Modeling MYC-dependent regulation of gene expression and cell metabolism in B-cell lymphomas

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von

Franziska Taruttis-Glagoleff

aus

Eisleben

im Jahr 2020

Der Promotionsgesuch wurde eingereicht am:
08.10.2020

Die Arbeit wurde angeleitet von:
Prof. Dr. Rainer Spang

Unterschrift:

Franziska Taruttis-Glagoleff

Declaration of personal contribution to published manuscripts

Parts of the thesis are composed of the following published manuscripts:

- Chapter 2 largely corresponds to the manuscript of Taruttis et al. (2017): I developed the method described together with my co-author Maren Feist. I performed the experimental design, generated the synthetic data and analysed the sequence data. I performed the statistical and bioinformatical analysis and prepared the figures except for Figure 2.1, which was kindly provided by Maren Feist.

  Maren Feist (Department of Haematology and Medical Oncology of the University Medical Center Göttingen) carried out all wet lab experiments. RNA-seq was done by Dr. Gabriela Salinas-Riester (Head of Core Microarray and Deep-Sequencing Core Facility, University Medical Center Göttingen)

- Chapter 3 largely corresponds to the manuscript of Taruttis et al. (2015): I developed the computational method described. I performed the experimental design and generated the synthetic data sets. I performed the statistical and bioinformatical analysis and prepared all presented figures.

Acknowledgement

Throughout working at my thesis at the Department of Statistical Bioinformatics at the Institute of Functional Genomics (University of Regensburg) I have received a great deal of support and assistance. The work was supported by the BMBF grant e:Bio:MMML-MYC-SYS and the Bavarian Genome Network BayGene.
I am very grateful to my supervisors Julia C. Engelmann and Rainer Spang for suggesting the topic, for their support and advice during the entire doctorate, for pushing me to sharpen my thinking and to write this thesis under their guidance. Especially, I am thankful to Rainer Spang for providing a friendly, motivating and inspiring place at the Department of Statistical Bioinformatics.
I would like to acknowledge my mentoring team, Tobias Müller and Wolfram Gronwald, for fruitful discussions and their helpful advice.
This work would not have been possible without collaboration. Thus, I want to thank Maren Feist and Dieter Kube, Philipp Schwarzfischer, Wolfram Gronwald and Katja Dettmer-Wilde and Karsten Kleo, Lora Dimitrova and Michael Hummel.
I want to thank Julia C. Engelmann and Katharina Hirsch for proofreading this thesis. Your advice brought my work to a higher level.
Many thanks to our secretaries Eva Engel and Sharon Petersen, who were always ready to lend an ear and for their help with both minor and major problems. I also want to thank Kinga Ay from the Regensburg International Graduate School of Life Sciences for her support during thesis submission.
During the time I worked on this thesis, I enjoyed fruitful discussions, about our work and about everything else with the whole group and the alumni of the Department of Statistical Bioinformatics. Special thanks got to Katharina Hirsch, Claudio Lottaz, Martin Pirkl, Anton Moll, Christian Hundsrucker, Chistian Kohler, Thorsten Rehberg and Paula Perez-Rubio.
Furthermore, I like to thank the "lunch club" of the Institute of Functional Genomics for encouraging me whenever it was needed. In particular, I would like to mention Nadine Aßmann, Claudia Samol, Nadine Nürnberger, Elke Perthen, Katja Dettmer-Wilde and Wolfram Gronwald.
Last but not least, I would like to thank my whole family for their continuing support. Especially, I am grateful to my husband Nico and my beloved children Kuno and Klara. I could not have completed this dissertation without their patience, their understanding and their constant encouragement.

Summary

The oncogene *MYC* plays an important role in B cell lymphoma pathogenesis. Despite more than 30 years of *MYC* research there are still open questions concerning its function and how to target *MYC* in lymphomagenesis. Thus, this work aims to examine the causal relationships between *MYC* and the transcriptome and metabolome in a B cell lymphoma cell line by computational methods. The data set covers RNA-seq and mass spectrometry measurements of the same cell line. The underlying data is purely observational, no intervention is needed since causal inference techniques enable virtual experiments in theory. The first part of this thesis addresses three issues:

First, the analysis of the RNA-seq data from cells with overexpressed *MYC* is challenging since *MYC* is a transcriptional amplifier. There is no de novo activation of genes by the elevated *MYC*, but an amplification of all presently expressed genes. This behavior is accompanied with an increase in cell size and an increase of RNA amount. Thus, the comparison of lymphoma cells with a high *MYC* expression with normal B cells by RNA-seq standard pipelines is difficult, since current normalization methods require a constant RNA amount across samples. I present a method that uses *Drosophila melanogaster* cells as a spike-in to calibrate the data to the number of cells in the sample (Taruttis et al., 2017). I demonstrate that, in case of transcriptional amplification in the B cell lymphoma cell line the use of an external spike-in is mandatory to observe the global gene expression changes. Furthermore, the *Drosophila melanogaster* spike-in normalization outperforms other calibration methods, including the use of the commercially available ERCC spike-ins.

