`). The statistics for the most influenced functions reported in this section are listed in the Table 2.3.

In our case study 3983 genes were selected for further analysis as distinguishing three experimental conditions, i.e. control, C2–ceramide treated and both C2–ceramide and PJ34 treated.

IPA analysis revealed significant enrichment in the *Cellular Growth and Proliferation* in both C2–ceramide and C2 with PJ34 treated cells this category was on the second position after *Cellular Movement* in C2 variant and *Cellular Development*. According to the p-value in cells treated with C2 and PJ34 *proliferation of tumor cell lines* was the first affected function and *proliferation of neuronal cells* was on the third position. In the C2–ceramide treated cells *proliferation of cells* was on the second position. It should be emphasized that in both variants proliferation scored very low p-value what indicates high data enrichment. On the basis of Z-score its activation was decreased in comparison to control samples. According to activation Z-score in C2–ceramide treated cells the highest decrease was designated for the functions related to the *organization of cytoskeleton* and *cytoplasm*, however *proliferation of cancer cells* was the third decreased function. While increased activation was predicted

Molecular function	Z-score	p-value
C2		
Proliferation of cells	-2.520	4.97E-11
Proliferation of cancer cells	-3.136	2.26E-05
Apoptosis	2.039	2.07E-04
C2 + PJ34		
Proliferation of tumor cell lines	-0.760	7.34E-09
Proliferation of neuronal cells	-1.808	4.70E-08
Cell survival	-2.661	1.87E-03
Proliferation of cells	-2.417	1.77E-06
Necrosis	2.482	1.34E-04
Cell death	3.338	6.48E-03
Apoptosis	3.777	7.44E-04

Table 2.3: Levels of statistical significance of each molecular function that was detected by IPA.

for the functions as: *morbidity or mortality* and *organismal death*, *apoptosis* was the fourth function. In the second experimental variant (C2–ceramide with PJ34) in the top of the relevant decreased functions were those related to cellular movement and development of neurons (e.g. *formation of cellular protrusions, development of neurons*). However activation of *cell survival* and *proliferation of cells* was predicted to be decreased in comparison to control calls. What is more, in this experimental variant *necrosis*, *cell death* and *apoptosis* were in the top of activated functions.

On the basis of the above findings, we conclude that the effect of the C2–ceramide enhanced by the PARP inhibition, affects both the ability of cells to divide, cellular viability as well as genetically controlled reorganization of cellular shape. This intriguing discovery should be subjected to the further analysis.

Discovery of Novel Marker Genes

In the last step of the MPH method we selected new markers which characterize cellular subpopulations of our interest. New markers were selected according to the top differences in genes involvement in the functional expression profiles supported by biological knowledge.

The method revealed that neuroblastoma cells which entered the cell death pathway are characterized by activation of EGR1, VEGF and GDF15. While the EGR1 is known to posses proapoptotic properties (Yu *et al.*, 2007), VEGF and GDF15 are usually described as prosurvival proteins (Mackenzie and Ruhrberg, 2012; Wang *et al.*, 2013). Nevertheless, recently

it was reported that over expression of GDF15 (which is also observed in our experiment) induces apoptosis of breast cancer cells (Kadara *et al.*, 2006). There is also some evidence that VEGF stimulates apoptosis through the ERK1/2 signaling pathway (Ferrari *et al.*, 2012). We believe that these three proteins provide unique markers for the ceramide–induced death of the neural cancer cells.

Simultaneously, some interesting genes were reported as significantly down-regulated in the proliferation profile. One of them is ASCL1, that in gthe literature was reported to play a direct role in promoting progenitor proliferation in the ventral telencephalon (Hardwick *et al.*, 2015), as well as, to coordinate gene expression in both proliferating and differentiating progenitors along the neuronal lineage (Raposo *et al.*, 2015). Furthermore, activation of CHRM3 was shown to promote the proliferation of BPH cells (Wang *et al.*, 2016) and is one of the most silenced genes in the proliferation profile. Based on the gene expression patterns from the W matrix, the most differentiating genes characterizing each functional profile were determined. These genes are listed in the Table 2.4 along with their fold changes based on raw gene expression levels.

Proliferation				Cell Death			
Symbol	FC C2 vs Ctrl	FC C2+PJ34 vs Ctrl	MPH FC	Symbol	FC C2 vs Ctrl	FC C2+PJ34 vs Ctrl	MPH diff
CCNA2	-1,1364	-1,3851	-17,0672	VEGF	2,6949	2,6949	16,6204
CD24	-1,7220	-2,8525	-17,0372	EGR1	5,1068	5,1068	16,5110
EDNRA	-1,5799	-2,6204	-16,6104	TNFRSF12A	1,2625	1,2625	14,5569
TXNIP	-2,9561	-3,4103	-16,2651	**ZMAT3**	1,0242	1,0242	14,3831
ASCL1	-1,3991	-4,0648	-15,1861	GDF15	9,8706	9,8706	14,1468
TNS3	-1,1940	-1,7839	-14,6120	CCL2	1,1454	1,1454	14,0414
TNFAIP6	-1,5649	-3,1068	-13,2426	**TGFB1**	1,5269	1,5269	13,9349
NRP1	-1,4113	-2,2428	-12,1654	**NQO1**	1,1745	1,1745	12,8296
GPR22	1,1253	-2,1723	-11,5563	**CTSB**	1,3108	1,3108	12,8086
MYC	-1,4409	-3,2422	-11,5151	**SQSTM1**	1,5379	1,5379	12,7907
MYCN	-1,1138	-1,1003	-11,4693	ZFP36	2,0750	2,0750	12,5204
CHRM3	-1,3253	-1,8256	-11,1344	NAB2	1,7434	1,7434	11,9886
HAND1	-1,3617	-1,5946	-10,9796	**SMPD1**	1,3307	1,3307	11,7540
DACH1	-1,4785	-1,7276	-10,9221	**CEBPB**	1,6915	1,6915	11,6918
ARRDC4	-1,8176	-2,1476	-10,9033	**ERN1**	1,5739	1,5667	11,5677

Table 2.4: List of top 15 genes characterizing activity of proliferation and cell death processes. Columns contain: gene symbols, fold change of C2 over control and C2+PJ34 over control (based on the transcriptomic data) and activity difference based on function characteristic profiles obtained from MPH method.

Discussion of the Results

In this chapter, a novel approach to the analysis of transcriptomic data from cellularly homogeneous samples aiming to describe their functional heterogeneity was proposed. The main innovation of this idea is based on a simple, yet important, assumption that in a population of cells each individual cell conducts its own molecular function. Consequently, in a type-homogeneous population of cells, the heterogeneity can be found at sub-cellular,

molecular process level. Moreover, each cell in a population may response in a different way to an external stimuli. These observations resulted in the development of the MPH method that estimates the proportion of cell sub-populations conducting specific cellular functions along with their transcriptomic description.

The above description of our results prove that our function characteristic profiles derived from the computational interpretation of transcriptomic data are functionally consistent. We have concluded that they actually correspond to the molecular sub-populations of both ovarian cancer and Neuroblastoma cells that were studied.

Using the MPH method and its results we were able to support the C2–ceramide related hypotheses based on the literature. Both experimental and computational experiments confirm that in the presence of C2–ceramide the cell death processes increase their activity and the fraction of proliferating cells is lower than in the control sample. The same trends were detected between the control and second experimental condition (C2–ceramide + PJ34). The control sample manifests higher proliferation activities and repressed cell death processes. However, our computational analysis suggests that the PARP inhibitor (PJ34) does not increase the cell viability in populations supplied with the C2–ceramide. This fact is not consistent with the results from the MTT assay reported in (Czubowicz and Strosznajder, 2014).

However, since the functional analysis provided in the results section supports our computational results we suppose that the cause for this inconsistence may be the assay used to measure cells viability. The MTT assay depends on concentration and activity of NAD(P)H-dependent cellular enzymes (Bernas and Dobrucki, 2002), while PARP actively reduces the concentration of NAD(P)H (Canto *et al.*, 2015). We suspect that the discrepancy between the results of the MPH method and the MTT assay may result from PARP dependent changes in NAD(P)H balance, PARP related silencing of cellular oxidative stress or might have other unknown cause. We suggest further verification of PJ34 effect on ceramide induced cell death.

Summarizing, the MPH method provides insightful interpretation of transcriptomic data, which is consistent with literature and functional analysis performed by IPA tools. However, in addition it expands our knowledge on the composition of the measured transcriptomic signal by the microarray experiments. We approximate the differences in gene expression levels for cells performing various molecular process and the proportions in which they are mixed.

The successful functional validation justify the applicability of the proposed novel approach to the analysis of transcriptomic data retrieved from homogeneous samples. We illustrate that the methods, used previously at the cellular level to determine the input of each cell type on the finally observed transcriptome, can be applied to the problem of a

sub-cellular, regulatory nature. The method itself is stable, computationally efficient and provides statistically significant outcome.

Finally, it should be noted that methods that engage the non-negative matrix factorization (NMF) techniques are also applicable in case of RNA-seq data analysis. Various modifications of NMF have been lately used for the inference from the RNA-seq results. For simple detection of the most differentiating genes between experimental and control ones, a method called discriminant NMF (DNMF) was suggested. The DNMF incorporates the Fisher's criterion into NMF by maximizing the distance among any samples from different experimental conditions, meanwhile minimizing the dispersion between any pairs of samples in the same class (Jia *et al.*, 2015).

Additionally, the already presented `DSection` method was used by (Dozmorov *et al.*, 2015) to describe the contribution of different cell types in systemic lupus erythematosus (SLE) pathogenesis where transcriptomic data were provided from RNA-seq experiments. Also, it should be mentioned that apart from the NMF there are other approaches to determination of cell-specific transcriptome, e.g. quadratic programming approach suggested by (Gong *et al.*, 2011). Nevertheless, none of these techniques were used to attempt the sub-cellular task that we address in this chapter. That is why we believe that the `MPH` approach is not only novel, but also highlights the fact that transcriptome analysis usually is based on multi-level assumptions and simplifications, which may obfuscate some subtle facts important in many biomedical aspects (e.g. diagnosis, drug resistance etc.).

2.6. Conclusions

In this study, we presented a novel methodological approach to quantify the functional heterogeneity of a homogeneous cell population based on transcriptomic data. To reach this aim, we adopted the methods originally proposed for quantification of cell proportions in heterogeneous tissue samples (e.g. mixed tissues) from expression data. Our model framework exploits the methodology designed for RNA expression microarrays applied for heterogeneous tissues. However, it should be emphasized that our novel approach can also be effectively applied to RNA-seq data by adapting the procedures proposed in (Li and Xie, 2013; Gong and Szustakowski, 2013).

The presented and discussed case study was focused on the role of C2–ceramide in mediating cell death processes in neuroblastoma cells. Our computational method quantified the activity of C2–ceramide in time and its influence on cell metabolic activity in a consistent way with the MTT assay. Finally, the method allowed to obtain biologically relevant results describing the biologically meaningful interdependence between C2–ceramide and the PJ34 compound. In this case, the specificity of molecular experiment, requires the use

of additional methods to verify the results of the MTT assay referring to cells viability.

Here, it should be mentioned that the presented framework could be further improved. Both `DSection` and `ssKL` procedures are characterized by the bias towards discrimination based mainly on fold changes in a given expression data set. One should note, that in the current implementation the `ssKL` method prevents the use of markers that differ on the direction of expression changes. The appropriate improvement of the decomposition method can constitute a very useful strategy for further research.

3

Low-rank Matrix Estimation of Transcriptomic Signal

"Everything should be made as simple as possible, but no simpler."

— Albert Einstein

Sequencing technologies are constantly improving and, thanks to that, new and more efficient methods dedicated to the field of transcriptomics are emerging. In the previous chapter, we discussed the problem of gene expression measurements performed on a bulk sample, that is a sample composed of many different cells. In such case, the variability among cells may occur on different levels. In the literature, the most explored field covers the problem of a sample consisting of various cell types e.g. tissue. Moreover, it is known that the transcriptomic pattern differs between cell types. Therefore, the gene expression level measurements performed on different samples, collected from the potentially same tissue site, may vary depending on the proportion of different cell types in these samples. On the other hand, this variability may be detected only at the sub-cellular level related to the activity of molecular processes and functions. Our remedy to the second possibility is the `MPH` method that helps to track the functional composition of sample, quantitative differences in proportions of cells and the most characteristic genes for a molecular process.

As already mentioned, the improvement of the high-throughput technologies is significant and reached the level of single cell resolution. Therefore, all computational methods, that aim to reconstruct and understand the composition of transcriptomic signal derived from a bulk sample, can be substituted with more adequate experimental techniques of gene expression measurements. One of the latest experimental alternatives is the single-cell RNA sequencing (sc-RNA seq), which has an unprecedented capability to evaluate the RNA content of an individual cell. This property naturally makes it possible to describe differences between various cell types at more precise level.

However, as usual in such cases, a new type of data entails the possibility, or sometimes necessity, of introducing new methods and techniques for their processing and analysis. In this chapter an algorithm dedicated to the analysis of scRNA-seq data is introduced along with a data analysis case study. The main aims of the algorithm is to retrieve the regularity of genes activity among cells, as well as, to filter out those probes or samples, that break the expected pattern of expression.

3.1. High-throughput Technologies in Transcriptomics

Modern medicine and biology rely on the constant progress among the high-throughput technologies. There exists a vast pallet of *-omics* fields, that try to deepen our understating of the complex biological world including transcriptomic, genomic, metabolomic, proteomic, etc. Thanks to huge amounts of data derived from high-throughput experiments in all of these fields and suitable statistical methods applied to them, we are able to draw conclusions about various biological phenomena on many levels. Starting from organism development, its evolution and proliferation (Fritz *et al.*, 2013; Beller and Oliver, 2006) through cell specialization events, fate decisions and its cycle progression (von Kriegsheim *et al.*, 2009; Usoskin *et al.*, 2015; Chu *et al.*, 2016; Qiu *et al.*, 2017) up to genome dynamics, its organization and regulatory mechanism (Wang *et al.*, 2010; Ewing and Kazazian, 2010; Tang *et al.*, 2009). Among multiple fields that benefit from advances in the Next Generation Sequencing (NGS), transcriptomics is the branch that has developed significantly (Spies and Ciaudo, 2015).

As we have discussed in the previous Chapter, conventionally, bulk RNA-seq data, that are analysed using microarray or RNA-seq techniques, provide mean gene expression values from a large number of cells in that biological sample. However, a mixture of multiple cells that often have different functions or origins may hide relevant information, carry high variance related to their cellular composition, and might not be reproducible in separate studies (Novelli *et al.*, 2008; Wills *et al.*, 2013; Gogolewski *et al.*, 2017).

Nonetheless, during the last few years there was a notable progress in a field of single

cell experiments and in particular single cell RNA-seq (scRNA-seq) where introduction of new protocols allowed scientists to determine a transcriptome of a single cell (Tang *et al.*, 2009; Wang and Navin, 2015; Liu and Trapnell, 2016; Ramskold *et al.*, 2012). A generalized protocol of scRNA-seq experiments involves following steps:

- Isolation of individual cells from a provided sample;
- Extraction of the RNA content and whole-genome-amplification (WGA);
- Construction of sequencing libraries;
- Sequencing using a chosen next-generation sequencer.

Each of these steps can be conducted in few alternative ways, which are described and compared in details in the literature (Shapiro *et al.*, 2013; Wang and Song, 2017; Ziegenhain *et al.*, 2017). In order to have an idea about how the experimental pipeline works and how the biological material is processed into data see the Figure 3.1 that summarizes it.

Thanks to scRNA-seq we are now able to perform statistical data analysis and more accurately describe the transcriptional activity of individual cells, pointing out important or even crucial behaviors concerning cell fate decisions, specialization or cycle progression according to its type or origin (Ramskold *et al.*, 2012).

3.2. Unsupervised Data Analysis for Single Cell RNA-seq

So far, in the literature there are described many studies that have carried out a broad analysis of scRNA-seq data from different perspectives. One of the most intensively explored approaches are the unsupervised techniques which allow to: reduce the dimensionality and visualize the data (Bartenhagen *et al.*, 2010; DeTomaso and Yosef, 2016), deduce the inner structure of the data often by finding a latent structure or variables, clustering and discovering other sources of variability generated by unknown mechanisms (e.g. biological or technical) (Usoskin *et al.*, 2015; Ramskold *et al.*, 2012; Li *et al.*, 2009; Chung and Storey, 2015; Ilicic *et al.*, 2016).

First of all, using the relatively basic, but very popular and constantly studied Principal Component Analysis (PCA) method was sufficient for discovery of various cell biology facts. The PCA was used through clustering to track Definite Endoderm cells and explain their linage from Embryonic Stem Cells (Chu *et al.*, 2016). Similarly, the classification of sensory neuron types was done using PCA (Usoskin *et al.*, 2015). In (Ilicic *et al.*, 2016), authors use PCA to find potentially damaged cells. The other linear unsupervised methods were also used to analyze scRNA-seq data: SVD in discovery of melanoma circulating

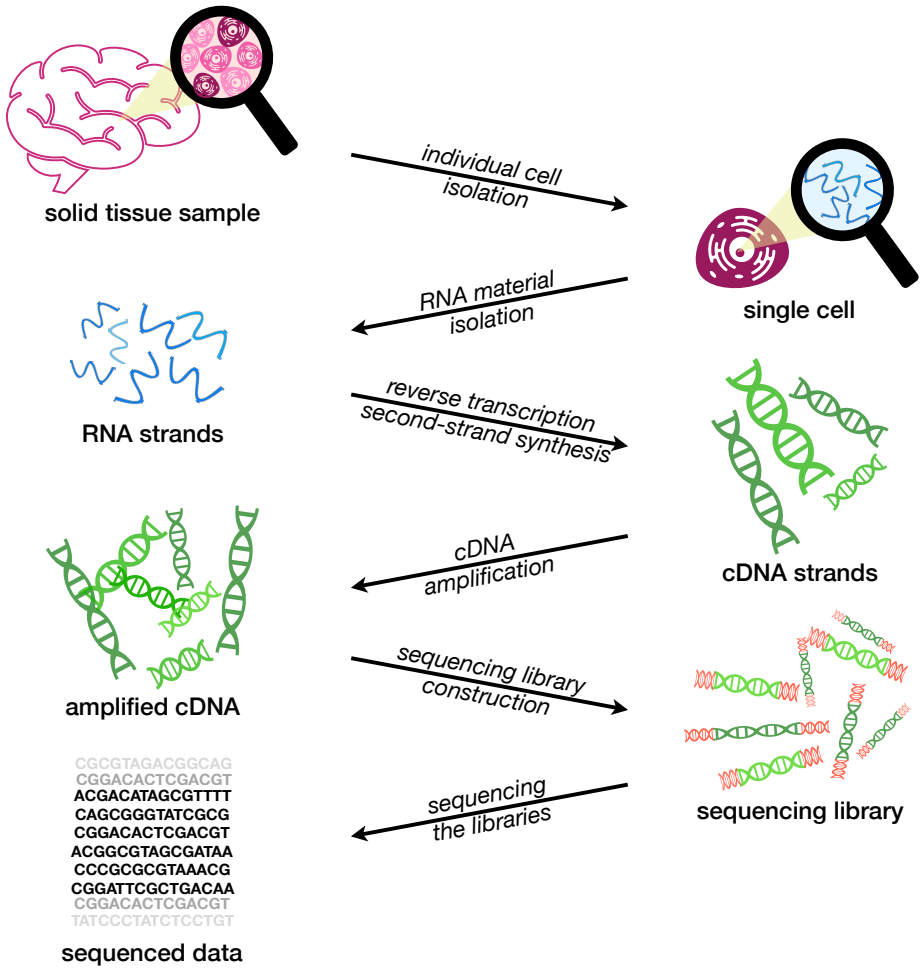


Figure 3.1: The general pipeline of the single cell RNA seq experiments. The figure presents main steps in a protocol for the single cell sequencing. First, cells need to be isolated, then the RNA content needs to be retrieved. Next, the RNA material is reverse transcribed and turned into cDNA, to be finally amplified. The amplified genetic content is then gathered into a library and sequenced. Finally, the sequenced data are ready for the further bioinformatical and statistical analysis.

tumor cells marker genes techniques (Ramskold *et al.*, 2012), ICA in describing the trajectories of cell fate decisions (Trapnell *et al.*, 2014). Furthermore, non-linear methods such as reversed graph embedding (RGE), t-distributed stochastic neighbor embedding (tSNE) were used to classification (Qiu *et al.*, 2017) but also visualization (DeTomaso and Yosef, 2016) of scRNA-seq data. Finally, there are also other matrix factorization methods. Non-negative factorization was used for cell types classification (Shao and Hofer, 2017; Zhu *et al.*, 2017) and pseudo-time cells ordering (Wang *et al.*, 2017b), while (Witten *et al.*, 2009) use the penalized matrix decomposition to correlate gene activity with DNA copy number change in breast cancer.

However, to the best of our knowledge, there does not exist a matrix decomposition method that provides information about truncated low-rank component that explains the most significant variability in the data along with its sparse signal and remaining low-importance signal or noise. Furthermore, robust PCA (`RPCA`) proposed in (Candès *et al.*,

2011) as well as extensions presented in this chapter have not been applied to scRNA-seq data. The promising results of our study justify the usefulness of this novel approach, especially in the context of the analyzed high-dimensional scRNA-seq data.