Second, Maathuis et al. (2010) presented the first high throughput analysis of virtual intervention experiments. Their ground-breaking IDA method (Maathuis et al., 2009) will have a lasting effect on the field of systems biology. Further developments of the IDA method recommended a subsampling strategy for the estimation of causal effects from observational data (Stekhoven et al., 2012). I extend IDA and its extension CStaR by analyzing the distribution of causal effects and call the method Accumulation IDA (aIDA) (Taruttis et al., 2015). aIDA improves the prediction of causal effects in comparison to Maathuis et al. (2009) and (Stekhoven et al., 2012).

Third, causal structure learning by the PC algorithm (Spirtes and Glymour, 1991; Kalisch and Bühlmann, 2007), the first step of both IDA and aIDA, assumes that the underlying structure is sparse. However, the application of the spike-in methods to B cell lymphoma data sets with *MYC* overexpression results in highly correlated data. Thus, the underlying causal structure is very likely not sparse. I assume that this is a consequence of the global role of Myc in gene expression (Lin et al., 2012; Nie et al., 2012). Thus, we

observe no technical artifact but a real biological process. I show that using the MMHC algorithm instead of the PC algorithm together with my accumulation method outperforms aIDA for highly correlated datasets. However, the MMHC-aIDA method breaks down, too, when the density of the underlying causal structure becomes too high.
The second part of the thesis presents a causal inference analysis of a B cell lymphoma cell line. We decided for the P493-6 cell line due to its doxycycline-dependant promoter to switch *MYC* on or off, which allows for an examination of the causal relationships of *MYC* under the same epigenetic conditions. RNA-seq and mass spectrometry data are measurements of the transcriptome and the metabolome of the cell line and are the input of the causal inference analysis. I show that the selection of the method to estimate the causal effects highly depends on the data structure. While the highly correlated RNA-seq dataset shows the best results with the MMHC-aIDA method, the mass spectrometry data performs well with aIDA. The analysis of RNA-seq data shows that *MYC* upregulates the majority of genes in the dataset. *MYC* further shows a positive causal effect on the majority of the metabolites. These findings are in line with the hypothesis that *MYC* is a trancriptional amplifier. Some of the causal effects of *MYC* on the transcriptome and metabolome are already known, others can be high priority candidates for future wet lab experiments.

Contents

CHAPTER 1

Introduction

1.1 The oncogene MYC and its role in B cell lymphoma pathogenesis

1.1.1 B cell lymphoma - a very heterogeneous group of cancer

Cancer is a disease characterized by abnormal cell growth and the ability to spread (metastasis) or to invade other neighboring tissues, which may occur in every tissue of the body. The World Cancer Report documents approximately 18.1 million new cases and 9.6 million cancer related deaths in 2018 (Wild et al., 2020). Moreover, the World Health Organization (WHO) assumes that the number of new cases increases by approximately 50 % within the next 20 years. From a genetic point of view cancer is caused by genomic alterations. Lengauer et al. (1998) organizes the various amount of these alterations into four groups: (i) the insertion or deletion of only a small number of nucleotides, (ii) a gene amplification (increase of the number of copies of a gene), (iii) the chromosomal translocation, which is a rearrangement of parts of two different chromosomes and (iv) an alteration of chromosome number by gain or loss of whole chromosomes (aneuploidy). There are more than 100 different types of cancer due to the high number of different tissues that might be affected and the complexity of genetic arrangement. Ciriello et al. (2013) summarizes the analysis of genomic alterations in the following two statements: (i) tumors occurring in the same tissue can be highly diverse with respect to their genomic profile (Network et al., 2012), (ii) whereas tumors from very different tissues can show similar genetic patterns (Verhaak et al., 2010). Hanahan and Weinberg (2011) reduced

the complexity of cancer to eight hallmarks, which provide a logical framework how cancer emerges and develops:

1. Cancer cells stimulate their own growth.
2. Cancer cells disturb or switch off programs in the cell, that control proliferation.
3. Cancer cells avoid apoptosis.
4. Cancer cells have unlimited replicative potential.
5. Tumors require the induction of angiogenesis.
6. Cancer cells are able to invade neighboring tissues and to metastasize.
7. Cancer cells undergo a reprogramming of cellular energy metabolism.
8. Cancer cells defend themselves against attack by immune cells.

Lymphoma is a group of cancer, which affects the human cells that defend the body against pathogens and infected cells, the lymphocytes. In 2018, 589.580 new cases of lymphoma and about 274.891 lymphoma related deaths have been reported worldwide (Bray et al., 2018). There are two groups: Hodgkin and Non-Hodgkin Lymphoma. Hodgkin Lymphoma are characterized by the existence of Reed-Sternberg cells, a distinctive cell type used for diagnostic purposes. While the five year survival rate of Hodgkin lymphoma is about 90 %, the survival rate of Non-Hodgkin Lymphoma highly depends on the subtype. There are more than 60 different subtypes of Non-Hodgkin Lymphoma.
About 90 % of Non-Hodgkin Lymphomas originate from a B cell, a type of white blood cells, that plays a central role in the adaptive immune system. This specific immune response eradicates invading pathogens, called antigens, without damaging any endogenous molecules. These antigens are bound by the B cell receptor (BCR), which B cells express on their cell membrane. A BCR consists of two heavy (H) polypeptide chains and two light (L) polypeptide chains of immunoglobulines (Ig) covalently connected by a disulfit bridge and CD79A and CD79B molecules, that transmit signals after BCR cross linking (Küppers et al., 1999a). Each B cell produces exactly one highly specific kind of immunoglobulines. Specific genetic processes during B cell development, that allow the H and L chains for rearrangement and point mutations enable the B cells to build a wide variety of more than 10^{12} different antibodies (Mårtensson et al., 2010). However, while these mechanisms are responsible for the high flexibility and specificity of the human adaptive immune response, they also carry a great danger. The need of these genetic rearrangements includes the high potential to end up in unwanted genetic alterations and

with that may trigger lymphomagenesis (Basso and Dalla-Favera, 2015). Chromosomal translocations, which do often involve an Ig gene and an oncogene, are a hallmark of B cell lymphoma pathogenesis (Küppers and Dalla-Favera, 2001; Willis and Dyer, 2000). The oncogene comes under the control of an Ig promoter, which leads to an uncontrolled upregulation of the oncogene and with that triggers lymphoma pathogenesis. When a lymphoma is developed at a certain state of the B cell development process, the genetic state of the cell of origin is "frozen" and this maturation state defines the lymphoma subtype (Küppers et al., 1999b; Shaffer et al., 2002).
Thus, to understand the huge diversity of B cell lymphomas we need to get insights into the B cell developmental process.

How does a B cell develop, which processes cause the genetic and functional diversity of the antibodies and how do errors within these processes lead to lymphoma? There are three specific breakpoints of chromosomal translocations which can be assigned to three different stages of the developmental process (Küppers, 2005; Robbiani and Nussenzweig, 2013):

(i) V(D)J recombination The B cell development starts in the bone marrow, where a lymphoid progenitor cell evolves into a pre B cell that expresses a pre-BCR. This pre-BCR consists of two Ig heavy (H) and surrogate light (SL) polypeptide chains. The H and L chains consist of variable (V) and constant (C) regions. According to their constant region the heavy chains are divided into five classes: the $\alpha, \delta, \epsilon, \gamma$ and μ chains. And depending on that grouping the antibodies are classified as IgA, IgD, IgE, IgG and IgM. The variable regions are completed by a Joining (J) and Diversity (D) region.
To build a functional antibody, the pre-BCR undergoes a process called V(D)J recombination of the μ-chains which rearranges these V, J and D gene segments. This rearrangement enables the B cells to produce very diverse antibodies, which are important for the adaptive immune response to new pathogens. Even if this process is well controlled by different mechanisms (Jung et al., 2006), errors during V(D)J recombination process can drive lymphomagenesis by introducing unwanted chromosomal translocations. A well-examined example is the BCL2-IgH translocation in follicular lymphoma. The *BCL2* gene is located on chromosome 18 and encodes for the BCL2 family of regulatory proteins (Adams and Cory, 2007). The translocation brings the BCL2 gene under the control of a Ig heavy chain promotor (Jäger et al., 2000; Tsujimoto et al., 1985, 1988). This leads to a dramatically increase of BCL2 expression and an increased BCL2 expression increases the probability of a cell to avoid apoptosis (Cleary et al., 1986; Tsujimoto

et al., 1984), which is a hallmark of carcinogenesis (Hanahan and Weinberg, 2011).

(ii) Somatic hypermutation After V(D)J recombination the surrogate L chains are replaced by corresponding valid L chains to form a IgM antibody. Thus, the IgM antibody is the first antibody which is produced during B cell development. The affinity of these IgM antibodies to bind their corresponding antigen is still low in comparison to the other classes. To account for that, this class of antibodies is able to form pentamers when it is secreted and with that the antigen is tackled by 10 antigen binding sites. If this pre-BCR is functional and non-autoreactive, the immature naive B cell leaves the bone marrow, starts to express IgD antibodies on its membrane and moves to the secondary lymphoid organs, for example the spleen, the lymph nodes or the tonsils. Immature B cells which do not pass this checkpoint undergo apoptosis (Rajewsky, 1996). After activation by an antigen and mediated by T cells the naive B cells proliferate intensively and form the germinal center. During its maturatin the germinal centers form two separate compartments: the dark zone and the light zone. In the dark zoe of the germinal centers the B cells undergo a process to remodel their immunglobulin genes and to express high-affinity antibodies: the somatic hypermutation. The B cells are extremely proliferative and undergo point mutations, gene amplifications and deletions in the V regions of the BCR. This ensures a huge variety of different antibodies with different affinity to the antigen. However, Pasqualucci et al. (2001) demonstrated, that hypermutated regions of some protooncogenes are prone to chromosomal translocations in Diffuse Large B Cell Lymphomas (DLBCLs). Thus, errors during somatic hypermutation may introduce unwanted chromosomal translocations.