In the next section, the detailed description of RPCA algorithm is introduced along with novel extensions that are suitable for analysis of transcriptomic data.

3.3. Truncated Robust PCA with L_2 Regularization

PCA is one of the most popular methods for dimension reduction and unsupervised learning. Given a dataset A containing m samples described by n variables, the main objective of PCA is to find a linear transformation, which maps each sample from the matrix A onto a new coordinate system. In this new system, the coordinates, corresponding to principal components, are ordered by decreasing variances explained. With such representation, we can reduce the dimensionality of our data with a minimal loss of information as well as determine important sources of variability. However, PCA has its limitations. With an increasing size and sparsity of genomic data, PCA becomes inefficient. Furthermore, the outcome of PCA may be easily biased by outliers, which is not an expected behaviour. The following extensions are an attempt to reduce these limitations during the analysis of high-dimensional data.

Robust PCA.

Our work is based on the decomposition algorithm proposed by (Candès *et al.*, 2011) called robust PCA (RPCA). The aim of the RPCA is to decompose the input matrix A, into low-rank matrix L and sparse matrix S components. Simultaneously, the algorithm should minimize the following optimization problem:

$$\min_{L,S} ||L||_* + \lambda_1 ||S||_1, \text{ where } A = L + S$$

Here we denote $||A||_*$ as the nuclear norm of matrix A and $||A||_1$ as the first norm of a vectorized A matrix which are given by the following formulas:

$$||A||_* = \sum \sigma_i = \text{tr}\left(\sqrt{AA^T}\right)$$
$$||A||_1 = \sum_{i,j} |a_{ij}|$$

In their work, authors discuss the assumptions that matrix A should follow for the decomposition to exist. Moreover they prove that the parameter λ_1 can be set to $1/\sqrt{\min(m,n)}$, where m, n are dimensions of the input matrix A, which, under weak probabilistic assumptions, guarantees a proper decomposition into low-rank and sparse components as

$m, n \to \infty$ (Candès *et al.*, 2011). However, it is shown that the spectrum of feasible values of λ_1 parameter is broader.

In order to solve the above optimization problem, as proposed in (Yuan and Yang, 2009), we use an implementation of a special case of the Alternating Directions method, which belongs to a more general class of augmented Lagrangian (AGL) multiplier algorithms. In general, the approach is based on minimizing the following AGL operator with respect to L and S matrices alternately:

$$l(L, S, Y) = ||L||_* + \lambda_1 ||S||_1 + \left\langle Y, A - L - S \right\rangle + \frac{\mu}{2} ||A - L - S||_F^2$$

where Y is the Lagrange multiplier matrix, the inner product of matrices $\left\langle \cdot, \cdot \right\rangle$ is defined as the trace of their product, i.e. $\left\langle A, B \right\rangle = \text{tr}(AB^T)$, $||A||_F$ is the Frobenius norm of the form $||A||_F = \sqrt{\sum_{i,j} a_{i,j}^2}$ and μ is the penalty coefficient.

The outline of the solution is presented in the Algorithm 1, in which two shrinkage operators are used:

$$\mathcal{S}_\tau(x) = \text{sgn}(x) \cdot \max(|x| - \tau, 0)$$
$$\mathcal{D}_\tau(X) = U \mathcal{S}_\tau(\Sigma) V^*$$

where τ is the shrinkage threshold value and $U\Sigma V^*$ is the `SVD` of matrix X. Operator $\mathcal{S}_\tau$ when applied to a matrix is equivalent to $\mathcal{S}_\tau$ applied to each of the matrix elements.

In case of initialization of the μ parameter and convergence condition, we set $\mu = \frac{m \cdot n}{4 \cdot ||A||_1}$, as suggested in (Yuan and Yang, 2009) and terminate the algorithm when $||A - L - S||_F \leq \delta ||A||_F$ where $\delta = 10^{-7}$. The implementation of the `RPCA` algorithm, that we further extend in this work, is publicly available, stable R package in CRAN repository (Sykulski, 2015).

Algorithm 1 RPCA by Alternating Directions

1: **procedure** RPCA(λ_1)
2: $S_0, Y_0 \leftarrow 0; \mu > 0$
3: **while** not converged **do**
4: compute $L_{i+1} = \mathcal{D}_{\mu^{-1}}(A - S_i + \mu^{-1} Y_i)$
5: compute $S_{i+1} = \mathcal{S}_{\lambda_1 \mu^{-1}}(A - L_{i+1} + \mu^{-1} Y_i)$
6: compute $Y_{i+1} = Y_i + \mu \cdot (A - L_{i+1} - S_{i+1})$
7: **end while**
8: **end procedure**

Truncated Version of RPCA.

First, we consider a truncated version of the algorithm, which calculates the L matrix in the $L + S$ decomposition in such a way that it is of a given rank k or the lowest possible

rank grater than k_0, for which the problem has a solution that meets all its criteria. In order to achieve that behavior, we use the truncated version of SVD (implementation from the *irlba* R package (Baglama *et al.*, 2018)) instead of a full SVD and iteratively modify the μ parameter according to the following rule:

$$\mu_{i+1}^{-1} = \max(c \cdot \mu_i^{-1}, \sigma_{k+1})$$

where σ_k is the k-th singular value from the truncated SVD and $c < 1$ is the AGL constraints penalty growth rate.

The change of μ is significant for the algorithm convergence. As μ_i^{-1} decreases, both threshold operators shrink less elements in S and singular values of L. Furthermore, the increase of the penalty coefficient for $A = L + S$ speeds up the convergence. However, in theory, AGL algorithm converges to the constraint problem even when $\mu_i^{-1} \not\to 0$. Simultaneously, when μ_{i+1}^{-1} is set to the value of σ_{k+1} we increase k, i.e. the number of computed SVD vectors, which is the expected rank of L matrix in i-th iteration of the algorithm.

Algorithm 2 truncated-RPCA

procedure tRPCA(λ_1, k_0, c)
  $S_0, Y_0 \leftarrow 0$; $\mu_0 > 0$; $k = k_0$
  while not converged **do**
    compute $L_{i+1} = \mathcal{D}_{\mu_i^{-1}}(A - S_i + \mu_i^{-1} Y_i)$
    compute $S_{i+1} = \mathcal{S}_{\lambda_1 \mu_i^{-1}}(A - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $Y_{i+1} = Y_i + \mu_i \cdot (A - L_{i+1} - S_{i+1})$
    compute $\mu_{i+1}^{-1} = \max(c \cdot \mu_i^{-1}, \sigma_{k+1})$
    if $\mu_i^{-1} == \sigma_{k+1}$ **then** increase k
    end if
    end while
  end procedure

The Algorithm 2 significantly reduces the computation time compared to the original RPCA, while preserving its accuracy. However, in the case of biomedical data, the decomposition into low-rank and sparse matrices is not always feasible or easily obtainable. The input matrix usually has more than a few k significant singular values that may come from biological activities, technical reasons, or other unknown sources. This prevents the recovery of low-rank component as when subtracted from the input A matrix, they do not constitute a sparse matrix. We may interpret these perturbations in the L matrix as a noise or low-importance information. Since it does not have a sparse nature, we extend the decomposition into $L+S+E$, where the matrix E contains a dense noise controlled for using the L_2 norm on vectorized matrix A (i.e. Frobenius norm).

Noise Reduction for tRPCA.

In order to relax the assumptions on the input matrix we introduce an additional matrix E to the decomposition. Now, the extended problem can be reformulated as follows:

$$A = L + S + E$$
$$\min_{L,S,E} ||L||_* + \lambda_1||S||_1 + \lambda_2||E||_F$$

In this setting, the E matrix is meant to contain the information of low importance or noise, which is mostly carried by the lowest singular values in the SVD decomposition of L matrix. To solve this problem we extend the Alternating Directions approach and we minimize the newly defined AGL operator also with respect to the E matrix:

$$\begin{aligned} l(L,S,E,Y) = \quad & ||L||_* + \lambda_1||S||_1 + \lambda_2||E||_F + \\ & + \left\langle Y, A - L - S - E \right\rangle + \frac{\mu}{2}||A - L - S - E||_F^2 \end{aligned}$$

In order to find E matrix that minimizes l, we solve $\frac{\partial l}{\partial E} = 0$. First, since $\frac{\partial ||L||_*}{\partial E} = \frac{\partial ||S||_1}{\partial E} = 0$ we can write:

$$\begin{aligned} \frac{\partial l(L,S,E,Y)}{\partial E} = \quad & \lambda_2 \frac{\partial ||E||_F}{\partial E} + \frac{\partial \left\langle Y, A - L - S - E \right\rangle}{\partial E} \\ & + \frac{\mu}{2} \frac{\partial ||A - L - S - E||_F^2}{\partial E} \end{aligned}$$

Let us reduce each of the partial derivatives. By definition of the Frobenius norm, for any i, j we can write:

$$\frac{\partial ||E||_F}{\partial E_{i,j}} = \frac{\partial \sqrt{\sum_{i,j} E_{i,j}^2}}{\partial E_{i,j}} = \frac{1}{2\sqrt{\sum_{i,j} E_{i,j}^2}} \cdot 2E_{i,j} = \frac{E_{i,j}}{||E||_F}$$

thus, consequently, for E we have:

$$\frac{\partial ||E||_F}{\partial E} = \begin{bmatrix} \frac{\partial ||E||_F}{\partial E_{1,1}} & \frac{\partial ||E||_F}{\partial E_{1,2}} & \cdots & \frac{\partial ||E||_F}{\partial E_{1,n}} \\ \frac{\partial ||E||_F}{\partial E_{2,1}} & \frac{\partial ||E||_F}{\partial E_{2,2}} & \cdots & \frac{\partial ||E||_F}{\partial E_{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial ||E||_F}{\partial E_{m,1}} & \frac{\partial ||E||_F}{\partial E_{m,2}} & \cdots & \frac{\partial ||E||_F}{\partial E_{m,n}} \end{bmatrix} = \begin{bmatrix} \frac{E_{1,1}}{||E||_F} & \frac{E_{1,2}}{||E||_F} & \cdots & \frac{E_{1,n}}{||E||_F} \\ \frac{E_{2,1}}{||E||_F} & \frac{E_{2,2}}{||E||_F} & \cdots & \frac{E_{2,n}}{||E||_F} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{E_{m,1}}{||E||_F} & \frac{E_{m,2}}{||E||_F} & \cdots & \frac{E_{m,n}}{||E||_F} \end{bmatrix} = \frac{E}{||E||_F}$$

Next, let us observe that based on additive property of the matrix transposition operator ($(A + B)^T = A^T + B^T$) and the fact that the trace tr is a linear operator ($\text{tr}\,(cA + B)^T = \text{tr}\,(A) + \text{tr}\,(B)$) we can write:

$$\begin{aligned} \left\langle Y, A - L - S - E \right\rangle &= \text{tr}\left(Y\left(A - L - S - E\right)^T\right) = \\ &= \text{tr}\left(Y\left(A^T - L^T - S^T - E^T\right)\right) = \\ &= \text{tr}\left(YA^T - YL^T - YS^T - YE^T\right) = \\ &= \text{tr}\left(YA^T\right) - \text{tr}\left(YL^T\right) - \text{tr}\left(YS^T\right) - \text{tr}\left(YE^T\right) \end{aligned}$$

three of these components are independent on E, thus their derivatives with respect to E are $\frac{\partial \operatorname{tr}(YA^T)}{\partial E} = \frac{\partial \operatorname{tr}(YL^T)}{\partial E} = \frac{\partial \operatorname{tr}(YS^T)}{\partial E} = 0$ and we can write:

$$\frac{\partial \langle Y, A - L - S - E \rangle}{\partial E} = \frac{\partial \operatorname{tr}\left(YE^T\right)}{\partial E} = Y$$

because the k-th element on a trace of YE^T is of the form $\sum_{l=1}^{n} Y_{k,l}E_{l,k}$ and thus the i-th row and the j-th column of $\frac{\partial \operatorname{tr}(YE^T)}{\partial E}$ has the value:

$$\frac{\partial \sum_{k=1}^{m}\sum_{l=1}^{n} Y_{k,l}E_{l,k}}{\partial E_{i,j}} = \sum_{k=1}^{m}\sum_{l=1}^{n}\frac{\partial Y_{k,l}E_{l,k}}{\partial E_{i,j}} = Y_{i,j}$$

Finally, the last component is simply derivative of quadratic function since the square of the Frobenius norm of a matrix is just a sum of squares of the matrix, thus:

$$\frac{\partial ||A - L - S - E||_F^2}{\partial E} = \frac{\partial \sum_{i,j}\left(A_{i,j} - L_{i,j} - S_{i,j} - E_{i,j}\right)^2}{\partial E} = 2\left(A - L - S - E\right)$$

Gathering it all together we have:

$$\frac{\partial l(L, S, E, Y)}{\partial E} = \lambda_2 \frac{E}{||E||_F} + Y + \mu\left(A - L - S - E\right)$$

equating this value to 0, to find E minimizing the l operator, and ordering the equation we get:

$$E\left(\frac{\lambda_2}{||E||_F} + \mu\right) = Y + \mu(A - L - S)$$

Now, let $C = Y + \mu(A - L - S)$. Then $\exists_{d\in\mathbb{R}}\, E = d \cdot C$ and we can reformulate and simplify the equation:

$$d \cdot C\left(\frac{\lambda_2}{||d \cdot C||_F} + \mu\right) = C$$

$$C\left(\frac{d \cdot \lambda_2}{||d \cdot C||_F} + d \cdot \mu - 1\right) = 0$$

Assuming that $C \neq 0$ we have

$$\frac{d \cdot \lambda_2}{|d|||C||_F} + d \cdot \mu = 1$$

Now we finally, determine the value of d. First, if $d < 0$ we get a contradiction, because from $-\frac{\lambda_2}{||C||_F} + d \cdot \mu = 1$ we calculate d as $d = \frac{||C||_F + \lambda_2}{\mu||C||_F} > 0$. On the other hand, for $d > 0$ we have:

$$d = \frac{||C||_F - \lambda_2}{\mu||C||_F} = \frac{1}{\lambda_2}\left(1 - \frac{\lambda_2}{||C||_F}\right) > 0$$

which holds for $||C||_F \geq \lambda_2$.

Therefore, the matrix that minimizes the operator l with respect to E has the following analytical formula:

$$E = \frac{1}{\mu}\left(1 - \frac{\lambda_2}{||Y + \mu(A - L - S)||_F}\right) \cdot (Y + \mu(A - L - S))$$

Based on this observation we define an operator:

$$\mathcal{E}_\tau(X) = \max\left(0, 1 - \frac{\tau}{||X||_F}\right) \cdot X$$

which describes how to determine the matrix E which minimizes the l operator.

Finally, we extend the algorithm of `tRPCA` by applying the defined operator $\mathcal{E}_\tau$. In our approach, we apply the operator twice, both, after minimization with respect to L and S, in order to filter out the potential mismatched components from both matrices. It is worth to emphasize, that in case of large $\lambda_2 > ||C||_F$ we end up with the previously introduced `tRPCA` procedure. Moreover, in every iteration we adjust k parameter to be a minimal value such that $\mathcal{D}_{\mu^{-1}}$ operator can be properly applied. Algorithm 3 presents the pseudo-code of the whole decomposition procedure, which we call `tRPCAL2`.

Algorithm 3 truncated-RPCA with L2 regularization

procedure tRPCAL2(λ_1, $\boldsymbol{\lambda_2}$, k_0, c)
  $S_0, Y_0, E_0 \leftarrow 0$; $\mu_0 > 0$; $k = k_0$
  while not converged **do**
    compute $L_{i+1} = \mathcal{D}_{\mu_i^{-1}}(A - S_i - E_i + \mu_i^{-1} Y_i)$
    compute $E^*_{i+1} = \mathcal{E}_{\lambda_2 \mu_i^{-1}}(A - S_i - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $S_{i+1} = \mathcal{S}_{\lambda_1 \mu_i^{-1}}(A - E^*_{i+1} - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $E_{i+1} = \mathcal{E}_{\lambda_2 \mu_i^{-1}}(A - S_{i+1} - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $Y_{i+1} = Y_i + \mu_i \cdot (A - E_{i+1} - L_{i+1} - S_{i+1})$
    compute $\mu_{i+1}^{-1} = \max(c \cdot \mu_i^{-1}, \sigma_{k+1})$
    if $\mu_i^{-1} == \sigma_{k+1}$ **then** increase k
    else $k = 1 + \text{argmax}_j\,(\sigma_j > \mu_{i+1}^{-1})$
    end if
  end while
end procedure

Low-Rank Matrix Clustering.

In order to examine the resulting decomposed matrix $A = L + S + E$ we use the following clustering procedure. Since L is a low-rank matrix (of rank k) with a known `SVD` decomposition $L = U\Sigma V^*$, we cluster all cells by their k dimensional representation $U\Sigma$ using the K-means algorithm, with the most suitable number of clusters (Macqueen, 1967; Hartigan and Wong, 1979; Lloyd, 1982). To visualize the clustering outcome in 2-dimensions, we apply the t-SNE algorithm (van der Maaten, 2014).

3.4. Single Cell Transcriptomic Data

In this study, we analysed the publicly available single cell RNA-seq datasets by 10x Genomics (https://www.10xgenomics.com/solutions/single-cell/). Specifically, our results were obtained using the scRNA-seq datasets experiments performed on peripheral blood mononuclear cells (PBMCs) from a healthy donor. PBMCs are primary cells with relatively small amounts of RNA (1pg RNA/cell). The final dataset contains 2.7k individual single cells, sequenced on Illumina NextSeq 500 with approx. 69k reads per cell.

Along with the 2.7k PBMCs dataset, we have used the scRNA-seq data retrieved from homogeneous samples of specific cell types that constitute the PBMC sample. Each type-specific dataset has over 90% of purity for each subtype by Fluorescence Activated Cell Sorting (FACS) (Basu *et al.*, 2010). The transcriptomes were used in (Zheng *et al.*, 2017) and described the following cell types and subtypes: $CD14^+$ Monocytes, $CD56^+$ Natural Killer cells, $CD19^+$ B cells, $CD34^+$ cells and subfamilies of T cells: $CD8^+$ Cytotoxic T cells, $CD8^+/CD45RA^+$ Naive Cytotoxic T cells, $CD4^+/CD45RO^+$ Memory T cells, $CD4^+$ Helper T cells (see Fig. 3.2).

Each of the above datasets is given in the form of a count matrix A, where the i-th row represents a gene and the j-th column represents an individual cell. The value of a_{ij} is the number of counts of the i-th gene for the j-th cell. Since our method is meant to filter out the sparse signal in S and the dense noise in E, we do not apply the typical quality control step. All cells are used in the analysis and we expect all perturbations (e.g. biological or technical outliers or fluctuations) that break the linear behavior to be captured by $S + E$ component of the decomposition.

Additionally, for each dataset, we filter out genes that had zero counts for all cells in a given set. Finally, the number of counts for each cell was normalized by its total number of counts and log-scaled. Furthermore, on the processed 2.7k PBMCs data matrix is consequently denoted as A. Out of over 32k genes, 16634 genes that had non-zero number of counts mapped for at least one cell are retained.

Test Set Construction.

In order to test our method, we set the labelling of cells from the PBMCs dataset. For each available type-specific dataset, we calculate its average transcriptome. However, since the correlation between averaged subtype-specific transcriptomes within T-cell family is relatively high, for the purpose of this work, we label the cells with one of the five possible types: (i) Monocytes, (ii) Natural Killers, (iii) B cells, (iv) T cells, (v) Unknown. T cells family transcriptome is designated as an average among all T cells subtypes transcriptomes.

The criteria for labelling consist of two conditions. First, a cell is assumed to be of an unknown type if it does not correlate with any of the given profiles with a Pearson

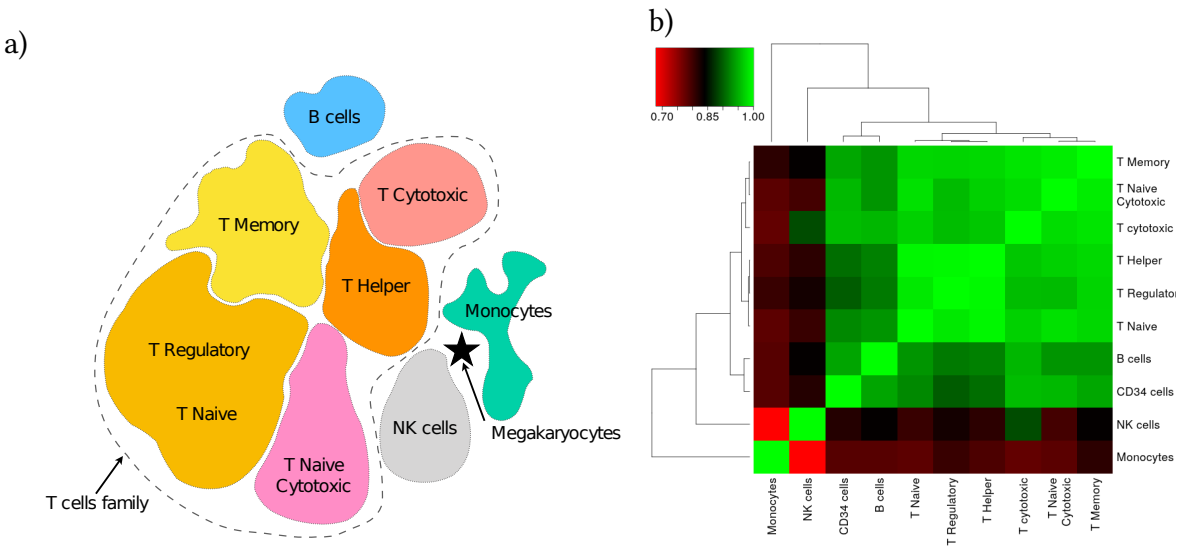


Figure 3.2: PBMC cells overview. (a) Schematic representation of t-SNE projection of 68k PBMCs dataset with cell subtypes clusters detected by correlation to type-specific transcriptomes adapted from (Zheng *et al.*, 2017). (b) The correlation heatmap of all PBMCs type-specific (averaged, normalized, log-scaled) transcriptomes.

correlation > 0.5. Second, the cell is assumed to be of a specific type if the difference between its correlation and correlations with other types is greater than a threshold value set to 0.025, otherwise it is assumed to be unknown.