(iii) Class switch recombination The B cells with the BCR altered by somatic hypermutation leave the dark zone of the germinal centers and enter the light zone. Here the cells are selected by T helper cells and follicular dendritic cells for an improved antigen binding. B cells which do not improve their affinity to the antigen by somatic hypermutation undergo apoptosis. Some B cells reenter the dark zone to further improve their affinity (De Silva and Klein, 2015). Before they leave the light zone these B cells can do a class switch recombination, where the Ig gene of the constant region is changed to build IgG, IgE and IgA antibodies. This results in the same antigen-binding domain, but a different class of antibodies (Küppers, 2005). The five different classes of antibodies can attack the antigens in different ways. They occur in different proportions and the IgG is the largest group of antibodies. However, class switch recombination is the third process which may generate chromosomal translocations. A well-described example is the t(8;14) mutation, which brings the *MYC* oncogene under the control of an Ig gene.

This translocation is a primary event during sporadic Burkitt's lymphoma pathogenesis (Taub et al., 1982; Janz et al., 2003; Küppers and Dalla-Favera, 2001; Dalla-Favera et al., 1983; Ramiro et al., 2006). Finally, the high-affinity B cells mature to effector B cells, which secrete huge amounts of antibodies to neutralize and destroy antigens and memory B cells, that trigger a more effective immune response in case of re-infection with the antigen.

Beyond the chromosomal translocations mentioned before, there are many other different kinds of mutations, like deletions, insertions or other genetic alterations that may trigger lymphomagenesis (Seifert et al., 2013). Each subtype of B cell lymphoma can be assigned to a specific maturation state based on mutations in the Ig V region and gene expression profiling. This assigned maturation state is termed the "cell of origin" (Shaffer III et al., 2012). For example, Burkitt's lymphoma share a similar genetic profile with the B cells in the dark zone of the germinal centers, while the DLBCLs share similarities with the GC light zone B cells (Küppers, 2005). However, Shaffer III et al. (2012) criticize the term "cell of origin", since pathogenesis may also start at an earlier state of the maturation process. Thus, after the initial oncogenic event the maturation of the B cell continues until secondary oncogenic hits determine the lymphoma subtype (Shaffer III et al., 2012).

1.1.2 The MYC oncogene is a hallmark of B cell lymphoma pathogenesis

MYC is located on chromosome 8 and plays a central role in B cell lymphomagenesis. It encodes for transcription factor C-MYC and influences metabolism, proliferation and cell differentiation. Together with MAX, MYC builds a complex, which binds DNA at the E-box promotor region (Kretzner et al., 1992; Amati et al., 1993). *MYC*'s oncogenic potential appears, when it comes under the control of an Ig gene by chromosomal translocation. The resulting upregulation of *MYC* can induce aggressive B cell lymphomas (Dalla-Favera et al., 1982; Persson and Leder, 1984; Adams et al., 1985). Upregulated MYC is detectable in 30% to 40% of DLBCLs and 70% to 100% of Burkitt lymphomas (Sesques and Johnson, 2017; Johnson et al., 2012; Chisholm et al., 2015; Agarwal et al., 2015; Perry et al., 2013). MYC plays an important role in every hallmark of B cell lymphoma pathogenesis, which is underpinned by the following examples:

(i) Cancer cells stimulate their own growth. Even if proliferative signaling in cancer cells is well understood, many details especially with respect to therapeutic approaches requires further research. Hanahan and Weinberg (2011) present three ways how

cancer cells maintain the proliferative signaling. First, they produce ligands for growth factors themselves. Second they send signals to normal cells in their environment, which causes these normal cells to produce growth factors for the cancer cells. And third, they increase the amount of growth factor receptors on their surface to become hyperresponsive (Hanahan and Weinberg, 2011).
A major growth factor signaling pathway in humans is the PI3K-AKT-mTOR pathway and the deregulation of this pathway may lead to cancer. In a nutshell, activated Phosphoinositid-3-Kinase (PI3K) causes cell growth via Proteinkinase B (AKT), increases metabolic activity and promotes survival (Cantley, 2002). The mechanistic Target of Rapamycin (MTOR), a downstream effector of AKT consists of two different complexes. The mTOR complex 1 (MTORC1) plays a major role in cell growth (Ma and Blenis, 2009) and the mTOR complex 2 (MTORC2) inhibits a MTORC1 repressor, the tuberous sclerosis complex 2 (TSC2) via AKT. MTOR directly activates protein synthesis, lipid synthesis and ribosomal biogenesis (Efeyan et al., 2012).
Deregulated MYC in Burkitt's lymphoma increases the expression of MIR17HG a precursor RNA for the micro-RNA miR-19 (Olive et al., 2009; Xiao et al., 2008). miR-19 is an inhibitor of Phosphatase and Tensin homolog (*PTEN*), a well-known tumor suppressor gene which downregulates PI3K. Thus, the inhibition of *PTEN* increases the expression of PI3K, which drives the survival of Burkitt's lymphoma cells (Schmitz et al., 2012).