Even though there are no transcriptomic profiles available for other cell types, such as Megakaryocytes (depicted in Fig. 3.2a) we are aware that they may exist in our dataset and thus expect to find them using our decomposition method. Please note that the above described correlation-based labelling in case of 68k PBMC dataset resulted in low percent of clearly assigned cell types, thus all results are presented for 2.7k cells. In the following section, we present the outcome of the analysis of the data using our truncated version of `RPCA` with Gaussian noise reduction.

3.5. Validation: Case-Study of Single Cell RNA-seq Data

The proposed `trPRCAL2` explains the input data (A) in terms of compressed, low-rank information (L), sparse signal (S) and noise (E). To validate our method on real data and evaluate its suitability for genomic data analysis, we use the scRNA-seq 2.7k PBMCs dataset. We report that `tRPCAL2` algorithm converged after 49 iterations, taking about 97 sec (compared to 20 sec PCA from `R prcomp`). Due to the high background variance, `tRPCA` and `RPCA` did not converge before 1000 iterations. All algorithms were run on AMD Opteron(tm) Processor 6380, 64 x 2.5GHz CPU, 256GB RAM, Gentoo Linux.

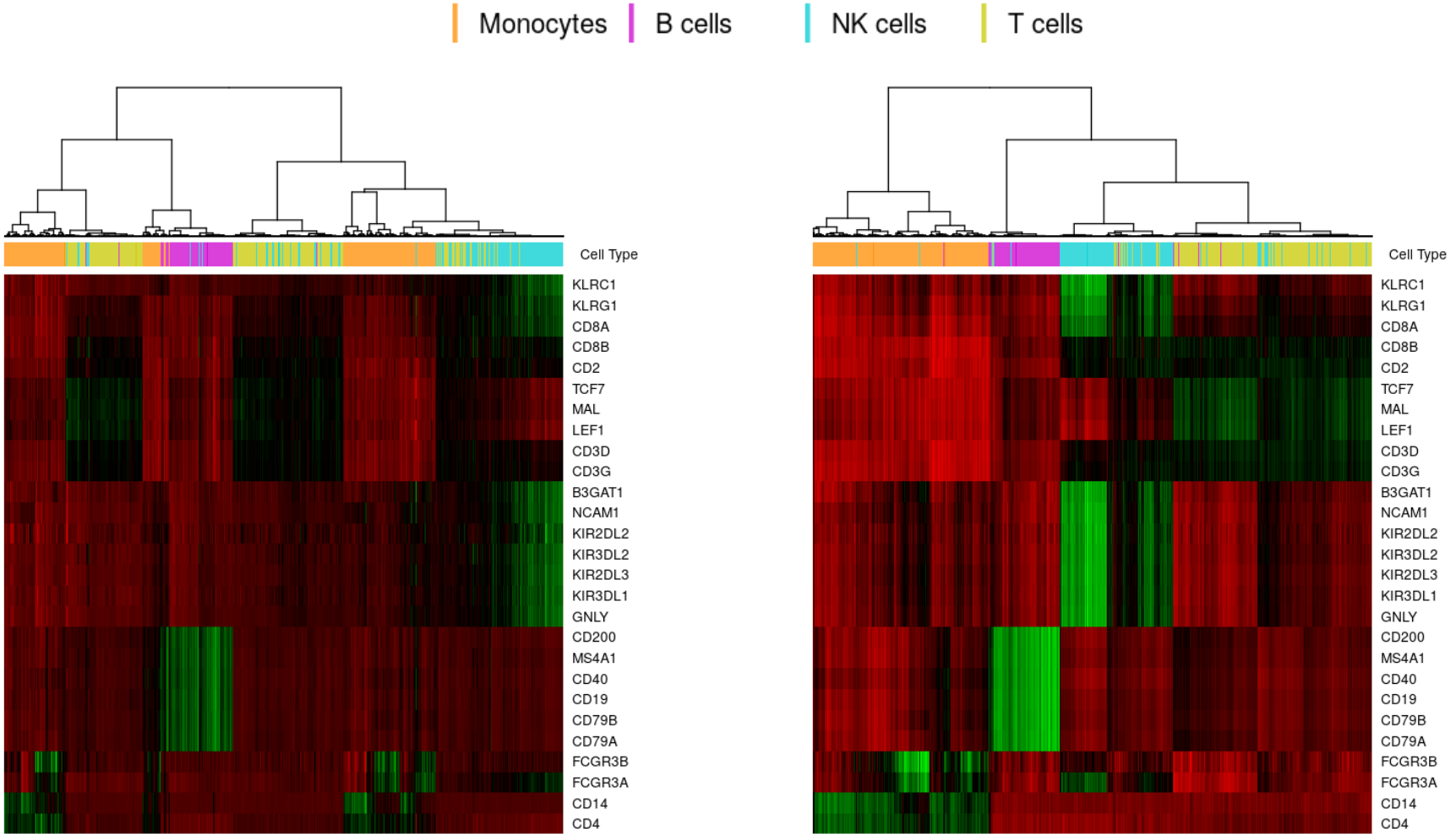


Figure 3.3: Marker gene based clustering comparison. The figure compares clustering of cells of known type with literature-based marker genes characterizing the analysed types of PBMC cells. The left panel is related to the signal represented in terms of the truncated PCA (10 highest singular values used). The right panel corresponds to the signal stored in the L matrix from `trPCAL2`. Top bars encode the original correlation-inferred cell types. Colours in the heatmap describe the activity level of a gene from lowest (red) through average (black) up to highest (green).

Clustering via Low-Rank Matrix.

First, we validate the quality of the dimension reduction by clustering cells basing on their low-rank representation in the L matrix. Using the hierarchical clustering algorithm (Johnson, 1967; Murtagh, 1985) we determined 5 clusters, which were visualized using t-SNE (van der Maaten, 2014) (see Fig. 3.4). In contrast to the expected cell types (derived from correlation with type-characteristic transcriptomes), we observed that the obtained clustering determines four main families of cells from the PBMCs dataset. Additionally, one more cluster separating NK and T cell family clusters was discovered. The cluster is described by increased activity of CD8A and CD8B (Bonferroni adjusted p-value $< 10^{-3}$) and regular activity of CD4, CD45 and CD25 genes in contrast to other cells. This characteristics suggests a cluster of cells mostly composed od $CD8^+$ T cytotoxic cells and explains its similarity to NK cells (Zheng *et al.*, 2017; Ohkawa *et al.*, 2001). Other dimension reduction, clustering and visualization techniques were also compared, e.g. PCA, Isomap (Tenenbaum *et al.*, 2000) or SIMLR (Wang *et al.*, 2017a), but since their quality was at most comparable we present results for the commonly used and already mentioned t-SNE algorithm.

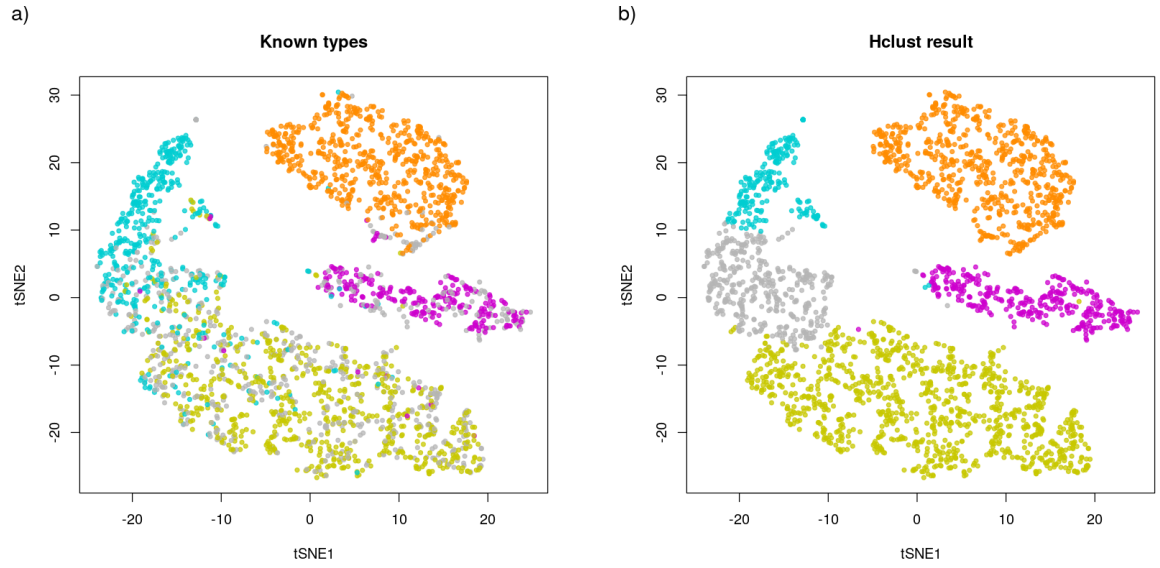


Figure 3.4: Clustering of 2.7k PBMCs. In both panels, cells are visualized using t-SNE (perplexity=35) ran on the 10-dimensional representation of the original input data (A) derived from L matrix. (a) Colours correspond to cell types inferred from correlation of each cell original transcriptome (columns of A) with type-specific PBMCs transcriptomes. We have determined: 630 Monocytes (orange), 251 B-cells (pink), 437 Natural Killer cells (blue) and 700 T-cells (yellow). Remaining 682 (gray) are assumed to be an unknown or tentative type. (b) Colours correspond to 5 clusters determined by hierarchical clustering method. Colours of the clusters correspond between predicted and original clusters for clarity.

Next, we compared our method of dimension reduction with the method analogous to the one used in (Zheng *et al.*, 2017). With SVD, we calculate top 10 singular values (in pursuance of the L matrix rank) of the PBMC data matrix (A) using R `irlba` package. Then, the input data was approximated through the reduced 10-dimensional space. We perform the hierarchical clustering of all cells for the most characteristic marker genes

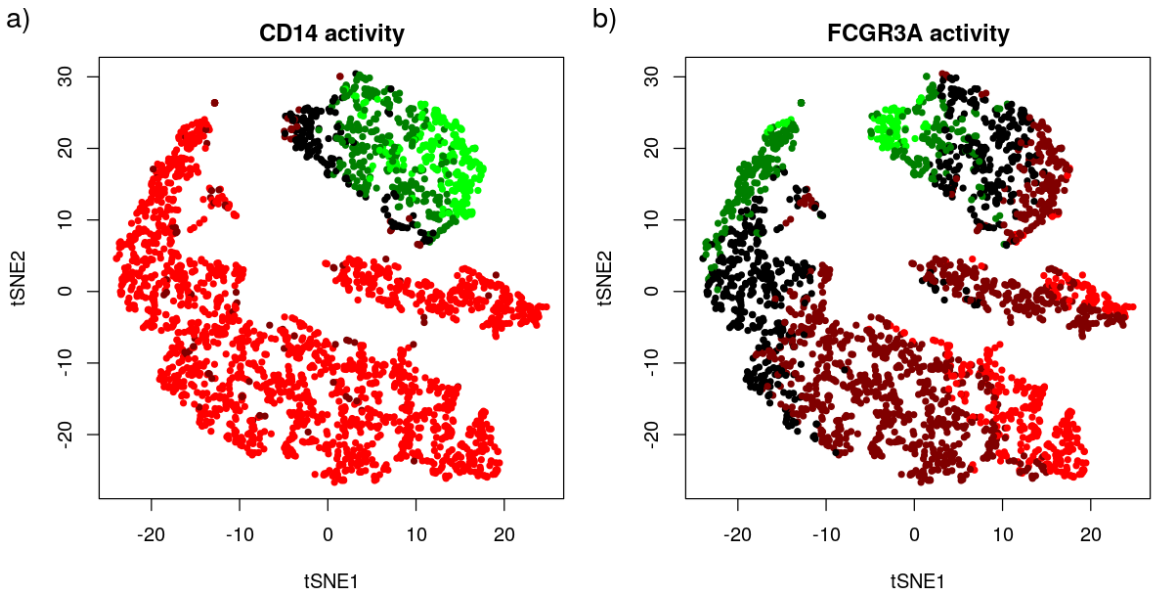


Figure 3.5: CD14 and FCGR3A activity levels. Panels present the activity of monocytes marker genes. (a) and (b) figures present the activity of CD14 and FCGR3A genes among all cells, respectively. The level of gene activity (lowest to highest) is spanned from red, through black, to green color scale.

per cell type (selected from the literature) on the described SVD-based approximation and the L matrices. The aim is to verify how well the dimensionality reduction preserves the most reliable, biological information related to type-specific marker genes. It appeared that not only the L matrix guarantees more accurate clustering, but also it contains more pronounced differences of the signal between clusters of both cells and genes, Fig. 3.3.

Monocyte Subtypes and Co-Expression Detection.

The literature suggests existence of at least three subtypes of monocytes in PBMCs (Ziegler-Heitbrock *et al.*, 2010). Their characterization can be based on the presence of CD14 (coded by CD14 gene) and CD16 (coded by FCGR3A, FCGR3B genes) clusters of differentiation: (i) the classical monocyte with high activity of CD14 ($CD14^{++}$ $FCGR3A^{-}$); (ii) the intermediate monocyte with high activity of CD14 and low activity of FCGR3A ($CD14^{++}$ $FCGR3A^{+}$); (iii) the non-classical monocyte with low activity of CD14 and co-expressed FCGR3A ($CD14^{+}$ $FCGR3A^{++}$).

Interestingly, such classification of subtypes can be found using the low rank signal from the L matrix (see Fig. 3.5). The activity of CD14 is almost uniquely distributed among the cluster of monocyte cells and, simultaneously, the activity of FCGR3A changes with the gradient defining the cell subtype progression among all monocytes. Moreover, Fig. 3.6 shows how the original expression values are distributed among decomposition matrices. The sparse peaks of activity are stored in S and the linear part in L. E matrix contained remaining noise of mean 0 and the standard deviation of order 10^{-4} for both CD14 and FCGR3A.

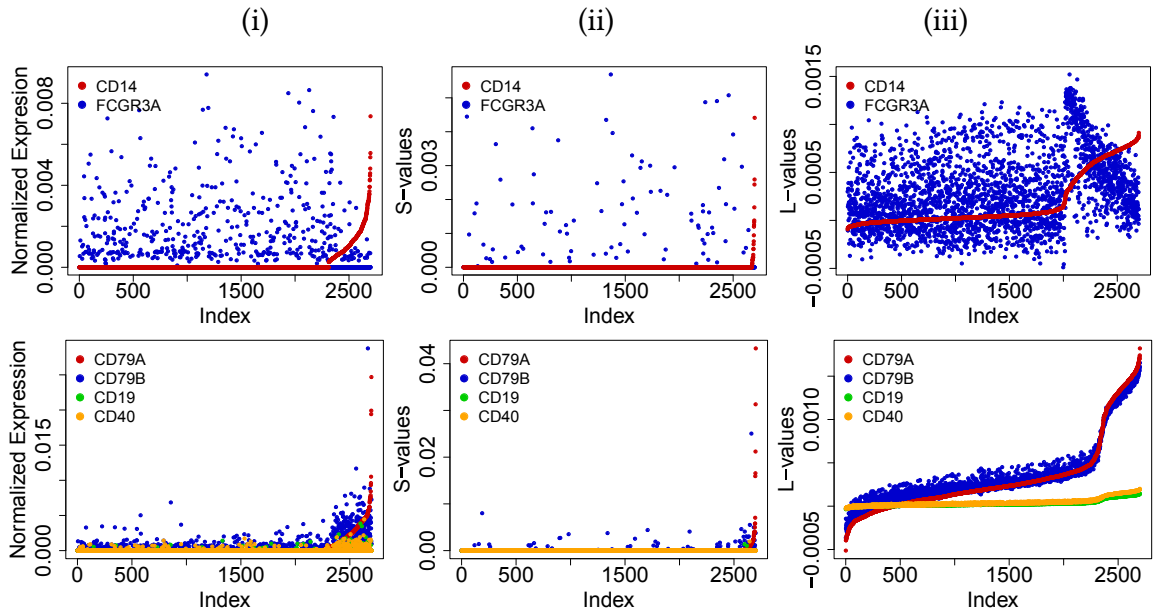


Figure 3.6: Co-expression patterns. The distribution of the original expression levels among S and L matrices for marker genes of monocytes (top) and B-cells (bottom). Consecutive panels present: (i) the normalized, log-transformed input data from A matrix; (ii) low-rank signal in L matrix; (iii) sparse signal in S matrix. In each panel cells (x-axis) are sorted by the activity level (y-axis) of first marker gene (CD14 for Monocytes and CD79A for B-cells).

Additionally, the low-rank L matrix well tracks and recovers co-expression patterns between genes. Namely, the activity of B cells can be detected by the presence of CD79 heterodimer composed of CD79A and CD79B proteins (Chu and Arber, 2001). Their co-expression measured in terms of correlation was at the level of 0.227, while after the decomposition their low-rank signal had correlation of level 0.995 (see Fig. 3.6). Similarly, the correlation between FCGR3A and GNLY characterizing Natural Killer cells (Crinier *et al.*, 2018) increased from 0.400 to 0.949. Naturally, these observations are possible thanks to filtering out the sparse and noise signals. Nonetheless, this type of information is retrieved by the proposed method in an unsupervised manner, and may suggest new co-expression patterns.

Sparse Signal Interpretation.

The presence of megakaryocytes in our PBMC dataset, reported in the population of PBMCs sample from (Zheng *et al.*, 2017), was not evident using the low-rank L matrix. Even though, a small cluster of cells of unknown type was separated by t-SNE (see Fig.3.4) and an analogous cluster depicted in Fig. 3.2a for 68k PBMCs data). Aiming in Megakaryocytes detection, we performed the hierarchical clustering on the subset of unknown type cells and only genes that had at least one non-zero entry in the sparse S matrix. This resulted in recovery of a well-separated cluster of 9 cells. Further analysis confirmed that the cluster is characterized by high over-expression of PF4 gene, which is a well known marker for

mature megakaryocytes (Adachi *et al.*, 1991), in comparison to other unknown cell types.

3.6. Noise Reduction and Algorithm Parameters.

Finally, we want to discuss the importance of the noise matrix E and selection of λ_1 and λ_2 parameters. The final decomposition quality, in terms of information distribution among three matrices, is mainly based on the choice of these crucial parameters.

For the purpose of this study, we have set $\lambda_1 = 0.016$ and $\lambda_2 = 10.0$, which resulted in the $L+S+E$ decomposition characterized by the norms of vectorized matrices summarized in the Table 3.1. Selection of the mentioned values was supported by the grid-based search

	$\|\|\cdot\|\|_*$	$\|\|\cdot\|\|_1$	$\|\|\cdot\|\|_2$
L	5.753	3398.162	4.265
S	60.289	60.289	2.826
E	57.881	2670.012	1.440

Table 3.1: Summary of all norm values calculated for each matrix resulting from the L+S+E decomposition with initial parameters set to $\lambda_1 = 0.016$ and $\lambda_2 = 10.0$.

through the parameter space. We have run `tRPCAL2` decomposition on the PBMC data for 150 different, evenly distributed, pairs $(\lambda_1, \lambda_2) \in [0.001, 0.05] \times [5, 15]$ setting $\mu_0 = 147.28$ using the improved formula for the initiation of the μ_0 parameter that takes into account the sparsity of an input data matrix A with k rows and l columns:

$$\mu_0 = \frac{|\{a_{i,j} : \ a_{i,j} \neq 0\}|}{4 \cdot k \cdot l \cdot \sum_{i,j} |a_{i,j}|}$$

The new formula is thus a ratio of the percent of non-zero values to 4 times the sum of absolute values in the data matrix. To determine the order of magnitude and search ranges for both parameters we have made use of the theory described in (Candès *et al.*, 2011) as well as estimations based on the properties of the AA^T matrix trace operator.

Since `tRPCAL2` algorithm mixes L_1 and L_2 norms, which describe different mathematical properties and in this sense are incomparable, the final decomposition depends not only on relative or absolute values of chosen λ parameters, but also on distributions of elements in the decomposed matrix. In order to approximate the relationship between λ_1 and λ_2 and their influence on the final composition of L, S and E matrices we summarized the results from simulation study and we conclude about properties such as: the rank of the resulting L matrix, relative and absolute sparsity of the S matrix, Fig. 3.7.

First, we systematize the boundary behaviours of the `tRPCAL2` algorithm. Namely, when both $\lambda_1 \to \infty$ and $\lambda_2 \to \infty$ the decomposition will result in $L = A, S = E = 0$. Next, for fixed λ_1 and $\lambda_2 \to 0$ the information shifts to E matrix and $E = A, L = S = 0$.

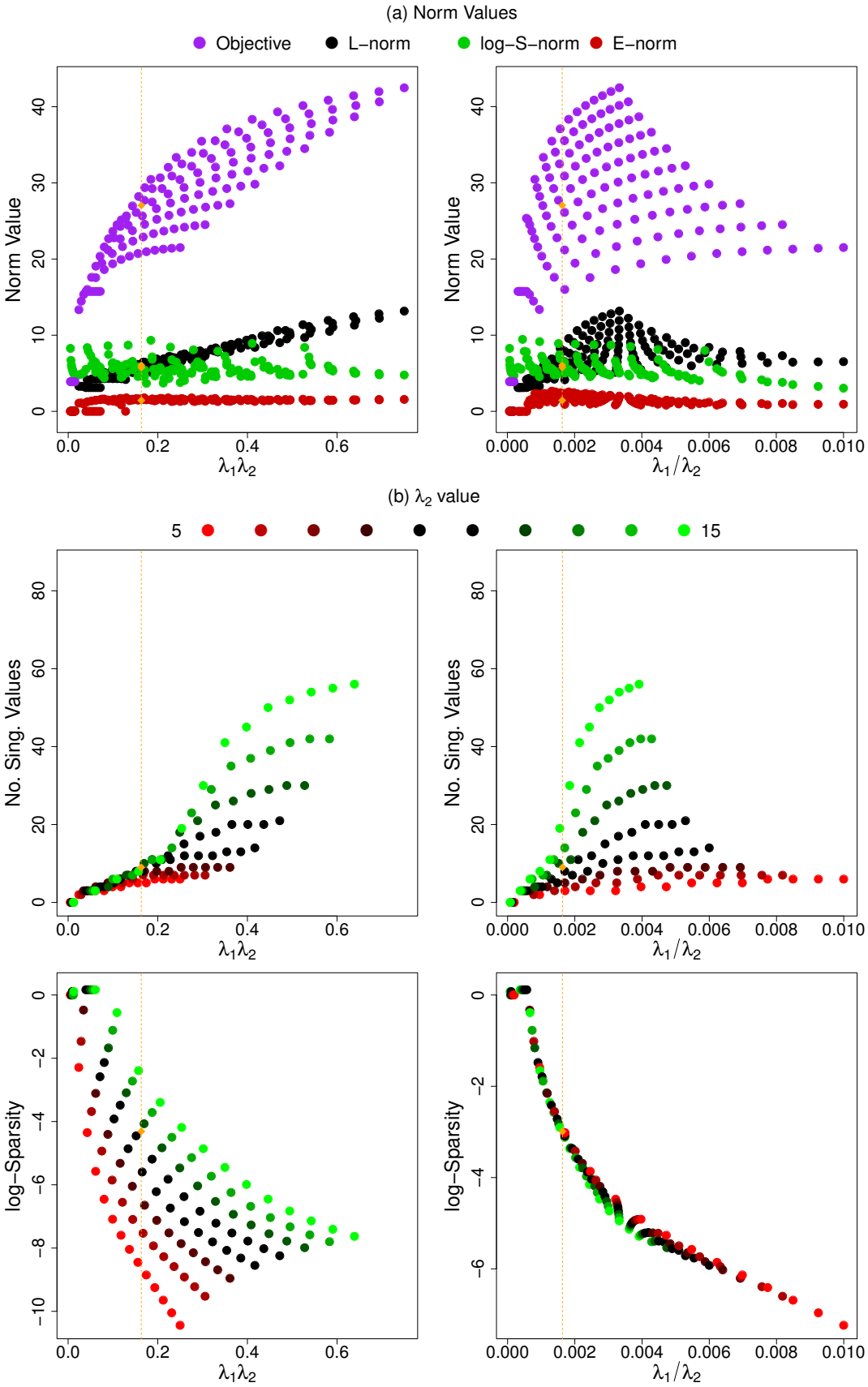


Figure 3.7: Properties of the algorithm. Each row presents value of some measurement as a function of product (left) and quotient (right) of λ_1 and λ_2 parameters, that `tRPCAL2` was run with. (a) Norm values of each matrix and the objective function value. (b) the number of singular values of L matrix (top) and logarithm of sparsity (percent of non-zero matrix entries) of S matrix (bottom). On each plot orange line corresponds to λ parameters finally used in our study.

Similarly, fixed λ_2 and $\lambda_1 \to 0$ results in $S = A, L = E = 0$ (see Fig. 3.7a). Although intuitive, these observations depend on different convergence rate, and thus flow of the information among resulting matrices. Here, based on the described set of simulations (see Fig 3.7b), we indicate several observations regarding signal distribution: (i) the rank of matrix L increases sub-linearly as a function of λ_1 and fixed λ_2; (ii) polynomially as a function of λ_2 and fixed λ_1; (iii) for fixed λ_2 sparsity (percent of non-zero elements) of S decreases exponentially as a function of λ_1.

Finally, we observed that the distribution of the E matrix elements is a mixture of zero-centered Gaussian and low-variance Gaussian distributions concentrated around non-negative, λ_2-dependent value. We reason, that since `tRPCAL2` is an optimization of linear combination of norms, E matrix captures parts of low-rank and sparse components bringing such mixture. One way to overcome this effect was described in (Zhang *et al.*, 2016) through their truncated nuclear norm minimization for `RPCA`.

More formal investigation of the `tRPCAL2` theoretical properties with respect to λ_1, λ_2 and L, S, E matrices could be of high interest in terms of future research.

3.7. Conclusions

In this chapter, we introduce an extension of robust PCA. We propose to decompose the input matrix into low-rank L, sparse S and noise E components. Thanks to the reduction of noise using the L_2 penalty, we restore the inner structure of the matrix. Our results suggest that our algorithms may better approximate the underlying systematic variation in the input data, as well as recognize the sparse perturbation signal of the data. We present the case study based on the scRNA-seq data from 2.7k PBMCs. The method provides relatively fast and accurate dimension reduction and clustering of the high-dimensional data detecting different subtypes within a given cell type, co-expression patterns and novel subtypes.

One possible direction for the further research is to derive precise formulas for λ_1 and λ_2 parameters that guarantee optimal solutions of the decomposition problem. So far, simulation-based selection of the parameters is time consuming. Ideally, selection method of λ parameters should result in the most natural $L + S + E$ decomposition, taking into account user's expectations in terms of, for example, Bayesian priors to relative magnitudes, and to other components' statistics. The applicability of our method to other types of data we also see as a promising direction of a further research. Preliminary results of video and image analysis, not described in this paper, suggest that the method can be successfully harnessed in the field of video-surveillance and image analysis. The current implementation of the `tRPCAL2` algorithm is available on-line: `https://github.com/macieksk/rpca` as a development R package.

4

Integrative Analysis of Metabolic Landscape Matrix

"Essentially, all models are wrong, but some are useful."

— George E. P. Box

The huge amount, of biological data gained from many different sources opens up new possibilities alongside with some dangers in data integration. The latter are not so widely taken into account. In this chapter, we are going to see what problems may arise with integration of metabolomic knowledge with gene expression datasets. We analyse common pitfalls and provide solutions, exemplifying them by a case study of the renal cell carcinoma (RCC).

In particular we provide a metabolic description of known morphological RCC subtypes and suggest a possible existence of poor-prognosis cluster of patients, which are characterized by common low activity of drug transporting enzymes important in chemotherapy.

Finally, the goal of this work is also to point out the problem that arises from integration of high-throughput data with inherently biased manually curated low-throughput data. In such cases overrepresented information may potentially overshadow the non-trivial discoveries.

4.1. Introduction

As we have already mentioned in the previous chapters, it can be easily observed, that in the past two decades the technological progress has provided a huge amount of biomedical data from various molecular levels (the so-called *-omics* data). Consequently, our understanding of biological processes gets more profound. These facts open up a broad field of data integration, that aims to infer from various data taking into account known biological dependencies between them.

The modern studies and corresponding literature emphasize the significant role of integration of various *-omics* data for better comprehension of a phenomenon of interest (Huang *et al.*, 2017). However, these integration procedures, thanks to a wider perspective, are not only meant to provide a new insight into a specific phenomena, but above all should constitute a deeper understanding of the genotype-phenotype mapping, that was recalled in the introduction of this thesis. The principal aim of all integrative studies is thus a better description of the relationship between genotype and phenotype layers (Gjuvsland *et al.*, 2013). Importantly, each unique set of *-omics* data can be integrated using various statistical methods, and thus may result in an unprecedented outcome. As a consequence, each year a number of reviews are published to track and summarize the current state of the art in the field of *-omics* integration, which one can check for more detailed discussions (Wanichthanarak *et al.*, 2015; Fondi and Lio, 2015).

Even though, as presented in the literature, there exists a plethora of methods designed for integration of transcriptomic and metabolomic data and their usage has provided interesting biological outcomes, in our work we want to focus on a particular problem of transcriptomic and metabolomic data integration. We report an interesting phenomenon related to the analysis of individual metabolic networks supported by transcriptomic data. To the best of our knowledge, yet, no one has commented on the bias that the structure of a metabolic network can introduce when used along with integrative methods, in particular when used in the context of the flux balance analysis (FBA) (Orth *et al.*, 2010).

Integration Strategies for Transcriptomic and Metabolomic Data.

One of the first attempts to this type of integration was suggested by (Covert and Palsson, 2002), where authors infer the binary enzymatic activity from the transcriptomic continuous signal. Next, a cascade of methods was proposed to harness transcriptomic data for analysis of metabolic networks. Among the most discussed methods we can point out: E-Flux (Colijn *et al.*, 2009), GIMME, (Becker and Palsson, 2008) GIMMEp (Bordbar *et al.*, 2012), iMAT (Zur *et al.*, 2010), INIT (Agren *et al.*, 2012), MADE (Jensen and Papin, 2011), mCADRE (Wang *et al.*, 2012), PROM (Chandrasekaran and Price, 2010), RELATCH (Kim and Reed, 2012). Apart from some minor details, fundamentally these approaches differ at

three basic levels:

- inferring enzymatic activity from transcriptomic data;
- inferring flow capacity boundaries from transcriptomic data;
- number of samples needed to create a mathematical model.

These methods along with their other features, were broadly discussed and compared in few reviews where detailed descriptions can be found (Mardinoglu and Nielsen, 2012; Masoudi-Nejad and Asgari, 2015; Kim and Lun, 2014; Blazier and Papin, 2012). Therefore, here, we focus on a summary of what type of biomedical outcomes they have provided so far.

Using this type of models, where a general metabolic network becomes context-specific through integration with a transcriptome of specific tissue or organism few groups of researchers have already reported some interesting discoveries. With nearly 2000 samples of breast tumor researchers discovered a novel poor prognosis cluster characterized by local production of serotonin along with active extracellular matrix and membrane remodelling reactions (Leoncikas *et al.*, 2016). In (Li *et al.*, 2010), by integrating transcriptomic knowledge with human metabolic network authors suggest a supervised method to predict novel drug-target interaction. In their work they predict related metabolic reactions and enzyme targets for approved cancer drugs, and predict drug targets with statistically high confidence rate. Reconstruction of a genome-scale metabolic models for 126 human tissues and cell types including healthy and tumor type was derived using the Recon 1 human metabolic network along with transcriptomic data (Wang *et al.*, 2012). Among all, the set of models includes 26 tumour-specific models accompanied by their normal counterparts, in particular 30 models of brain tissue subtypes were determined. Finally, using the modified version of iMAT, differential fatty acid uptake into mitochondria along with arachidonic acid and eicosanoid metabolism were suggested to explain different proliferative rates and invasiveness between PC-3/M (highly proliferative, cancer stem cells) and PC-3/S (highly invasive, epithelial-mesenchymal-transition-like properties) subpopulations derived from prostate cancer cell line (Marin de Mas *et al.*, 2018). In summary, the general biomedical aim of this approach is to model the gene-protein-reaction interactions in a form of a metabolic network and infer multiple biological properties, e.g. post-transcriptional gene activities or intensity and activity of metabolic reactions, that can explain the nature of analyzed sample.

In this chapter, we want to draw attention to a specific bias related to the task of data integration and analysis using general metabolic networks along with transcriptomic data. First, we present how the structure of a metabolic network can easily increase the importance of specific groups of enzymatic reactions, which may lead to incomplete or questionable conclusions. Next, in order to cope with that obstacle, we suggest a possible routine

that can eliminate the unwanted bias preserving the metabolic knowledge obtained from integration procedure. Finally, we present results from the analysis of the TCGA kidney cancer dataset, that point out reactions discriminating patients in the context of observed clinical factors. Additionally, we report discovery of a poor-prognosis cluster of patients along with its characterization.

4.2. Human Metabolism Modelling

In this section of the chapter, we describe in details the topological and structural properties of the most recent version of the model of human cell metabolic network (RECON 2.2). Next, we will describe how to personalise a general metabolic network using an information about transcriptomic activity.

RECON 2.2: The Human Genome-Scale Metabolic Reconstruction

In general, a metabolic network is a set of reactions among elements of a given set of metabolites. Each reaction may be associated with a specific genetic rule that needs to be met in order for reaction to occur. If a reaction has no genetic rule assigned it can be triggered whenever all substrates are available. A genetic rule usually describes which genes/proteins need to be active/present in the system to carry out the reaction. In particular, a rule may require simultaneous presence of several enzymes (e.g. the transfer of L-Oh-Proline by the Apical Imino Amino Acid Transporters in Kidney And Intestine requires both Transmembrane Protein 27 and Solute Carrier Family 6) or alternative enzymes that can catalyze the reaction. (e.g efflux of 2-hydroxy-atorvastatin-lactone into bile that is supported by ATP binding cassette subfamily C member 2 or subfamily B member 1). Finally, each reaction has also a lower and upper bound describing minimal and maximal flow of metabolites through this reaction.

On the other hand, formally, the network can be considered as a Petri net with conditional (gene rules) transitions (reactions) of places (metabolites), where by a conditional transition we mean additional boolean formula that needs to be met in order for transition to occur (see Figure 4.2 for the example).

In our study we use the human genome-scale metabolic reconstruction model, RECON 2.2 (Swainston *et al.*, 2016). The model describes 7785 reactions (transitions), out of which 4742 depend on a genetic rule (condition) and these are called enzymatic reactions. Next, each genetic rule is a boolean formula in a disjunctive normal form (DNF), where each boolean variable corresponds to binary activity state (active or not) of one of 1670 unique genes. Additionally, there are 2654 metabolites distributed among 10 compartments. In total, we consider 6048 compartment-specific metabolites (see Table 4.1 and Figure 4.1).

all reactions	7785	enzymatic	4742
		transport	3043
		exchange[a]	701
		demand[b]	44
		reverse	3782
		directed	4003
all metabolites	6048	unique substances	2654
		boundary	722
		compartments	10
all genetic rules	4742	simple rules	2912
		complex rules	1830
		unique rules	1341
		unique genes	1670

Table 4.1: Summary of the RECON 2.2 metabolic model. [a] exchange reactions describe in- and outflow of metabolites through the system boundary; [b] demand reactions are intra-network, unlimited sinks or sources of metabolites degradation or production.

Personalized Genome-Scale Metabolic Models

In order to construct a sample-specific metabolic model from a general metabolic network for a given transcriptomic pattern the following procedure is applied. First, specify thermodynamic conditions describing acceptable capacity of reactions fluxes, i.e. set minimal and maximal flow levels through each reaction. Next, evaluate each genetic rule, based on the transcriptomic data to decide which reactions can occur in the initial model (see an example in Figure 4.2). Finally, the obtained model can be further studied. In our work, we formulate a linear programming problem to find a steady-state flux distribution. This procedure provides two types of information. First, it suggests new pattern of fulfilment among genetic rules (binary valued), where each change, in contrast to the original value, can be considered as a post-transcriptional change. Second, it outputs a vector of fluxes that met the criteria of the linear problem that was solved. These fluxes can provide an information about metabolic reactions or pathways that are mostly exploited given such expression pattern.

TCGA Transcriptomic Data

In this study we have used 20 RNA-seq transcriptomic datasets provided by TCGA, each composed of cancerous and control samples and accompanied by clinical data, that provide a palette of categorical and numerical values describing all samples. Apart from tumorous samples there are also control (derived from the healthy tissue) samples, which allow to track potential, cancer-specific markers. For the purpose of our work, each dataset was

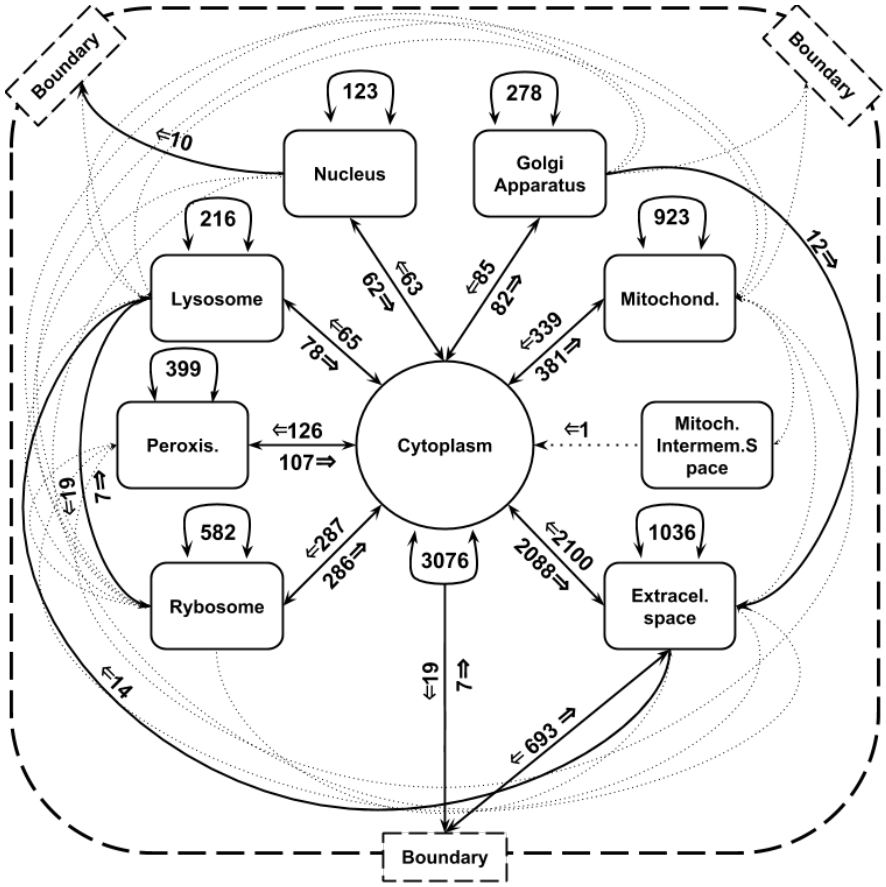


Figure 4.1: The figure presents an outline of the reactions distribution in the metabolic model of RECON 2.2. There are 9 intercellular compartments depicted and the additional boundary component which represents external entry and exit of metabolites into the metabolism network model. Each line connecting two compartments represents all reactions that involve metabolites from these compartments. The number of these reactions and their direction is assigned to each line. To keep the figure readable, for less than 10 reactions we use a dotted, thinner line.

subjected to a standard preprocessing routine with `recount` R package. Next, only genes that are composing genetic rules in the RECON 2.2 model were selected. Finally, we select a number of 5 read counts to be a threshold for a gene to be considered as active.

In the presented case study we used the kidney cancer dataset (Renal Cell Carcinoma, RCC; 897 tumor, 129 normal samples) Tumor tissues were also classified by their morphological type subtype into four groups: 527 Clear Cell RCC, 290 Papillary RCC, 66 Chromophobe RCC and 14 unclassified RCC samples. For 201 patients the survival data were also available. Additionally, we used the brain cancer dataset (707 samples) a showcase of a biased clustering results. The above procedure resulted in 20 binary activity matrices that were used to create sample-specific metabolic network models. Importantly, each cancer type is supported by clinical data, that provide a palette of categorical and numerical values including: cancer subtype, morphology, survival time, etc.