(ii) Cancer cells disturb or switch off programs in the cell that control proliferation. Differentiated, mature cells have programs to inhibit cell growth and proliferation. A lot of these programs involve tumor suppressor genes. A typical tumor suppressor gene is the *TP53* gene which is regulated by *MDM2*. During homeostasis, without stress exposure *MDM2* ubiquitinates and with that downregulates TP53. After stress exposure, for example by oxidative stress or DNA damage, *TP53* is activated. When a cell is damaged *TP53* triggers cell cycle arrest and, in case of irreparable cell damage, activates apoptosis. In spite of the huge efforts during the past three decades the detailed regulation mechanism is still unknown (Aubrey et al., 2016). *MYC* has a contradictory function. On the one hand *MYC* increases proliferation, while on the other hand *MYC* also drives apoptosis via TP53 (Hermeking et al., 1994). This process acts as a failsafe programm to circumvent uncontrolled proliferation. However, in lymphoma cells this second function is most often inactive. For example, about 30%-60% of Burkitt's lymphoma carry a *TP53* mutation (Gaidano et al., 1991; Newcomb, 1995; Preudhomme et al., 1995). In transgenic mice with a translocation similar to the translocation in the majority of Burkitt's lymphoma, Schmitt et al. (2002) observed that these phenotypes are chemoresistant and develop fast to the terminal stage. In summary, the loss of TP53

and the activation of *MYC* are an example for mutations that cooperate to increase proliferation and survival.

(iii) Cancer cells avoid apoptosis. Normal cells share the ability to induce a programmed cell death to preserve the homeostatic balance of their tissues. Whenever a cell is damaged and not able to fulfill its normal tasks anymore the program is activated, too. BCL2 plays an important role in the regulation of programmed cell death, since it consists of pro- and antiapoptotic units and primarily promotes cell survival (Vaux et al., 1988). Usually, in absence of *BCL2*, *MYC* induces apoptosis via BIM (encoded by *BCL2L11*). However, in more than 50% of DLBCLs *BCL2* is upregulated and inhibits *MYC* mediated apoptosis with no effect on the proliferative function (Hoffman and Liebermann, 1998). Moreover *BCL2* mediates chemoresistance (Schmitt and Lowe, 2001). While Reed et al. (1988) demonstrated the oncogenic potential of *BCL2*, additional genomic alterations are needed to develop an aggressive lymphoma and a chromosomal translocation of *MYC* could take on this role (McDonnell et al., 1989). Neither a mutation of *MYC* nor a mutation of *BCL2* alone causes aggressive lymphomas. However, the combination of *BCL2* mediated survival signals and proliferation signals provided by *MYC*, also known as a type of Double-hit lymphomas results in low survival rates and a bad prognosis (Johnson et al., 2009).

(iv) Cancer cells have unlimited replicative potential. The ends of the chromosomes consist of a special DNA structure, called telomers, which prevent the chromosome ends from DNA damage (Muller, 1938; McClintock, 1941). The length of the telomers is related to the life span of the cell since they shorten with increasing age (Harley et al., 1990). Further telomere shortening directly causes cells to pass into a state called senescence (Nelson and Kastan, 1994; Vaziri et al., 1997), characterized by viable cells which cannot divide anymore (Sedivy, 1998). The enzyme complex telomerase which adds telomeric repeats to the chromosome ends (Greider and Blackburn, 1987), hampers this process. Kim et al. (1994) summarize that telomerase is not present in the majority of normal mortal cell cultures, whereas it is present in the majority of immortal and cancer cells. 85 % of human cancers upregulate telomerase to maintain their telomere length and with that avoid to pass to senescence and crisis, while the remaining cancer types use other processes to avoid telomere shortening (Cesare and Reddel, 2010). The telomerase enzyme complex consists of the RNA part called telomerase RNA component (*TERC*) and the catalytic subunit telomerase reverse transcriptase (TERT). Different studies suggest that *MYC* acitvates TERT directly (Greenberg et al., 1999; Wang et al., 1998; Wu et al., 1999). Thus, *MYC* has a direct influence on keeping limitless replicative

potential. Further, inversely Koh et al. (2015) found, that high levels of TERT positively influence MYC stability and with that regulate lymphomagenesis.

(v) Malignant tumors induce angiogenesis Angiogenesis is the ability to form new blood vessels from preexisting ones. This property is essential for the survival of the tumors, since they have to ensure supply of oxygen and nutritients and they have to maintain the possibility to get rid of carbon dioxide and metabolic waste (Hanahan and Weinberg, 2011). *MYC* triggers the expression of the proangiogenetic factor *VEGF* and represses thrombospondin, an antiangiogenetic factor (Baudino et al., 2002). Further in lymphoma that carry a translocation between *MYC* and an Ig gene (*MYC* positive), the repression of *MYC* puts a stop to angiogenesis and thrombospondin expression which is essential for tumor regression. Thus, *MYC* induces angiogenesis in lymphoma and plays an important role in tumor maintenance (von Eyss and Eilers, 2011).