Gene	Activity	Reactions	
G_1:	0	$R_1 : M_1 + M_2 \xrightarrow{G_1 \vee G_2} M_4$	$R_4 : M_1 \xrightarrow{G_3} M_3 + M_4$
G_2:	1	$R_2 : M_3 \xrightarrow{G_3} M_2$	$R_5 : M_4 \xrightarrow{\top} M_1$
G_3:	1	$R_3 : M_1 \xrightarrow{G_1 \wedge G_3} M_2$	$R_6 : M_4 \xrightarrow{G_1 \vee G_2 \wedge G_3} M_3$

Individual Model

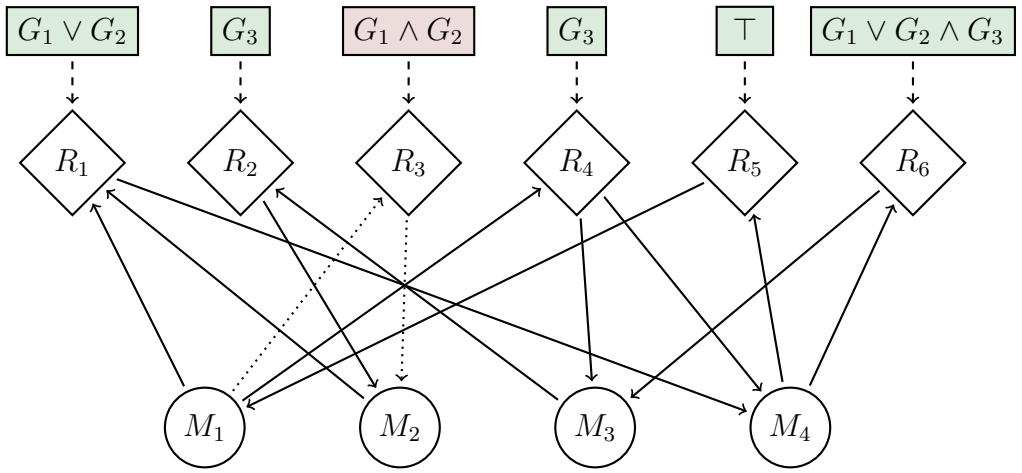


Figure 4.2: An example of metabolic network with genetic rules. The example is composed of five reactions R_i among four metabolites M_j. Among these reactions there are four enzymatic that are coordinated by enzymes related to three specific genes G_k. All arrows describe the flow of metabolites through reactions according to the above list of reactions. Using the gene activity pattern, a general network can be turned into transcriptom specific and represented as a Petri net with conditional transactions. Here we can see that inactivity of gene G_1 results in silencing (dotted line) reaction R_3, which represses the production of metabolite M_2 in the system.

Steady-State Fluxes Distribution

In order to determine the activity state of each reaction in a personalized model we make use of the approach proposed by (Shlomi *et al.*, 2008).

$$\max_{v,y^+,y^-} \sum_{r\in\mathcal{A}} (y_r^+ + y_r^-) + \sum_{r\in\mathcal{I}} y_r^+ \quad \text{s.t.} \qquad \text{(a)}$$

$$\begin{aligned}
& S \cdot v = 0 && && \text{(b)} \\
& v_{min} \le v \le v_{max} && && \text{(c)} \\
& v_{min,r}(1 - y_r^+) \le v_r - y_r^+ && r \in \mathcal{A} && \text{(d)} \\
& v_{max,r}(1 - y_r^-) \ge v_r + y_r^- && r \in \mathcal{A} && \text{(e)} \\
& v_{min,r}(1 - y_r^+) \le v_r && r \in \mathcal{I} && \text{(f)} \\
& v_{max,r}(1 - y_r^+) \ge v_r && r \in \mathcal{I} && \text{(g)} \\
& v \in \mathbb{R}^m, \;\; y_r^+, y_r^- \in \{0,1\} &&
\end{aligned} \tag{4.1}$$

First we process the RNA-seq data and use it to determine the state of genetic rules of the metabolic network, as described in the previous subsection. Specifically, each gene is assumed to be active (1) or in-active (0) and used to evaluate the boolean genetic rule as a

standard logical expression. As a result, each enzymatic reaction r belongs to one of two disjoint sets: active $\mathcal{A}$ or inactive $\mathcal{I}$ reactions.

Next, we formulate a mixed-integer linear programming (MILP) problem (see Equation 4.1) and solve it using Gurobi solver (Gurobi Optimization, 2018). The goal is to maximize an objective function (4.1a) with respect to stoichiometric (4.1b), thermodynamic (4.1c) and transcriptomic (4.1d-g) constraints. The solution of the described problem provides a sample-specific, binary vector of activity states for all reactions (both enzymatic and transports) which we refer to as a metabolic landscape. For detailed description see the Methods section in (Shlomi *et al.*, 2008).

4.3. Binary Data Analysis

In the literature there are described various statistical methods that are specifically dedicated to the analysis of binary data, such as non-negative matrix factorization (Zhang *et al.*, 2007), sparse logistic PCA method proposed by (Lee *et al.*, 2010) or variational method for the factorization of 0–1 data, employing independent beta latent densities (Li, 2005).

Nonetheless, in our case, in order to explore the data, track the potential latent variables and visualize results we use the standard principal component analysis (PCA), which provides numerically stable and robust results and is also used to define a function for separating variables selection.

PCA Loadings-based Variables Selection

Let $M \in \{0,1\}^{m\times n}$ be a binary data matrix with m observations and n variables, $L = (L^{(1)}, ..., L^{(n)})$ be a loadings matrix and P scores (or principal components) matrix, satisfying $P = ML$. By definition P is a representation of M in a new basis that is composed of vectors of L. For the purpose of binary landscapes analysis, we construct a set of highest valued coordinates from first k directions of the rotation matrix. The general parametrized function $\mathcal{L}$ describes the selection procedure:

$$\mathcal{L}_{f,k}(L) = \bigcup_{i=1}^{k} \left\{ j \ : |L_k^{(i)}| \geq f\left(L^{(i)}\right) \right\}$$

where f is the function determining the threshold for the value of a coordinate to be assumed significant and k is the number of vectors from loading matrix that are considered. The function is used to find the groups of redundant groups reactions activities.

We have also used two measures to determine differentiating variables in our datasets.

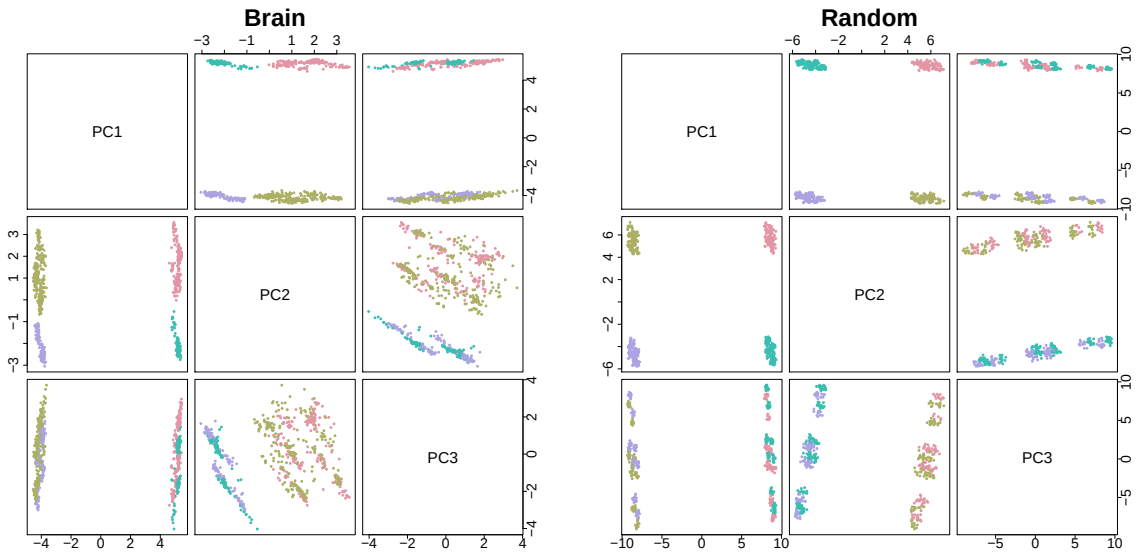


Figure 4.3: Comparison of the first three principal components of metabolic landscapes determined for brain cancer (left) and random (right) datasets. In both cases samples form well-separating clusters that can be identified by the activity pattern of overrepresented gene rules. For the brain dataset: SLC7A9 and SLC28A3. For the random dataset: SLC7A6 and SLCO1B1.

Jaccard Index

Let $z \in \{0,1\}^n$, we define $z^1 = \sum_i z_i$ and $z^0 = \sum_i (1 - z_i)$. The Jaccard Index (JI) of two binary vectors $\mathcal{J} : \{0,1\}^n \times \{0,1\}^m \to [0,1]$ is defined as:

$$\mathcal{J}(x,y) = \frac{\min(x^1, y^1) + \min(x^0, y^0)}{\max(x^1, y^1) + \max(x^0, y^0)}$$

The index evaluates to 0 when two vectors don't share any element, increases with an increasing number of shared elements between vectors and reaches 1 when vectors have both the same elements and the same lengths. JI is used to determine differentiating reactions between two groups of samples.

Tanimoto Similarity

Tanimoto similarity measure (TSM) of two binary vectors $\mathcal{T} : \{0,1\}^n \times \{0,1\}^n \to [0,1]$ is defined as:

$$\mathcal{T}(x,y) = \frac{n - |\{i : x_i \neq y_i\}|}{n + |\{i : x_i \neq y_i\}|}$$

The measure evaluates to 0 when two vectors differ at each coordinate, increases with an increasing number of compatible coordinates, reaching 1 when vectors are equal. TSM measures the distance between metabolic landscapes.

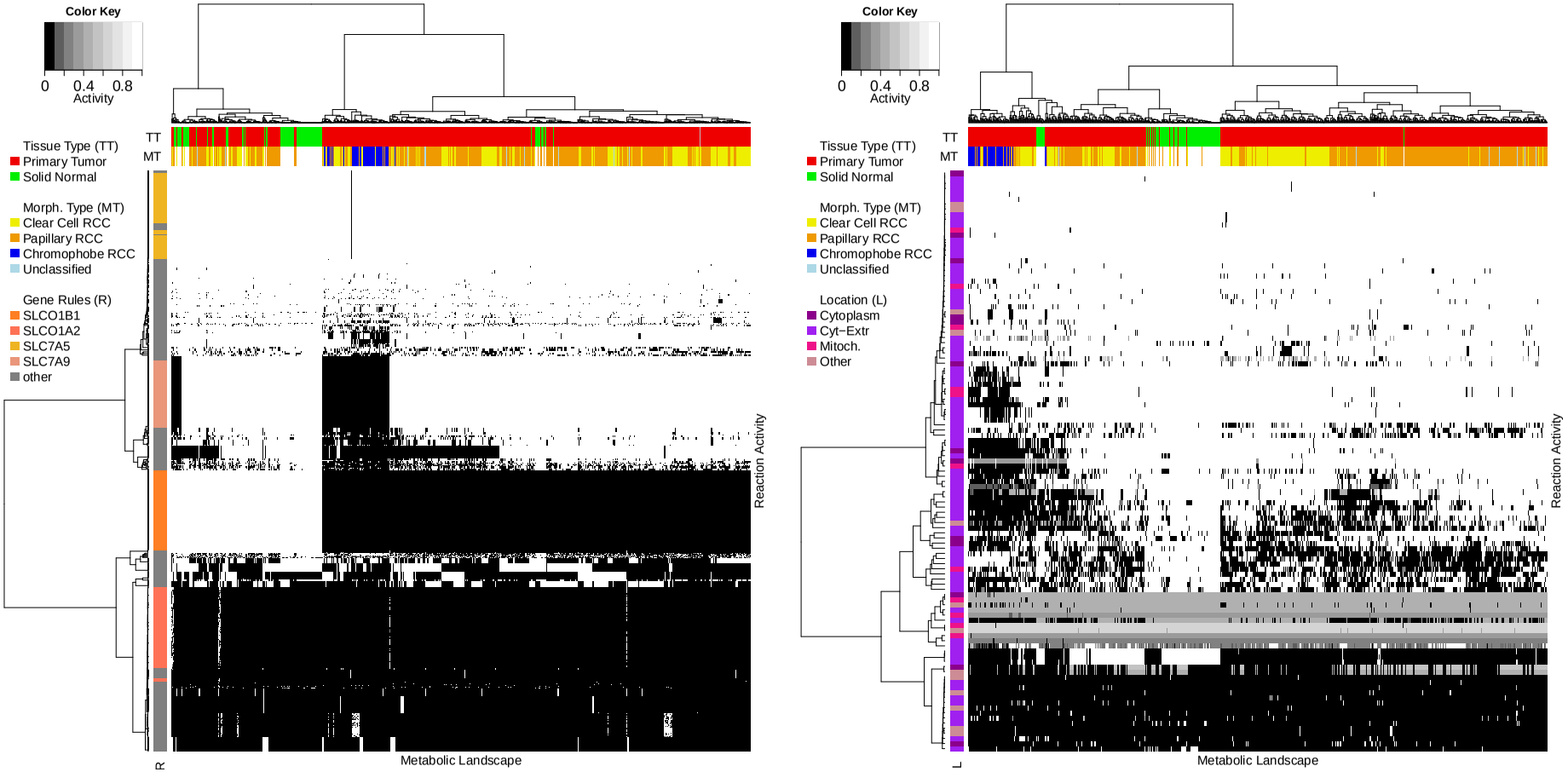


Figure 4.4: The comparison of reactions activity before (left) and after (right) bias reduction. On the left panel, the vertical strip marks reactions associated with the same genetic rule (orange scale colors) noticeably determining the clustering of all landscapes. In both panels, horizontal stripes represent the Tissue Type (TT) and Morphological Type (MT) of all samples. One can see, how the bias reduction improves the correlation of data with clinical variables, especially the morphological type. Finally, the vertical strip on the right panel presents that no bias related to compartments (purple scale colors) was introduced.

4.4. Integrative Analysis of Cancer Data

Our first step was to apply the procedure suggested by (Shlomi *et al.*, 2008) to the cohort of TCGA data in order to verify if there exists a common metabolic pattern among various types of tumors, or alternatively, if there exist specific metabolic biomarkers that may be useful during diagnosis procedures.

Universal Metabolic Biomarkers for Various Cancer Types

Our preliminary results suggested that, surprisingly, each data set has from 2 up to 4 clearly separable clusters of patients. In 13 out of 20 cancer landscapes datasets samples we identify on average 180 differentiating reactions separating samples into four clusters. Additionally, in remaining 7 landscape datasets there are on average 95 reactions differentiating samples into two clusters (see Figure 4.3 for an example based on brain cancer dataset). Moreover, it turned out, that almost all (>95%) discriminating (using $\mathcal{L}_{\max,2}$ function) reactions these clusters are coordinated by two main enzymes that are encoded by: SLCO1A2 (solute carrier organic anion transporter family member 1A2, coordinating superfamily of 94 Amino Acid-Polyamine-Organocation superfamily reactions (Jack *et al.*, 2000)) and SLC7A9 (solute carrier family 7 member 9, coordinating superfamily of 79 Resistance-Nodulation-Cell Division reactions (Tseng *et al.*, 1999)) genes. These observations may lead to a conclusion that cancers in general have natural subfamilies that can be described by the activity of specific groups of metabolic reactions and thus also activity of particular enzyme encoding genes. Activation of SLCO1A2 is related to the development and functioning of the immune system, organismal system for calibrated responses to potential internal or invasive threats and was highly over expressed in breast cancer tissues (Zhou *et al.*, 2015). On the other hand, SLC7A9 enables the transportation of substances (such as macromolecules, small molecules, ions) into, out of or within compartments of a cell, or between cells. Additionally, as a co-enzyme with SLC3A1 coordinates group of five transportation/exchange reactions of L-Cystine, L-Alanine and L-Ornithine, that were also detected as differentiating the cancer data. This observation implies activation of SLC3A1, that was recently reported to promote breast cancer tumorigenesis (Jiang *et al.*, 2017).

4.5. Metabolic Network Structure Bias

Even though the above literature reports may sound promising, we report another observation related to the analysis performed on a dataset with 500 randomly generated gene activities that were subjected to the metabolic analysis (see the right panel of Figure 4.3). Even though the gene activity dataset does not include any relevant information we are able

to identify two groups of reactions differentiating samples into four well-separable clusters. This observations, undoubtedly, put in question all results related to clustering of metabolic landscapes performed on all cancer datasets, since the analysis infers about knowledge that does not come from the data but from the topology and structure of metabolic network.

The analysis of the RECON 2.2 network structure revealed that there exist groups of reactions that are associated with the same genetic rule (top 10 most common genetic rules are related to over 900 reactions). Even though these reactions are biologically non-redundant (each of them is functional), they cause statistical redundancy which introduces network bias and may lead to biologically irrelevant conclusions. For this reason we propose two methods of bias reduction. The aim is to transform the data so that the statistical analysis does not detect artificially induced data separation, but rather may result in a discovery of subtle differences in metabolic activity between samples possibly related to novel metabolic biomarkers.

Due to the discovery of artifacts in a metabolic landscape analysis we propose the following pipeline of procedures, that aims to reduce the statistical redundancy resulting from the RECON 2.2 network structure.

Computational Procedure for Bias Reduction

First, we process the RNA-seq transcriptomic data, convert it to gene activity matrix using the threshold of 5 counts. The matrix is used to create a personalized metabolic network model, formulate a MILP problem and solve it with Gurobi solver. The solution is composed of sample-specific metabolic landscapes, which are subjected to statistical analysis. The analysis of data is initiated by verification if redundancy among groups of reactions exists (using the $\mathcal{L}$ function) and its removal, if needed.

The data transformation is based on aggregating the activity state for groups of reactions with respect to compartments to which belong their substrates and products. Namely, each reaction is labeled with a name of form $s_1 ... s_k$-$p_1 ... p_j$, where $s_1, ... s_k$ and $p_1, ..., p_j$ are alphabetically ordered names of compartments that, respectively, substrates and products belong to. Next, all reactions with the same genetic rule and assigned label are aggregated into one represent reaction with the level of activity equal to the average of activities in the group.

Finally, inference from the transformed data structure through hierarchical clustering using binary distance measures (e.g. TSM), correlation with clinical data, selection of discriminatory features and functional analysis of determined clusters is performed. The source code of scripts and functions used in the described workflow are available on Github `https://github.com/storaged/metabolic-landscape`

4.6. Validation: Bias Reduction for Renal Cell Carcinoma

In order to validate the proposed workflow we perform the metabolic landscape analysis on the TCGA dataset of Renal Cell Carcinoma. After performing the bias reduction we report significant improvement in samples clustering, both in the sense of unwanted network structure-dependent clusters composition and correlation with clinical data. (see Figure 4.4). We remove the amplified activity pattern of reactions coordinated by the same genetic rule, that influenced the clustering of samples in an unwanted way. After the reduction there are no significant, discriminating reactions associated with the same genetic rule. This step also results in more reliable clustering of data according to clinical observations, e.g. normal or tumor tissue type or morphological type of a tumor sample.

Biomarkers of Renal Cell Carcinoma

The results of our analysis of the Renal Cell Carcinoma (RCC) TCGA dataset are consistent with latest reports, that indicate SLC6A3 as a experimentally confirmed biomarker for RCC (Hansson *et al.*, 2017). The transcriptomic signal of SLC6A3 in our data is clearly discriminating biomarker of Clear Cell RCC subtype (6.89 logFC)

However, thanks to the analysis of metabolic landscapes we suggest further potential biomarkers that correspond to specific, known RCC subtypes (Muglia and Prando, 2015). The literature so far reports CXCL16 gene as a significantly expressed in papillary RCC with others still waiting for their validation (McGuire and Fitzpatrick, 2009). Nonetheless, the analysis of metabolic landscapes suggests two transport reactions that discriminate the Papillary RCC subtype. Both of them are supported alternatively by the already reported SLC6A3 or SLC6A2, other member of the same Solute Carrier family. The transportation activity state of dopamine and norepinephrine via sodium symport between cytoplasm and extracellular space are well separating the Papillary RCC subtype from other samples. Namely, these reactions are predicted to be inactive in Papillary samples. Our results report 247 out of 291 ($\approx 85\%$) papillary samples and 42 out of 737 ($\approx 6\%$) other samples characterized by inactivation of these reactions. This observation, is also confirmed by pure transcriptomic data, which suggests down regulation of both genes (-3.62 and -2.2 logFC of SLC6A2 and SLC6A3, respectively) for papillary samples. This observation suggests simultaneous drop of activity of both SLC6A2 and SLC6A3 as a potential diagnostic biomarker.

In case of Chromophobe RCC (ChRCC) subtype, before the bias reduction two genetic rules were in fact separating this subtype from other cancer samples: (i) extracellular space and cytoplasm exchange reactions supported by the complex of SLC3A1 and SLC7A9 (5 reactions involving L- Cystine, L-Alanine and L-Leucine) and (ii) reactions controlled by SLC7A9 that is involved in 79 reactions. However, after the bias reduction, we find two

additional biomarker candidates: inactivation of sodium-dependent transport of (i) phosphate, supported by SLC17A1 and (ii) ascorbate supported by SLC23A1. Inactive phosphate and ascorbate transport characterizes, respectively, 65 and 62 out of 66 ChRCC samples and 45 and 28 out of 774 other tumor samples. Even though in the literature reports it is still not clear how the phosphate transportation or concentration level and absorption via SLC23A1 of ascorbate influences the cancer cells (Wohlrab *et al.*, 2017). We point these factors, both reaction activity and transcriptomic/proteomic levels, as possible biomarkers of ChRCC.

Poor Prognosis Cluster

Finally, we report a candidate for new subtypes resulting from the statistical and functional analysis of metabolic landscape clusters. Based on samples clustering after the bias reduction we have further studied the obtained 6 clusters of samples. Among them, there are four homogeneous clusters composed mainly of: healthy tissues (Control), Chromophobe RTCC subtype (Chr-basal), Clear Cell RTCC (CC-basal) and Papillary RTCC (Pap-basal). Even though, the pattern of remaining two clusters was not related to their morphological type we have noticed a statistically significant (p-value: 0.002) difference in survival time in one of the clusters (see Figure 4.5), which we address as a poor prognosis cluster.