(vi) Cancer cells are able to invade neighboring tissues and to metastasize Cancer cells may on the one hand penetrate adjacent tissues, but, on the other hand, spread to organs different and distant from the primary tumor and form secondary tumors, which is known as metastasis. The ability of cancer cells to invade neighboring tissues and to metastasize causes approximately 90% of cancer related deaths (Canel et al., 2013; Li and Li, 2014; Mehlen and Puisieux, 2006). Metastatic spread is a complex molecular process known as the invasion-metastasis cascade (Fidler, 2003; Gupta and Massagué, 2006; Talmadge and Fidler, 2010). Following Leber and Efferth (2009) the five steps of this cascade include: (i) invasion and migration of individual cells from the primary tumor to neighboring, healthy tissues, (ii) the invasion of blood and lymphatic vessels by the cancer cells, (iii) the circulation of cancer cells via the blood and the ability of these specialized, aggressive cells to tolerate the high concentration of oxygen and cancer-cell-toxic lymphocytes, (iv) the potential of the cancer cells to leave the blood stream again, and (v) the colonization of distant organs together with proliferation and angiogenesis. E-cadherin, encoded by *CDH1* (Huntsman and Caldas, 1998), mediates cell-cell adhesions and the loss of E-cadherin is associated with increased cell motility, tumor progression and metastasis (Birchmeier and Behrens, 1994; Canel et al., 2013). Ma et al. (2010) have shown that MYC directly activates *MIR-9*, a miRNA, which represses the expression of E-cadherin. These findings suggest that *MYC* also influences invasion and metastasis. Roehle et al. (2008) found that *MIR-9* is upregulated in follicular lymphoma. Further, Di Lisio et al. (2012) found an upregulation of *MIR-9* in DLBCLs and Marginal zone B cell lymphoma and Szenthe et al. (2013) point out, that *MYC* activates miR-9 and miR-9 targets E-cadherin. These findings highlight the importance of miR-9

for lymphomagenesis.

(vii) Cancer cells undergo a reprogramming of cellular energy metabolism Normal healthy mammalian cells primarily convert glucose to pyruvat via glycolysis in the presence of free oxygen to generate energy. Pyruvat is further processed to acetyl-CoA, a key molecule in protein and lipid metabolism, by oxidation in the mitochondria (Oxidative phosphorylation). In contrast, if oxygen is absent, normal cells transform pyruvat to lactate (anaerobic glycolysis). Both metabolic pathways produce adenosine triphospahte (ATP), a molecule used for intracellular energy transfer. However, while oxidative phosphorylation produces 36 mol ATP per mol Glucose, the anaerobic glycolysis supplies 2 mol ATP per mol Glucose. Warburg (1956) observed that cancer cells, even under aerobic conditions, increase lactate production. This process is known as Warburg effect or aerobic glycolysis. The benefits of the Warburg effect for the cancer cells and whether its a cause or a consequence of cancer development is not fully resolved, yet (Liberti and Locasale, 2016; Devic, 2016). However, the findings of Le et al. (2012) suggest that the overexpression of *MYC* increases the glycolytic flux in *MYC*-driven Burkitt lymphoma. Beyond these findings *MYC* also influences glutamine metabolism, fatty acid and cholesterol metabolism and regulates ribosome biogenesis, protein and nucleotide biosynthesis (Hsieh et al., 2015). Thus, *MYC* is a key manipulator of cellular energy metabolism.

(viii) Cancer cells defend themselves against attack by immune cells The immune system prevents cancer development at three levels: (i) it impedes the outbreak of viral infections and with that lowers the probability for virus-induced cancer development, (ii) it neutralizes pathogens fast and prevents chronic inflammation, which might lead to DNA damage increasing the risk to develop cancer, and (iii) cells of the immune system are able to detect cancer cells and to remove them (Swann and Smyth, 2007). Recently, Casey et al. (2016) showed that *MYC* hampers this detection of tumor cells by inducing the expression of *CD47* and *PD-L1*. While *PD-L1* sends a "'don't find me"' signal mainly induced by IFN-γ, a pro-inflammatory molecule (Casey et al., 2016; He et al., 2015), *CD47* sends a "'don't eat me"'-signal to macrophages and dendritic cells (Casey et al., 2016; Spranger et al., 2016). Thus, *MYC* influences both, the adaptive immune system via *PD-L1* and the innate immune system via *CD47*.

These are just a few examples for the various functions of *MYC* during tumorigenesis. *MYC* is deregulated in many different ways. In addition to gross genetic events like insertional mutagenesis, amplification and chromosomal translocation (Meyer and Penn,

2008), *MYC* is also activated by many signaling pathways, e.g. BCR- or NFκB-signaling, without structurally affecting the location of the *MYC* gene (Sesques and Johnson, 2017). A deregulation of *MYC* can be a primary event for example in Burkitt's lymphoma or a secondary event for example in *MYC* positive DLBCLs, which is often associated with a poor prognosis (Johnson et al., 2009). However, deregulated *MYC* alone is not sufficient for lymphomagenesis (Sander et al., 2012). Although upregulated *MYC* induces proliferation in BL, *MYC* also triggers apoptosis by the activation of the p53 pathway and by upregulating the proapoptotic gene *BIM* (Evan et al., 1992; Hoffman and Liebermann, 2008; Jacobs et al., 1999; Schmitt et al., 1999). More than inducing lymphomagenesis, *MYC* cooperates with other mutations, as described above. But, the suppression of *MYC* can reverse lymphomagenesis (Marinkovic et al., 2004). However this reversion does not work in every context. Gabay et al. (2014) summarizes that for example a knock out of *TP53* expression (Giuriato et al., 2006) or a *RAS* activation (D'Cruz et al., 2001) hampers the reversion of tumorgenesis. Thus, the physiological state of the cell influences the response of the cells to *MYC* upregulation (Gabay et al., 2014).

1.1.3 MYC - a transcriptional amplifier and resulting consequences for data normalization

The work of Lin et al. (2012) and Nie et al. (2012) suggests that *MYC* is a transcriptional amplifier since they found that *MYC* leads to the induction of virtually all transcribed genes by a factor of two to three, accompanied with an increase in cell size. Roughly spoken, there is no de novo activation of genes by *MYC*, since *MYC* just amplifies the expression of the presently expressed genes depending on the epigenetic background. This behaviour directly influences the RNA amount in a cell. We can observe a much higher amount of RNA in cells with deregulated *MYC* expression (Figure 1.1). Thus, the RNA amount between cells with deregulated *MYC* and normal cells differs dramatically. This observation is crucial for data normalization (Lovén et al., 2012), since current normalization methods require the total amount of RNA between different conditions to be constant (see (Li et al., 2017) for a summary of current normalization methods).
These normalization methods are based on the assumption that only a small number of genes change their expression values, while the majority of genes stays constant. Thus, the global upregulation of the majority of presently expressed genes induced by *MYC* is a violation of this main assumption.
The effect of amplification on data normalization cannot be solved by a computational method alone, since the crucial step happens in the wet lab. Whenever a fixed amount of RNA is taken from our sample for further analysis, the information about the total

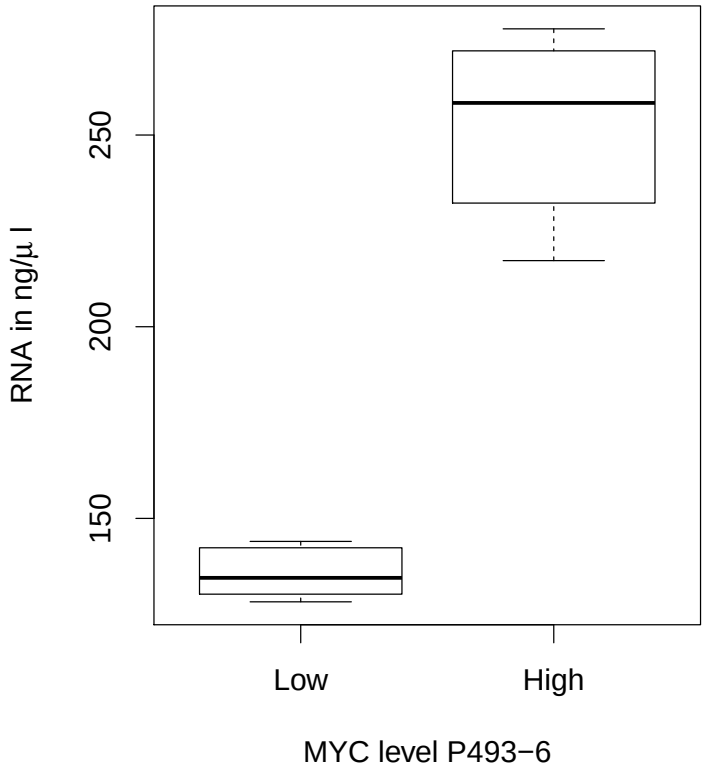


Figure 1.1: RNA concentration of P493-6 cells for two levels of *MYC*. The P493-6 cell line allows for the induction of *MYC* via a doxycycline depended promoter to switch *MYC* on or off. "'Low"' indicates the cells in the presence of doxycycline and represents a low level of *MYC* expression, while "'High"' indicates the cells in absence of doxycycline and a high expression level of *MYC*. The induction of *MYC* in one million P493-6 cells leads to an increase of the RNA amount per cell(n=10).