Since the unknown clusters are mixture of Clear Cell and Papillary subtypes, we performed a differential analysis of four clusters (using also CC-basal and Pap-basal) in order to describe a metabolic as well as genetic nature of this cluster. We determined a set of 106 differentiating reactions coordinated by four genes: OAT2 (3), PRODH2 (4), PKLR (5) and SLC1A2 (94). Among these reactions we report orotate-glutamate antiport, uptake of allopurinol and oxypurinol by the hepatocytes, mitochondrial proline Oxidase (NAD) and dehydrogenase and reactions involved in pyruvate metabolism. Genes coordinating these reactions reveal a specific expression pattern of the poor prognosis cluster, i.e. low expression and consequent inactivation of corresponding reactions.

Additionally, we performed the functional analysis of top 100 differentiating genes, using DAVID on-line tool. The analysis provided a consistent output indicating functions and keywords commonly related to modifications in transport and symport reactions highlighting transmembrane transport activity. Finally, pathway analysis performed with KEGG implies deregulations in Glycan Biosynthesis and Metabolism pathway.

The above observations can be preliminarily verified with literature reports. In the context of metabolomics, abnormalities in Glycan Biosinthesis that we observe were reported as a significant factor of cancer cells phenotype and biology almost two decades ago (Brockhausen, 1999). Moreover, the poor prognosis properties can be influenced by low activity of SLC1A2 belonging to the Organic-anion-transporting polypeptide (OATP) family, because it is responsible for transport of anticancer drugs (e.g. methotrexate used

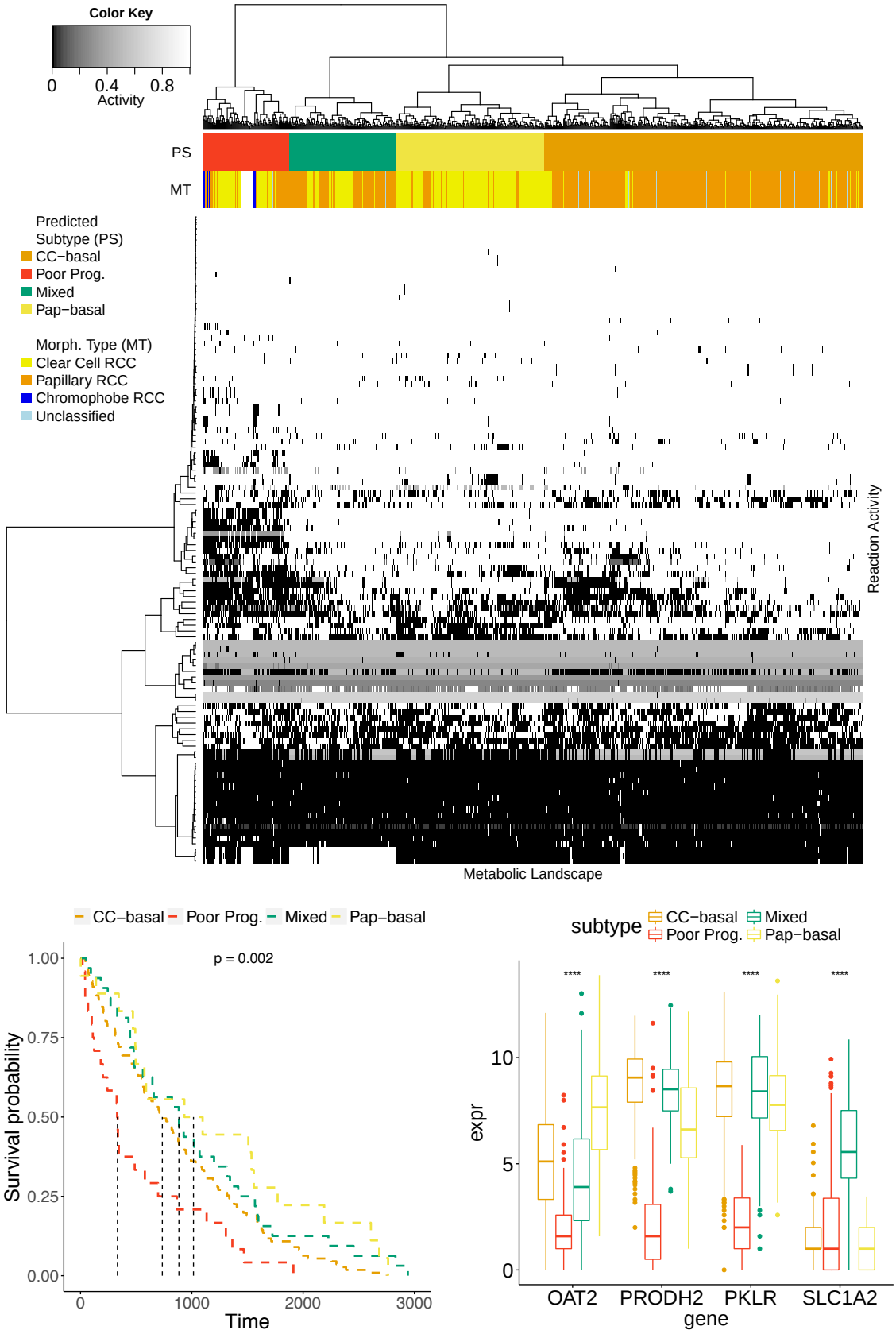


Figure 4.5: Analysis of the poor prognosis cluster. The heatmap presents reactions activity of samples from four clusters from hierarchical clustering composed mainly of Clear Cell and Papillary RCC. The horizontal stripes compare the predicted subtypes labeling with known morphological types. The bottom-left panel presents the Kaplan-Meier survival curves, which present the significantly lower survival time of patients from the poor prognosis cancer (p-value: 0.002). The bottom-right panel compares the expression level of genes responsible for the activity of four discriminating genes. For the poor prognosis cluster all 4 genes (and thus corresponding reactions) are characterized by low activity.

in chemotherapy) (Thakkar *et al.*, 2015) and overall uptake of. Similarly, low expression of OAT2 was reported to influence a poor response to antitumor UFT-based chemotherapy in colorectal cancer patients (Nishino *et al.*, 2013). Loss of PRODH2 was also reported in cancer (Loayza-Puch and Agami, 2016), however no links with cancer prognostics were reported. The above summary may constitute an introductory justification for the possible existence of the poor prognosis cluster in RCC mainly conditioned by the chemoresistance dictated by the activity of its potential biomarker genes.

4.7. Conclusions

Concluding, in this chapter we recall an important, yet barely discussed, data analysis and integration problem of twofold nature. Using the TCGA datasets and metabolic network model we presented how bioinformatical and statistical data analysis may lead to outwardly interesting biological and medical observations, which may be justified by the literature reports. However, not only do we prove these observations are the inevitable consequence of the model assumptions (as genetic rules induce clusters artificially), but also remind that these assumptions conceal the current state of knowledge about cell metabolism, as many discovery claims may testify.

To overcome this problem we propose a computational procedure for metabolic landscapes data processing and further analysis. We present an outcome of Renal Cell Carcinoma case-study, that proves the importance of thought-out data analysis and reasoning. Additionally, we report a possible existence of poor prognosis cluster of patients characterized by the low activity of drug transporting enzymes and thus possibly limited activity of absorption reactions.

5

Biomedical Applications

"When finally interpreted, the genetic messages encoded within our DNA molecules will provide the ultimate answers to the chemical underpinnings of human existence."

— James D. Watson

Bioinformatics is one of these fields of science, in which we can explore the meaning and significance of interdisciplinarity. When computational methods meet real-life, medical cases, and can help the world of medicine in diagnosis or therapeutics, we can assume that our work is going in the right direction.

In this chapter, we will see the results from scientific cooperation with Baylor Collage of Medicine. Our contribution includes results from transcriptomic data analysis, as well as some additional types of bioinformatical analysis of data related to the FOXF1 gene, for which all the specific functions have not yet been fully determined. However, our studies contributed to better understanding of its role for the organism from early state of embryonic development and the consequences that bring the manipulations of its expression pattern and copy number.

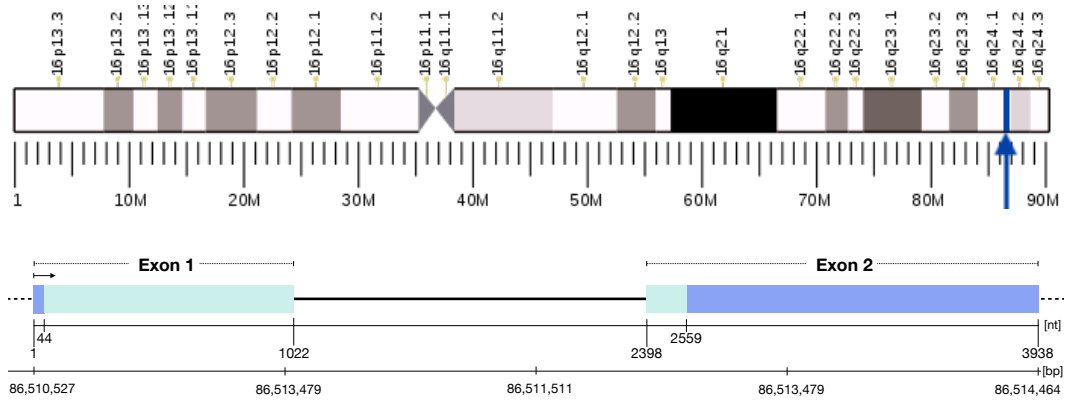


Figure 5.1: Schematic representation of position of FOXF1 on the chromosom 16 presented on the upper panel. The location is marked with blue stripe and arrow at 16q24.1. Additionally, the bottom panel shows the exon-intron organisation of the 3938 nucleotides long FOXF1 gene. The coding regions within the exons are depicted with light blue color, while darker one corresponds to non-coding sequences.

5.1. Biological Background of the FOXF1 Gene

In the genomic context FOXF1 (Forkhead Box F1) gene is quite well recognized. Its specific chromosomal location, on the current genome assembly GRCh38/hg38, is determined as chr16:86,510,527-86,515,418, cytogenic band 16q24.1, which is the long arm at position 24.1 (see Figure 5.1).

FOXF1 is a protein coding gene and it belongs to the forkhead family of transcription factors, which means it is responsible for binding to specific regions of DNA and helping in activity control of other genes. The family is additionally characterized by a distinct forkhead domain. All the specific functions of this gene have not yet been well-determined. Nonetheless, it is known to play a significant role in the pulmonary genes regulation and development of pulmonary mesenchyme as well as embryonic tissue development from which blood vessels of the lung arise (Maeda *et al.*, 2007).

From the signalling perspective, It is related to Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers and FOXA2/FOXA3 transcription factor pathways (Mahlapuu *et al.*, 2001). Consequently, when undergoing point mutations or genomic deletions, the gene is reported to be associated with Alveolar Capillary Dysplasia With Misalignment Of Pulmonary Veins (ACDMPV) and Persistent Fetal Circulation Syndrome diseases (Slot *et al.*, 2018; Shaw-Smith, 2010). Heterozygous deletions and point mutations in the FOXF1 gene locus were found in more than 40% of patients with ACDMPV (Stankiewicz *et al.*, 2009).

In this chapter, we present the results of the FOXF1 overexpression analysis and its consequences affecting the process of lung development. Additionally, starting from an identification of FOXF1 duplication adjacent to a large minisatellite, we will present the result of bioinformatical analysis of minisatellites of size greater than 1 kb among various

species, which demonstrated both its evolutionarily instability and population polymorphism that, possibly,may lead to structural variation due to DNA replication errors.

5.2. Hi-C and Transcriptomic Analysis of FOXF1 Knock-In

When it comes to expression patterns of FOXF1, its expression in fetal and adult lungs, placenta, and prostate was reported by (Bozyk *et al.*, 2011; van der Heul-Nieuwenhuijsen *et al.*, 2009; Hellqvist *et al.*, 1996). FOXF1 heterozygous point mutations and genomic deletions have been reported in newborns with the neonatally lethal lung developmental disorder, alveolar capillary dysplasia with misalignment of pulmonary veins (ACDMPV) with or without defects involving heart, gastrointestinal, or genitourinary systems (Sen *et al.*, 2013; Bishop *et al.*, 2011). Previous studies have also shown FOXF1 to be epigenetically inactivated in breast cancer (with potential role as a tumor suppressor gene) (Lo *et al.*, 2010) and overexpressed in basal cell carcinoma, medulloblastoma, and rhabdomyosarcoma (Armeanu-Ebinger *et al.*, 2011; Wendling *et al.*, 2008). Finally, overexpression was also tracked in lung fibroblasts from patients with idiopathic pulmonary fibrosis (Melboucy-Belkhir *et al.*, 2014). As one can see, while effects of the loss of function of FOXF1 in cancer and lung disease are well systematized, no gain-of-function mutations have been identified and effects of the overexpression of FOXF1, specifically in the context of lung development, are not currently known.

To study the effects of FOXF1 overexpression in lung development, biologists generated a FOXF1 overexpression mouse model by knocking-in a Cre-inducible FOXF1 allele into the ROSA26 (R26) locus. The Cre/lox site-specific recombination system is an laboratory tool for the generation of conditional somatic mouse mutants. The mice were phenotyped using micro-computed tomography (micro-CT), head-out plethysmography, ChIP-seq and transcriptome analyses, immunohistochemistry, and lung histopathology, out of which we have analysed ChIP-seq and transcriptomic data.

Statistical Analysis of Transciprtomic Microarray Data

The microarray data were analysed using the `lumi` bioconductor package (Du *et al.*, 2008), normalized by robust spline normalization and transformed using variance stabilization transformation (VST). Two-sample T-test was applied to determine differentially expressed genes between R26FOXF1; Tie2-cre and the R26-LSL-FOXF1 lung groups. Differential expression p-values were adjusted for false discovery rates (FDR). Fold changes were calculated using reverse VST. Although our sample size was relatively small for parametric tests, we justified the use of T-tests in this study by large effect sizes of our analysis (De Winter, 2013). DAVID tool (Huang *et al.*, 2009) was used for gene ontology and pathway analyses.

Figure 5.2: Consensus FOXF1 binding motif identified from analysis of DNA sequences underlying ChIP-seq peaks. The figure was generated using the on-line WebLogo tool (`http://weblogo.berkeley.edu/logo.cgi`).

The performed statistical analyses of transcriptomes from microarray studies showed 1242 deregulated genes in R26FOXF1; Tie2-cre lungs compared to lungs from littermate controls (FDR < 0.05, absolute fold-change ≥ 1.2). 519 genes (41.79%) were down-regulated and 723 genes (58.21%) were up-regulated. DAVID analyses identified gene ontology terms related to protein transport, protein localization, cell adhesion, and blood vessel morphogenesis to be associated with the deregulated genes. Some of the downregulated genes included IGFBP3, PPARG, RCAN1 and PRKCDBP, which were already reported as related to lungs development and functioning. Here it is worth to highlight, that IGFBP3 and PRKCDBP are also inversely upregulated in ACDMPV lungs (Sen *et al.*, 2014), suggesting that these genes might be relevant to the role of FOXF1 in the pathology of ACDMPV. On the other hand among upregulated genes were FGFR2, EGFL7 and ROBO4. Interestingly, upregulated FGFR2 is associated with nitrofen-induced pulmonary hypoplasia (Friedmacher *et al.*, 2012),while EGFL7 is shown to cause embryonic lethality due to impaired angiogenesis in mice (Nichol *et al.*, 2010). Finally, ROBO4 is a vascular specific receptor known to inhibit endothelial migration (Park *et al.*, 2003). Additionally, using a threshold of 1.2-fold-change, comparison of the R26FOXF1;Tie2-cre microarray dataset with FOXF1 knock-out P0.5 lung dataset from (Sen *et al.*, 2014) revealed 165 genes to be commonly deregulated in both datasets. These genes are potential targets of FOXF1 in the lung, as their expression changes reciprocally with the loss or gain of FOXF1.

Bioinformatical Analysis of ChIP-Seq Data

The second analysis was conducted ChIP-Seq data, that shortly speaking is a method used to analyze protein interactions with DNA. In a nutshell, ChIP-seq combines chromatin immunoprecipitation (ChIP) with DNA sequencing (seq) to track the binding sites of DNA-associated proteins. In our study it is used to map global binding sites precisely for any protein of interest.

Biologists performed an experiment for two biological replicates of pooled E18.5 and control samples and used The Rubicon ThruPlex DNA-Seq library preparation system to prepare ChIP-Seq libraries for sequencing on the Illumina HiSeq sequencing system. We mapped the sequence reads on to the mm10 mouse genome using Bow-tie2 tool (Langmead

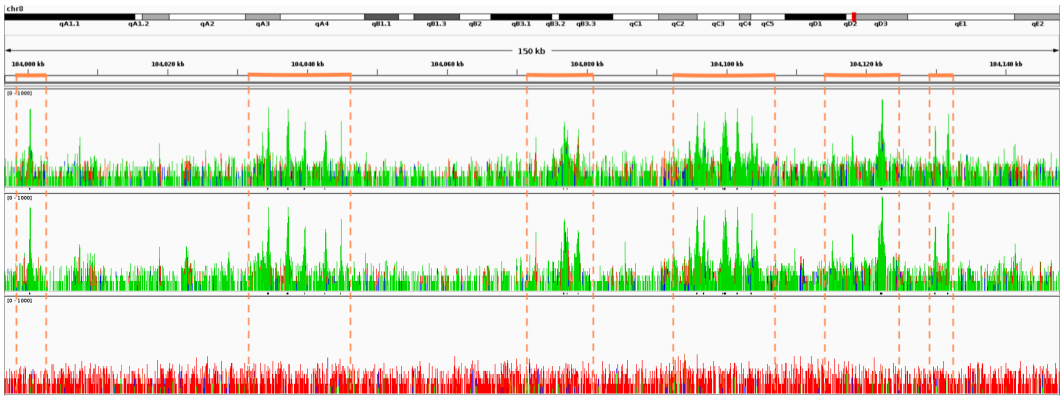


Figure 5.3: Visualization of peak calling procedure near CDH5 gene (approx. 150kb downstream) with important peaks highlighted in orange on the horizontal axis. Note, the strong conservation of peaks between replicates. Moreover, the peaks are easily observable compared to the background signal of the control sample.

and Salzberg, 2012). The obtained percentage of uniquely mapped reads was 77.85, 73.10 and 73.10%, respectively, for the two biological replicates and input control samples, which is an acceptable level according to (Bailey *et al.*, 2013). Next, we performed the peak calling procedure using the Model-based Analysis of ChIP-Seq (MACS2) tool (Zhang *et al.*, 2008) and assessed it using IDR-score (Li *et al.*, 2011). After detection of 696 peak regions, we have identified the consensus FOXF1 binding motif from the DNA sequences underlying the peaks 5.2.

Finally, the functional analysis of selected peak regions was performed using the GREAT tool (McLean *et al.*, 2010) and classification of genes common to the ChIP-seq and microarray datasets was done using LungGENS (Du *et al.*, 2015).

Conducted ChIP-seq analysis in E18.5 wild-type lungs identified binding of FOXF1 in the proximity of genes involved in biological processes such as blood vessel, cardiovascular and embryonic development. Among the genes associated with multiple FOXF1 binding sites, CDH5 and ITGB1 are genes involved in endothelial cell development (see Figure 5.3). Recently, it was shown that ITGB1 controls CDH5 localization and blood vessel stability (Yamamoto *et al.*, 2015). Additionally, NRP1 signaling has previously been shown to be essential for fetal pulmonary development (Joza *et al.*, 2013). Another forkhead gene, FOXA2 has been described to be required for the transition to breathing at birth (Wan *et al.*, 2004).

When the FOXF1 ChIP-seq data were compared to the FOXF1 knock-in and knock-out microarray datasets, 11 genes were found to have binding sites for FOXF1 and were reciprocally deregulated in the microarray datasets. These included the genes ARHGAP18, SOX11, ZSWIM6, TNFRSF19, EDNRB, GHR, 2510009E07RIK, OSTF1, SMARCA2, SLIT2 and NUP210. Interestingly, $SOX11^{-/-}$ mice exhibit lung hypoplasia and die at birth (Sock *et al.*, 2004). GHR signaling is involved in early lung growth, oxidative protection, and lipid metabolism in the developing lung (Beyea *et al.*, 2006). These genes could potentially be

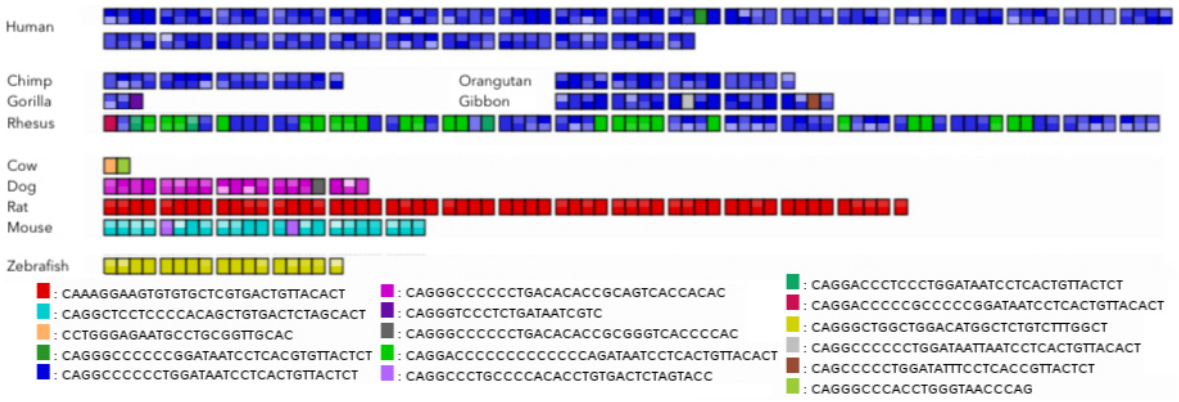


Figure 5.4: Cross-species visualization of syntenic sequences of the 8.6 kb minisatellite on chromosome 16q24.1. The figure presents the variation of motifs among different species and their conservation. Repeat sequences are represented as a row of multicolour strips and each strip maps to a particular group of motifs. The descending intensity of the color used to code the particular motif indicates the increasing number of differences from the motif corresponding to that color (mutational and deleterious differences are represented by upper and lower part of the strip, respectively). To increase the clarity of the figure, the human sequence was shortened to ∼4.1 kb, and divided into two rows. Moreover, the rat and dog sequences are represented by one of the insertional translocation sequences (Table 5.1).

direct targets of FOXF1 in the embryonic and early postnatal lung.

Summarising, our ChIP-seq and transcriptome analyses in E18.5 lungs identified SOX11, GHR, EDNRB, and SLIT2 as potential downstream targets of FOXF1. This interdisciplinary study shows that overexpression of the highly dosage-sensitive FOXF1 impairs lung development and causes vascular abnormalities. This has important clinical implications when considering potential gene therapy approaches to treat disorders of FOXF1 abnormal dosage, such as ACDMPV.

5.3. Minisatelites and Their Impact on DNA Replication

Four unrelated cases with duplications of 16q24.1 chromosomal site containing the entire FOXF1 were analysed by an interdisciplinary team. We are particularly interested in a fourth patient with speech and motor delay, and borderline intellectual disability. There was identified an ∼1.7 Mb FOXF1 duplication adjacent to a large minisatellite. By minisatelite we understand a Variable Number Tandem Repeat (VNTR) with repeat units between 10 and 100 nucleotides (nt) are defined as minisatellites. Additionally, we consider microsatellites with less than nine nt and macrosatellites with repeat units greater than 100 nt. VNTRs are extremely unstable, with mutation rates 10-100,000 times higher than non-repeat sequences. They tend to be highly polymorphic, expanding or contracting due to DNA strand replication or recombination slippage. The structure of the duplication, from the studied patient, was complex and arose *de novo* on the maternal chromosome. Probably, this was a consequence of a DNA replication error, that was initiated by the adjacent large tandem repeat.

Species	Orthologous genomic coordinates	Other genomic locations	Genome build
Human	chr16:85,437,697-85,446,384 (8688 bp)	N/A	hg19
Chimp	chr16(+):85,057,404-85,058,000 (597 bp)	chr6(−):171,902,631-171,903,434 (804 bp) chr16(−):87,565,917-87,566,290 (374 bp)	CSAC 2.1.4/panTro4
Gorilla	chr16(+):75,975,460-75,975,529 (70 bp)	chr7(+):156,341,820-156,342,473 (654 bp)	gorGor3.1/gorGor3
Orangutan	chr16(+):73,141,772-73,142,390 (619 bp)	chr1(−):12,527,885-12,528,916 (1032 bp)	WUGSC 2.0.2/ponAbe2
Gibbon	chr2(+):158,448,609-158,449,280 (672 bp)	chr17(−):95,524,509-95,526,825 (2317 bp) chr20(−):83,814,471-83,814,741 (271 bp)	GGSC Nleu3.0/nomLeu3
Rhesus	chr20(+):83,708,139-83,710,782 (2644 bp)	N/A	BGI CR_1.0/rheMac3
Cow	chr18(+):10,576,131-10.576,179 (49 bp)	N/A	Baylor Btau_4.6.1/bosTau7
Dog	chr5(+):67,222,084-67,222,093 (10 bp)	chr28(−):35,919,958-35,920,624 (667 bp)	Broad CanFam3.1/canFam3
Rat	chr14(+):3,644,385-3,645,161 (777 bp)	chr8(−):60,446,61-60,465,11 (1851 bp) chr12(+):82,176,55-82,180,67 (413 bp)	RGSC 5.0/rn5
Mouse	chr2(+):167,074,771-167,075,545 (775 bp)	chr3(+):79,000,591-79,000,853 (263 bp) chr3(−):144,489,926-144,491,349 (1424 bp) chr10(−):121,578,768-121,580,330 (1563 bp)	GRCm38/mm10
Zebra fish	chr16(+):8,911,924-8,912,616 (693 bp)	chr4(+):44,314,003-44,315,463 (1461 bp) chr8(+):41,186,318-41,186,865 (548 bp) chr15(−):11,992,633-11,992,908 (276 bp)	Zv9/danRer7

Table 5.1: Summary of the locations of orthologous sequences to the 8.6 kb minisatellite across various species. Visualization of these sequences is presented in the Figure 5.4

To determine the occurrence of polymorphism in the minisatellite region in the general population, the Database of Genomic Variants: Structural Variation track in the UCSC genome browser was used. Orthologous VNTR sequences to the minisatellite sequence in other genomes were obtained using the convert and blat functions in the UCSC genome browser. In the reference human genome, the large minisatellite is 8,688 kb in size, consisting of imperfect repeats of 33 bp sequence CAGGGCCCCCCGGATAATCCTCACTGTTACACT. The orthologous VNTRs of this minisatellite in several species are much shorter and range from 597 bp in Chimp to 2,644 bp in Rhesus (see Table 5.1), demonstrating its high instability during the evolution of the human genome.

In order to visualize syntenic (i.e. occurring on the same chromosome) sequences from the Table 5.1, main repeating motifs were extracted from each syntenic sequence and the pairwise alignment of them was done. Next, repeats composing each sequence were aligned to form one set consisting off all existing motifs among the considered species. Resulting set was clustered into several groups using the greedy clustering algorithm: the most numerous motifs were assumed to be centres, whilst number of differences was assumed to be the metric. Finally, each sequence was represented as a row of multicolour stripes and each strip maps onto particular group of motifs (see Fig. 5.4).

Interestingly, the duplication in the considered patient case mapping distal and adjacent to a large minisatellite contains in its junction fragment, truncated segments of this minisatellite. This minisatellite turns out to be highly polymorphic in the general population, with many deletions and duplications, indicating that it contracts and expands. The orthologous VNTRs of this minisatellite in several species are much shorter and range from 597 bp in Chimp to 2644 bp in Rhesus (see Figure 5.4, and Table 5.1), demonstrating its high instability during the evolution of the human genome.

Summarizing, this systematic form of data mining through genomes of different species along with biomedical interpretation of the collected data by our co-workers revealed an evolutionarily unstable and highly polymorphic minisatellite on 16q24.1. They propose that instability of minisatellites greater than 1 kb can lead to genomic structural variation due to DNA replication errors, which was supported by the case of studied patients. The lack of any pulmonary symptoms in the patient also suggests relatively benign paediatric pulmonary consequences of FOXF1 overexpression due to constitutional duplications.

6
Conclusions

"The measure of greatness in a scientific idea is the extent to which it stimulates thought and opens up new lines of research."

— Paul Dirac

In this dissertation several related topics in high-throughput data analysis have been approached with particular emphasis on transcriptomic and metabolomic data. Namely, we have discussed the analysis of expression activity data of cell-type homogeneous samples. We have introduced `MPH` method that enables estimation of the proportion of different sub-population conducting the same functional activity and their corresponding transcriptomic pattern. Next, two extensions of the robust PCA (`RPCA`) decomposition algorithm have been proposed. The first one is a truncated version of the algorithm (`tRPCA`) and the other (`tRPCAL2`) takes into account the dense noise component in the decomposition. Further on, we have analysed the transcriptomic data through integration with metabolomic data. For the needs of this task, we have introduced the procedure reducing the bias related to metabolic networks and approaches to feature selection in binary data. Finally, the results of two downstream bioinformatical analyses carried out for the purpose of cooperation with the Baylor Collage of Medicine have been were presented in the last chapter. For this purpose various bioinformatical tools have been used to infer from transcriptomic, ChIP-seq and microsatelite data.

As the reader could have noticed the presented scope of research is highly interdisci-

plinary. On one hand, we have theoretical approaches, methods and algorithms related to the matrix analysis and decomposition were presented, and on the other, each was supported by a case study with transcriptomic data from different high-throughput technologies. It is undeniable that it is the nowadays way in which the achievements of computer science supported by modern biology and biotechnology will set the direction for the field of bioinformatics.

Therefore, this work is naturally an input into the field of computer science, data analysis, feature selection and matrix decomposition techniques. However, let us hope that it is also a source of novel methods and workflows that constitute at least a small contribution to our understanding of the genotype–phenotype mapping mentioned in the Introduction.

To conclude let us recall the question stated by Walter Gilbert who received the Nobel Prize in Chemistry in 1980 for teaching us how to sequence DNA.

In the year 2020 you will be able to go into the drug store,
have your DNA sequence read in an hour or so,
and given back to you on a compact disc so you can analyse it.

To what extent he made a mistake by stating this sentence is a secondary matter. It is clear that with the current direction and intensity of research, the existence of the described reality is only a matter of time. Interesting seems to be a much more general question. Where, if there are any, are the limits of our comprehension of the reality that surrounds us, and in particular of the mechanisms that drive its living part? Are the capabilities of science unlimited, or alternatively there are realities beyond the cognitive abilities? Although it is difficult to respond to these questions, it seems that a form of an answer is to broaden the boundaries of science as far as possible, even by conducting research that has been presented in this dissertation.

Bibliography

Adachi, M., Ryo, R., Sato, T. and Yamaguchi, N. (1991). Platelet factor 4 gene expression in a human megakaryocytic leukemia cell line (CMK) and its differentiated subclone (CMK11-5). *Exp. Hematol.*, **19** (9), 923–927.

Aghdam, R., Ganjali, M., Zhang, X. and Eslahchi, C. (2015). CN: a consensus algorithm for inferring gene regulatory networks using the SORDER algorithm and conditional mutual information test. *Mol Biosyst*, **11** (3), 942–949.

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I. and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.*, **8** (5), e1002518.

Alberch, P. (1991). From genes to phenotype: dynamical systems and evolvability. *Genetica*, **84** (1), 5–11.

Armeanu-Ebinger, S., Bonin, M., Habig, K., Poremba, C., Koscielniak, E., Godzinski, J., Warmann, S. W., Fuchs, J. and Seitz, G. (2011). Differential expression of invasion promoting genes in childhood rhabdomyosarcoma. *Int. J. Oncol.*, **38** (4), 993–1000.

Baglama, J., Reichel, L. and Lewis, B. W. (2018). *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C. and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9** (11), e1003326.

Bartenhagen, C., Klein, H. U., Ruckert, C., Jiang, X. and Dugas, M. (2010). Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics*, **11**, 567.

Basu, S., Campbell, H. M., Dittel, B. N. and Ray, A. (2010). Purification of specific cell population by fluorescence activated cell sorting (FACS). *J Vis Exp*, (41).

Becker, S. A. and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.*, **4** (5), e1000082.

Beller, M. and Oliver, B. (2006). One hundred years of high-throughput Drosophila research. *Chromosome Res.*, **14** (4), 349–362.

Bernas, T. and Dobrucki, J. (2002). Mitochondrial and nonmitochondrial reduction of MTT: interaction of MTT with TMRE, JC-1, and NAO mitochondrial fluorescent probes. *Cytometry*, **47** (4), 236–242.

Beyea, J. A., Sawicki, G., Olson, D. M., List, E., Kopchick, J. J. and Harvey, S. (2006). Growth hormone (GH) receptor knockout mice reveal actions of GH in lung development. *Proteomics*, **6** (1), 341–348.

Bishop, N. B., Stankiewicz, P. and Steinhorn, R. H. (2011). Alveolar capillary dysplasia. *Am. J. Respir. Crit. Care Med.*, **184** (2), 172–179.

Blazier, A. S. and Papin, J. A. (2012). Integration of expression data in genome-scale metabolic network reconstructions. *Front Physiol*, **3**, 299.

Bordbar, A., Mo, M. L., Nakayasu, E. S., Schrimpe-Rutledge, A. C., Kim, Y. M., Metz, T. O., Jones, M. B., Frank, B. C., Smith, R. D., Peterson, S. N., Hyduke, D. R., Adkins, J. N. and Palsson, B. O. (2012). Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. *Mol. Syst. Biol.*, **8**, 558.

Borgan, E., Sitter, B., Lingjaerde, O. C., Johnsen, H., Lundgren, S., Bathen, T. F., Sorlie, T., Borresen-Dale, A. L. and Gribbestad, I. S. (2010). Merging transcriptomics and metabolomics–advances in breast cancer profiling. *BMC Cancer*, **10**, 628.

Borrageiro, G., Haylett, W., Seedat, S., Kuivaniemi, H. and Bardien, S. (2018). A review of genome-wide transcriptomics studies in Parkinson's disease. *Eur. J. Neurosci.*, **47** (1), 1–16.

Bozyk, P. D., Popova, A. P., Bentley, J. K., Goldsmith, A. M., Linn, M. J., Weiss, D. J. and Hershenson, M. B. (2011). Mesenchymal stromal cells from neonatal tracheal aspirates demonstrate a pattern of lung-specific gene expression. *Stem Cells Dev.*, **20** (11), 1995–2007.

Brockhausen, I. (1999). Pathways of O-glycan biosynthesis in cancer cells. *Biochim. Biophys. Acta*, **1473** (1), 67–95.

Brunet, J. P., Tamayo, P., Golub, T. R. and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.*, **101** (12), 4164–4169.

Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *J. ACM*, **58** (3), 11:1–11:37.

Canto, C., Menzies, K. J. and Auwerx, J. (2015). NAD(+) Metabolism and the Control of Energy Homeostasis: A Balancing Act between Mitochondria and the Nucleus. *Cell Metab.*, **22** (1), 31–53.

Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M. and Pascual-Montano, A. (2006). Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. *BMC Bioinformatics*, **7**, 78.

Chandrasekaran, S. and Price, N. D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (41), 17845–17850.

Chu, L. F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., Choi, J., Kendziorski, C., Stewart, R. and Thomson, J. A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, **17** (1), 173.

Chu, P. G. and Arber, D. A. (2001). CD79: a review. *Appl. Immunohistochem. Mol. Morphol.*, **9** (2), 97–106.

Chung, N. C. and Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, **31** (4), 545–554.

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T. Y., Moody, D. B., Murray, M. and Galagan, J. E. (2009). Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput. Biol.*, **5** (8), e1000489.

Covert, M. W. and Palsson, B. . (2002). Transcriptional regulation in constraints-based metabolic models of Escherichia coli. *J. Biol. Chem.*, **277** (31), 28058–28064.

Crinier, A., Milpied, P., Escaliere, B., Piperoglou, C., Galluso, J., Balsamo, A., Spinelli, L., Cervera-Marzal, I., Ebbo, M., Girard-Madoux, M., Jaeger, S., Bollon, E., Hamed, S., Hardwigsen, J., Ugolini, S., Vely, F., Narni-Mancinelli, E. and Vivier, E. (2018). High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. *Immunity*, **49** (5), 971–986.

Czubowicz, K. and Strosznajder, R. (2014). Ceramide in the molecular mechanisms of neuronal cell death. The role of sphingosine-1-phosphate. *Mol. Neurobiol.*, **50** (1), 26–37.

De Winter, J. C. F. (2013). Using the Student's t-test with extremely small sample sizes. *Pr. Assess. Res. Eval.*, **18**, 1–12.

DeTomaso, D. and Yosef, N. (2016). FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics*, **17** (1), 315.

Dharmadhikari, A. V., Gambin, T., Szafranski, P., Cao, W., Probst, F. J., Jin, W., Fang, P., Gogolewski, K., Gambin, A., George-Abraham, J. K., Golla, S., Boidein, F., Duban-Bedu, B., Delobel, B., Andrieux, J., Becker, K., Holinski-Feder, E., Cheung, S. W. and Stankiewicz, P. (2014). Molecular and clinical analyses of 16q24.1 duplications involving FOXF1 identify an evolutionarily unstable large minisatellite. *BMC Med. Genet.*, **15**, 128.

—, Sun, J. J., Gogolewski, K., Carofino, B. L., Ustiyan, V., Hill, M., Majewski, T., Szafranski, P., Justice, M. J., Ray, R. S., Dickinson, M. E., Kalinichenko, V. V., Gambin, A. and Stankiewicz, P. (2016). Lethal lung hypoplasia and vascular defects in mice with conditional Foxf1 overexpression. *Biol Open*, **5** (11), 1595–1606.

Dojer, N., Gambin, A., Mizera, A., Wilczyński, B. and Tiuryn, J. (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, **7**, 249.

Dozmorov, M. G., Dominguez, N., Bean, K., Macwana, S. R., Roberts, V., Glass, E., James, J. A. and Guthridge, J. M. (2015). B-Cell and Monocyte Contribution to Systemic Lupus Erythematosus Identified by Cell-Type-Specific Differential Expression Analysis in RNA-Seq Data. *Bioinform Biol Insights*, **9** (Suppl 3), 11–19.

Du, P., Kibbe, W. A. and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24** (13), 1547–1548.

Du, Y., Guo, M., Whitsett, J. A. and Xu, Y. (2015). 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax*, **70** (11), 1092–1094.

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., Wang, Y. and Wong, W. H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U.S.A.*, **115** (30), 7723–7728.

Erkkila, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I. and Lahdesmaki, H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, **26** (20), 2571–2577.

Ewing, A. D. and Kazazian, H. H. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.*, **20** (9), 1262–1270.

Ferrari, G., Terushkin, V., Wolff, M. J., Zhang, X., Valacca, C., Poggio, P., Pintucci, G. and Mignatti, P. (2012). TGF-β1 induces endothelial cell apoptosis by shifting VEGF activation of p38(MAPK) from the prosurvival p38β to proapoptotic p38α. *Mol. Cancer Res.*, **10** (5), 605–614.

Fondi, M. and Lio, P. (2015). Multi -omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiol. Res.*, **171**, 52–64.

Friedmacher, F., Doi, T., Gosemann, J. H., Fujiwara, N., Kutasy, B. and Puri, P. (2012). Upregulation of fibroblast growth factor receptor 2 and 3 in the late stages of fetal lung development in the nitrofen rat model. *Pediatr. Surg. Int.*, **28** (2), 195–199.

Fritz, J. V., Desai, M. S., Shah, P., Schneider, J. G. and Wilmes, P. (2013). From meta-omics to causality: experimental models for human microbiome research. *Microbiome*, **1** (1), 14.

Gaujoux, R. and Seoighe, C. (2013). CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, **29** (17), 2211–2212.

Gjuvsland, A. B., Vik, J. O., Beard, D. A., Hunter, P. J. and Omholt, S. W. (2013). Bridging the genotype-phenotype gap: what does it take? *J. Physiol. (Lond.)*, **591** (8), 2055–2066.

Gogolewski, K. and Gambin, A. (2018). Pca-like methods for the integration of single cell rna-seq data with metabolic networks. In *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pp. 1–1.

—, Kostecki, M. and Gambin, A. (2018a). Renal cell carcinoma classification: a case study of pitfalls associated with metabolic landscape analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 96–101.

—, Sykulski, M., Chung, N. C. and Gambin, A. (2018b). Truncated robust principal component analysis and noise reduction for single cell rna-seq data. In F. Zhang, Z. Cai, P. Skums and S. Zhang (eds.), *Bioinformatics Research and Applications*, Cham: Springer International Publishing, pp. 335–346.

—, Wronowska, W., Lech, A., Lesyng, B. and Gambin, A. (2017). Inferring Molecular Processes Heterogeneity from Transcriptional Data. *Biomed Res Int*, **2017**.

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S. and Szustakowski, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE*, **6** (11), e27156.

— and Szustakowski, J. D. (2013). DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, **29** (8), 1083–1085.

Gurobi Optimization, L. (2018). Gurobi optimizer reference manual.

Hansson, J., Lindgren, D., Nilsson, H., Johansson, E., Johansson, M., Gustavsson, L. and Axelson, H. (2017). Overexpression of Functional SLC6A3 in Clear Cell Renal Cell Carcinoma. *Clin. Cancer Res.*, **23** (8), 2105–2115.

Hardwick, L. J., Ali, F. R., Azzarelli, R. and Philpott, A. (2015). Cell cycle regulation of proliferation versus differentiation in the central nervous system. *Cell Tissue Res.*, **359** (1), 187–200.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means clustering algorithm. *Applied Statistics*, **28** (1), 100–108.

Hasin, Y., Seldin, M. and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.*, **18** (1), 83.

Hellqvist, M., Mahlapuu, M., Samuelsson, L., Enerback, S. and Carlsson, P. (1996). Differential activation of lung-specific genes by two forkhead proteins, FREAC-1 and FREAC-2. *J. Biol. Chem.*, **271** (8), 4482–4490.

Huang, D. A. W., Sherman, B. T. and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37** (1), 1–13.

Huang, S., Chaudhary, K. and Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet*, **8**, 84.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C. and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.*, **17**, 29.

Jack, D. L., Paulsen, I. T. and Saier, M. H. (2000). The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology (Reading, Engl.)*, **146 (Pt 8)**, 1797–1814.

Jensen, P. A. and Papin, J. A. (2011). Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, **27** (4), 541–547.

Jia, Z., Zhang, X., Guan, N., Bo, X., Barnes, M. R. and Luo, Z. (2015). Gene Ranking of RNA-Seq Data via Discriminant Non-Negative Matrix Factorization. *PLoS ONE*, **10** (9), e0137782.

Jiang, Y., Cao, Y., Wang, Y., Li, W., Liu, X., Lv, Y., Li, X. and Mi, J. (2017). Cysteine transporter SLC3A1 promotes breast cancer tumorigenesis. *Theranostics*, **7** (4), 1036–1046.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32** (3), 241–254.

Joza, S., Wang, J., Tseu, I., Ackerley, C. and Post, M. (2013). Fetal, but not postnatal, deletion of semaphorin-neuropilin-1 signaling affects murine alveolar development. *Am. J. Respir. Cell Mol. Biol.*, **49** (4), 627–636.

Kadara, H., Schroeder, C. P., Lotan, D., Pisano, C. and Lotan, R. (2006). Induction of GDF-15/NAG-1/MIC-1 in human lung carcinoma cells by retinoid-related molecules and assessment of its role in apoptosis. *Cancer Biol. Ther.*, **5** (5), 518–522.

Kim, J. and Reed, J. L. (2012). RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biol.*, **13** (9), R78.

Kim, M. K. and Lun, D. S. (2014). Methods for integration of transcriptomic data in genome-scale metabolic models. *Comput Struct Biotechnol J*, **11** (18), 59–65.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9** (4), 357–359.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401** (6755), 788–791.

Lee, S., Huang, J. Z. and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *Ann Appl Stat*, **4** (3), 1579–1601.

Leoncikas, V., Wu, H., Ward, L. T., Kierzek, A. M. and Plant, N. J. (2016). Generation of 2,000 breast cancer metabolic landscapes reveals a poor prognosis group with active serotonin production. *Sci Rep*, **6**, 19771.

Li, A., Walling, J., Ahn, S., Kotliarov, Y., Su, Q., Quezado, M., Oberholtzer, J. C., Park, J., Zenklusen, J. C. and Fine, H. A. (2009). Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res.*, **69** (5), 2091–2099.

Li, L., Zhou, X., Ching, W. K. and Wang, P. (2010). Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines. *BMC Bioinformatics*, **11**, 501.

Li, Q., Brown, J. B., Huang, H. and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5** (3), 1752–1779.

Li, S., Todor, A. and Luo, R. (2016). Blood transcriptomics and metabolomics for personalized medicine. *Comput Struct Biotechnol J*, **14**, 1–7.

Li, T. (2005). A general model for clustering binary data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, New York, NY, USA: ACM, pp. 188–197.

Li, Y. and Xie, X. (2013). A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics*, **14 Suppl 5**, S11.

Liu, S. and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res*, **5**.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, **28**, 129–137.

Lo, P. K., Lee, J. S., Liang, X., Han, L., Mori, T., Fackler, M. J., Sadik, H., Argani, P., Pandita, T. K. and Sukumar, S. (2010). Epigenetic inactivation of the potential tumor suppressor gene FOXF1 in breast cancer. *Cancer Res.*, **70** (14), 6047–6058.

Loayza-Puch, F. and Agami, R. (2016). Monitoring amino acid deficiencies in cancer. *Cell Cycle*, **15** (17), 2229–2230.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S. and Shafee, T. (2017). Transcriptomics technologies. *PLoS Comput. Biol.*, **13** (5), e1005457.

Mackenzie, F. and Ruhrberg, C. (2012). Diverse roles for VEGF-A in the nervous system. *Development*, **139** (8), 1371–1380.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Maeda, Y., Dave, V. and Whitsett, J. A. (2007). Transcriptional control of lung morphogenesis. *Physiol. Rev.*, **87** (1), 219–244.

Mahlapuu, M., Enerback, S. and Carlsson, P. (2001). Haploinsufficiency of the forkhead gene Foxf1, a target for sonic hedgehog signaling, causes lung and foregut malformations. *Development*, **128** (12), 2397–2406.

Mardinoglu, A. and Nielsen, J. (2012). Systems medicine and metabolic modelling. *J. Intern. Med.*, **271** (2), 142–154.

Marin de Mas, I., Aguilar, E., Zodda, E., Balcells, C., Marin, S., Dallmann, G., Thomson, T. M., Papp, B. and Cascante, M. (2018). Model-driven discovery of long-chain fatty acid metabolic reprogramming in heterogeneous prostate cancer cells. *PLoS Comput. Biol.*, **14** (1), e1005914.

Masoudi-Nejad, A. and Asgari, Y. (2015). Metabolic cancer biology: structural-based analysis of cancer as a metabolic disease, new sights and opportunities for disease treatment. *Semin. Cancer Biol.*, **30**, 21–29.

McGuire, B. B. and Fitzpatrick, J. M. (2009). Biomarkers in renal cell carcinoma. *Curr Opin Urol*, **19** (5), 441–446.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28** (5), 495–501.

Melboucy-Belkhir, S., Pradere, P., Tadbiri, S., Habib, S., Bacrot, A., Brayer, S., Mari, B., Besnard, V., Mailleux, A., Guenther, A., Castier, Y., Mal, H., Crestani, B. and Plantier, L. (2014). Forkhead Box F1 represses cell growth and inhibits COL1 and ARPC2 expression in lung fibroblasts in vitro. *Am. J. Physiol. Lung Cell Mol. Physiol.*, **307** (11), L838–847.

Muglia, V. F. and Prando, A. (2015). Renal cell carcinoma: histological classification and correlation with imaging findings. *Radiol Bras*, **48** (3), 166–174.

Murtagh, F. (1985). *Multidimensional Clustering Algorithms*. Wuerzburg: Physica-Verlag.

Nichol, D., Shawber, C., Fitch, M. J., Bambino, K., Sharma, A., Kitajewski, J. and Stuhlmann, H. (2010). Impaired angiogenesis and altered Notch signaling in mice overexpressing endothelial Egfl7. *Blood*, **116** (26), 6133–6143.

Nishino, S., Itoh, A., Matsuoka, H., Maeda, K. and Kamoshida, S. (2013). Immunohistochemical analysis of organic anion transporter 2 and reduced folate carrier 1 in colorectal cancer: Significance as a predictor of response to oral uracil/ftorafur plus leucovorin chemotherapy. *Mol Clin Oncol*, **1** (4), 661–667.

Novelli, G., Ciccacci, C., Borgiani, P., Papaluca Amati, M. and Abadie, E. (2008). Genetic tests and genomic biomarkers: regulation, qualification and validation. *Clin Cases Miner Bone Metab*, **5** (2), 149–154.

Ohkawa, T., Seki, S., Dobashi, H., Koike, Y., Habu, Y., Ami, K., Hiraide, H. and Sekine, I. (2001). Systematic characterization of human CD8+ T cells with natural killer cell markers in comparison with natural killer cells and normal CD8+ T cells. *Immunology*, **103** (3), 281–290.

Orth, J. D., Thiele, I. and Palsson, B. . (2010). What is flux balance analysis? *Nat. Biotechnol.*, **28** (3), 245–248.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5** (2), 111–126.

Park, K. W., Morrison, C. M., Sorensen, L. K., Jones, C. A., Rao, Y., Chien, C. B., Wu, J. Y., Urness, L. D. and Li, D. Y. (2003). Robo4 is a vascular-specific receptor that inhibits endothelial migration. *Dev. Biol.*, **261** (1), 251–267.

Pigliucci, M. (2010). Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **365** (1540), 557–566.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, **14** (10), 979–982.

Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P. and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30** (8), 777–782.

Raposo, A. A., Vasconcelos, F. F., Drechsel, D., Marie, C., Johnston, C., Dolle, D., Bithell, A., Gillotin, S., van den Berg, D. L., Ettwiller, L., Flicek, P., Crawford, G. E., Parras, C. M., Berninger, B., Buckley, N. J., Guillemot, F. and Castro, D. S. (2015). Ascl1 Coordinately Regulates Gene Expression and the Chromatin Landscape during Neurogenesis. *Cell Rep*, **10** (9), 1544–1556.

Ren, S., Shao, Y., Zhao, X., Hong, C. S., Wang, F., Lu, X., Li, J., Ye, G., Yan, M., Zhuang, Z., Xu, C., Xu, G. and Sun, Y. (2016). Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer. *Mol. Cell Proteomics*, **15** (1), 154–163.

Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J., Parida, S. K., Kaufmann, S. H. and Jacobsen, M. (2010). Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 27.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell*, **58** (4), 586–597.

Riekeberg, E. and Powers, R. (2017). New frontiers in metabolomics: from measurement to insight. *F1000Res*, **6**, 1148.

Schmidt, M. N., Winther, O. and Hansen, L. (2009). Bayesian non-negative matrix factorization. In H. Springer, Berlin (ed.), *Adali T., Jutten C., Romano J.M.T., Barros A.K. (eds) Independent Component Analysis and Signal Separation. ICA*, vol. 5441, pp. 540–547.

Sen, P., Dharmadhikari, A. V., Majewski, T., Mohammad, M. A., Kalin, T. V., Zabielska, J., Ren, X., Bray, M., Brown, H. M., Welty, S., Thevananther, S., Langston, C., Szafranski, P., Justice, M. J., Kalinichenko, V. V., Gambin, A., Belmont, J. and Stankiewicz, P. (2014). Comparative analyses of lung transcriptomes in patients with alveolar capillary dysplasia with misalignment of pulmonary veins and in foxf1 heterozygous knockout mice. *PLoS ONE*, **9** (4), e94390.

—, Yang, Y., Navarro, C., Silva, I., Szafranski, P., Kolodziejska, K. E., Dharmadhikari, A. V., Mostafa, H., Kozakewich, H., Kearney, D., Cahill, J. B., Whitt, M., Bilic, M., Margraf, L., Charles, A., Goldblatt, J., Gibson, K., Lantz, P. E., Garvin, A. J., Petty, J., Kiblawi, Z., Zuppan, C., McConkie-Rosell, A., McDonald, M. T., Peterson-Carmichael, S. L., Gaede, J. T., Shivanna, B., Schady, D., Friedlich, P. S., Hays, S. R., Palafoll, I. V., Siebers-Renelt, U., Bohring, A., Finn, L. S., Siebert, J. R., Galambos, C., Nguyen, L., Riley, M., Chassaing, N., Vigouroux, A., Rocha, G., Fernandes, S., Brumbaugh, J., Roberts, K., Ho-Ming, L., Lo, I. F., Lam, S., Gerychova, R., Jezova, M., Valaskova, I., Fellmann, F., Afshar, K., Giannoni, E., Muhlethaler, V., Liang, J., Beckmann, J. S., Lioy, J., Deshmukh, H., Srinivasan, L., Swarr, D. T., Sloman, M., Shaw-Smith, C., van Loon, R. L., Hagman, C., Sznajer, Y., Barrea, C., Galant, C., Detaille, T., Wambach, J. A., Cole, F. S., Hamvas, A., Prince, L. S., Diderich, K. E., Brooks, A. S., Verdijk, R. M., Ravindranathan, H., Sugo, E., Mowat, D., Baker, M. L., Langston, C., Welty, S. and Stankiewicz, P. (2013). Novel FOXF1 mutations in sporadic and familial cases of alveolar capillary dysplasia with misaligned pulmonary veins imply a role for its DNA binding domain. *Hum. Mutat.*, **34** (6), 801–811.

Shao, C. and Hofer, T. (2017). Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics*, **33** (2), 235–242.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14** (9), 618–630.

Shaw-Smith, C. (2010). Genetic factors in esophageal atresia, tracheo-esophageal fistula and the VACTERL association: roles for FOXF1 and the 16q24.1 FOX transcription factor gene cluster, and review of the literature. *Eur J Med Genet*, **53** (1), 6–13.

Shen-Orr, S. S. and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.*, **25** (5), 571–578.

Shin, T. H., Lee, D. Y., Lee, H. S., Park, H. J., Jin, M. S., Paik, M. J., Manavalan, B., Mo, J. S. and Lee, G. (2018). Integration of metabolomics and transcriptomics in nanotoxicity studies. *BMB Rep*, **51** (1), 14–20.

Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. . and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26** (9), 1003–1010.

Slot, E., Edel, G., Cutz, E., van Heijst, A., Post, M., Schnater, M., Wijnen, R., Tibboel, D., Rottier, R. and de Klein, A. (2018). Alveolar capillary dysplasia with misalignment of the pulmonary veins: clinical, histological, and genetic aspects. *Pulm Circ*, **8** (3), 2045894018795143.

Sock, E., Rettig, S. D., Enderich, J., Bosl, M. R., Tamm, E. R. and Wegner, M. (2004). Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Mol. Cell. Biol.*, **24** (15), 6635–6644.

Spies, D. and Ciaudo, C. (2015). Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput Struct Biotechnol J*, **13**, 469–477.

Stankiewicz, P., Sen, P., Bhatt, S. S., Storer, M., Xia, Z., Bejjani, B. A., Ou, Z., Wiszniewska, J., Driscoll, D. J., Maisenbacher, M. K., Bolivar, J., Bauer, M., Zackai, E. H., McDonald-McGinn, D., Nowaczyk, M. M., Murray, M., Hustead, V., Mascotti, K., Schultz, R., Hallam, L., McRae, D., Nicholson, A. G., Newbury, R., Durham-O'Donnell, J., Knight, G., Kini, U., Shaikh, T. H., Martin, V., Tyreman, M., Simonic, I., Willatt, L., Paterson, J., Mehta, S., Rajan, D., Fitzgerald, T., Gribble, S., Prigmore, E., Patel, A., Shaffer, L. G., Carter, N. P., Cheung, S. W., Langston, C. and Shaw-Smith, C. (2009). Genomic and genic deletions of the FOX gene cluster on 16q24.1 and inactivating mutations of FOXF1 cause alveolar capillary dysplasia and other malformations. *Am. J. Hum. Genet.*, **84** (6), 780–791.

Stempler, S., Yizhak, K. and Ruppin, E. (2014). Integrating transcriptomics with metabolic modeling predicts biomarkers and drug targets for Alzheimer's disease. *PLoS ONE*, **9** (8), e105383.

Swainston, N., Smallbone, K., Hefzi, H. and Dobson et al., P. D. (2016). Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, **12**, 109.

Sykulski, M. (2015). *rpca: RobustPCA: Decompose a Matrix into Low-Rank and Sparse Components.*

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6** (5), 377–382.

Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290** (5500), 2319–2323.

Thakkar, N., Lockhart, A. C. and Lee, W. (2015). Role of Organic Anion-Transporting Polypeptides (OATPs) in Cancer Therapy. *AAPS J*, **17** (3), 535–545.

Thurau, C., Kersting, K., Wahabzada, M. and Bauckhage, C. (2011). Convex non-negative matrix factorization for massive datasets. *Knowledge and Information Systems*, **29** (2), 457–478.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32** (4), 381–386.

Tseng, T. T., Gratwick, K. S., Kollman, J., Park, D., Nies, D. H., Goffeau, A. and Saier, M. H. (1999). The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. *J. Mol. Microbiol. Biotechnol.*, **1** (1), 107–125.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lonnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggstrom, J., Kharchenko, O., Kharchenko, P. V., Linnarsson, S. and Ernfors, P. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.*, **18** (1), 145–153.

van der Heul-Nieuwenhuijsen, L., Dits, N. F. and Jenster, G. (2009). Gene expression of forkhead transcription factors in the normal and diseased human prostate. *BJU Int.*, **103** (11), 1574–1580.

van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, **15**, 3221–3245.

von Kriegsheim, A., Baiocchi, D., Birtwistle, M., Sumpton, D., Bienvenut, W., Morrice, N., Yamada, K., Lamond, A., Kalna, G., Orton, R., Gilbert, D. and Kolch, W. (2009). Cell fate decisions are specified by the dynamic ERK interactome. *Nat. Cell Biol.*, **11** (12), 1458–1464.

Wan, H., Xu, Y., Ikegami, M., Stahlman, M. T., Kaestner, K. H., Ang, S. L. and Whitsett, J. A. (2004). Foxa2 is required for transition to air breathing at birth. *Proc. Natl. Acad. Sci. U.S.A.*, **101** (40), 14449–14454.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. and Batzoglou, S. (2017a). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, **14** (4), 414–416.

Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R. M., Hudson, J. R., Davis, R. W. and Han, J. (2010). High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (4), 1518–1523.

Wang, J. and Song, Y. (2017). Single cell sequencing: a distinct new field. *Clin Transl Med*, **6** (1), 10.

Wang, N., Dong, B. J., Quan, Y., Chen, Q., Chu, M., Xu, J., Xue, W., Huang, Y. R., Yang, R. and Gao, W. Q. (2016). Regulation of Prostate Development and Benign Prostatic Hyperplasia by Autocrine Cholinergic Signaling via Maintaining the Epithelial Progenitor Cells in Proliferating Status. *Stem Cell Reports*, **6** (5), 668–678.

Wang, S., MacLean, A. and Nie, Q. (2017b). Low-rank similarity matrix optimization identifies subpopulation structure and orders single cells in pseudotime. *bioRxiv*.

Wang, X., Baek, S. J. and Eling, T. E. (2013). The diverse roles of nonsteroidal anti-inflammatory drug activated gene (NAG-1/GDF15) in cancer. *Biochem. Pharmacol.*, **85** (5), 597–606.

Wang, Y., Eddy, J. A. and Price, N. D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol*, **6**, 153.

— and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Mol. Cell*, **58** (4), 598–609.

Wanichthanarak, K., Fahrmann, J. F. and Grapov, D. (2015). Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomark Insights*, **10** (Suppl 4), 1–6.

Watson, B. S., Bedair, M. F., Urbanczyk-Wochniak, E., Huhman, D. V., Yang, D. S., Allen, S. N., Li, W., Tang, Y. and Sumner, L. W. (2015). Integrated metabolomics and transcriptomics reveal enhanced specialized metabolism in Medicago truncatula root border cells. *Plant Physiol.*, **167** (4), 1699–1716.

Wendling, D. S., Luck, C., von Schweinitz, D. and Kappler, R. (2008). Characteristic overexpression of the forkhead box transcription factor Foxf1 in Patched-associated tumors. *Int. J. Mol. Med.*, **22** (6), 787–792.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W. and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.*, **31** (8), 748–752.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10** (3), 515–534.

Wohlrab, C., Phillips, E. and Dachs, G. U. (2017). Vitamin C Transporters in Cancer: Current Understanding and Gaps in Knowledge. *Front Oncol*, **7**, 74.

Yamamoto, H., Ehling, M., Kato, K., Kanai, K., van Lessen, M., Frye, M., Zeuschner, D., Nakayama, M., Vestweber, D. and Adams, R. H. (2015). Integrin Î²1 controls VE-cadherin localization and blood vessel stability. *Nat Commun*, **6**, 6429.

Yang, K., Xia, B., Wang, W., Cheng, J., Yin, M., Xie, H., Li, J., Ma, L., Yang, C., Li, A., Fan, X., Dhillon, H. S., Hou, Y., Lou, G. and Li, K. (2017). A Comprehensive Analysis of Metabolomics and Transcriptomics in Cervical Cancer. *Sci Rep*, **7**, 43353.

Yu, J., Baron, V., Mercola, D., Mustelin, T. and Adamson, E. D. (2007). A network of p73, p53 and Egr1 is required for efficient apoptosis in tumor cells. *Cell Death Differ.*, **14** (3), 436–446.

Yuan, X. and Yang, J. (2009). Sparse and Low-Rank Matrix Decomposition Via Alternating Direction Methods. *optimization-online.org*.

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J. K., Liu, Z. P. and Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, **28** (1), 98–104.

Zhang, Y., Guo, J., Zhao, J. and Wang, B. (2016). Robust principal component analysis via truncated nuclear norm minimization. *J. Shang. Jiaot. Uni.*, **21** (5), 576–583.

—, Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9** (9), R137.

Zhang, Z., Li, T., Ding, C. and Zhang, X. (2007). Binary matrix factorization with applications. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 391–400.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, **8**, 14049.

Zhou, Y., Yuan, J., Li, Z., Wang, Z., Cheng, D., Du, Y., Li, W., Kan, Q. and Zhang, W. (2015). Genetic polymorphisms and function of the organic anion-transporting polypeptide 1A2 and its clinical relevance in drug disposition. *Pharmacology*, **95** (3-4), 201–208.

Zhu, X., Ching, T., Pan, X., Weissman, S. M. and Garmire, L. (2017). Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ*, **5**, e2888.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell*, **65** (4), 631–643.

Ziegler-Heitbrock, L., Ancuta, P., Crowe, S., Dalod, M., Grau, V., Hart, D. N., Leenen, P. J., Liu, Y. J., MacPherson, G., Randolph, G. J., Scherberich, J., Schmitz, J., Shortman, K., Sozzani, S., Strobl, H., Zembala, M., Austyn, J. M. and Lutz, M. B. (2010). Nomenclature of monocytes and dendritic cells in blood. *Blood*, **116** (16), 74–80.

Zur, H., Ruppin, E. and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. *Bioinformatics*, **26** (24), 3140–3142.