amount of RNA per cell is lost. We cannot decide whether there was a real increase of the amount of RNA between two conditions or we just observe technical differences in processing the samples, where we could simply account for. A solution is to measure the amount of RNA before taking a fixed-sized aliquot from the probe (Aanes et al., 2014). However this method is very inaccurate due to the different proportions of different specimen of RNA within a sample. Moreover, there is a large spread in measurements between different measurement set-ups, e.g. different methods, different technical equipment, but also different operator time and skills (Aranda et al., 2009). By adding an external standard to the sample, we can overcome this problem. Lovén et al. (2012) used the commercially available External RNA control consortium (ERCC) spike in kit, which consists of 92 poly-adenylated transcripts derived from *Bacillus subtilis* with lengths between 250 and 2000 nucleotides. Adding a fixed amount of these transcripts to the total amount of RNA enables us to keep the information of RNA amount during the following

experimental steps. Roughly spoken, if we take a fixed amount of RNA from our sample after adding the ERCC spike in kit, we are able to reconstruct the initial quantity by rescaling the sample with the ERCC transcripts.
If the number of cells in the sample is known, this method offers a new fix point for data normalization. Now, we are able to rescale the gene expression levels to the total number of cells instead of a fixed sized aliquot. This new method will open new insights, especially for comparisons between samples where the cells under different conditions contain different amounts of RNA. However, since the ERCC spike-ins are added after cell lysis, the method cannot account for the difference in RNA extraction. The work of Buettner et al. (2015) suggests that also different stages in the cell cycle may lead to a difference in RNA content. This indicates that differences in RNA amount are not a *MYC*-specific, but rather a general phenomenon and a problem for the common normalization methods.

1.2 On the estimation of causal effects from observational data

1.2.1 Correlation does not imply causation

In science we distinguish between two kinds of data, observational (non-experimental) data and interventional (experimental) data. Interventional data is data achieved from a randomized experiment. In the simplest possible scenario the samples are (randomly) divided into a treatment and a control group and we are able to observe the causal effect of the treatment on our variable(s) of interest. Assume we want to examine how *MYC* affects other genes in a certain cell line. In a randomized experiment we measure the transcriptome of 10 samples where we knocked down *MYC* by the inhibitor 10058-F4 and 10 untreated control samples. We compare the results of knock-down and control samples to directly estimate fold changes. And with that we know which genes are mainly affected by the *MYC* knock down. In contrast to that, in an observational study the treatment is not randomly assigned to the samples or there is no treatment of the samples at all. We want to draw conclusions from a pure observation. As a counterpart to the experiment described above, we just observe 10 gene expression profiles of the cell line and we want to learn something about the causal effect of *MYC* on the other genes. What is the difficulty of that question?
Assume you want to find out whether there is a causal relationship between smoking and lung cancer of 30 year old women over 10 years. For an interventional study you have to find e.g. 1000 30 year old women, who never had smoked before, assign randomly 500 of them to smoke 20 cigarettes per day for the next ten years and assign 500 of them

to stay smoke-free in the same period. During this decade you register all cases of lung cancer. Now, we are able to estimate the causal effect of smoking on getting lung cancer within ten years for 30 year old women. Of course such a study is unethical and not legally warranted. How would an observational study look like? We could select 500 non smoking 30 years old women and 500 which smoked approximately 20 cigarettes per day and register all cases of lung cancer for the next ten years. This is a feasible study, which is compatible with basic moral principles and maybe we find a high correlation between smoking and having lung cancer. However, the main problem is that (without taking further assumptions (Peters et al., 2014)) there is no way to decide whether an observed association between 2 variables is causal or not. Thus, the estimation of causal effects from observational data is a very hard task.

A common example to illustrate the problem is the relationship between eating ice cream in summer and drowning deaths. Plotting the number of people who drowned against ice cream sales over the period of one or several years, we might infer that eating ice cream causes drowning. But if we have a closer look to our example we will find that we did not take the season and the air temperature into account. Assume that "I" is the variable for the ice cream sales, "D" represents the variable for the number of drowning deaths and "S" describes the season or air temperature. Although we find a correlation between "I" and "D", we find that "I" and "D" are independent given "S" (see Figure 1.2). This illustrates one of the main pitfalls in causal inference: the third common cause.

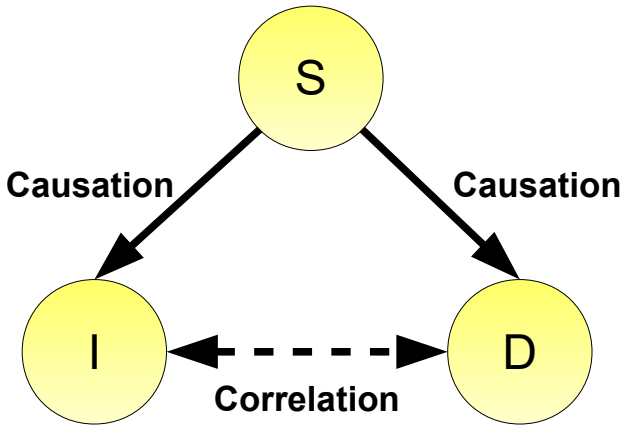


Figure 1.2: The third common cause. The observed correlation between I and D is not due to a direct causal connection between I and D, but due to a third common cause S, which has a causal effect on both I and D.

Another common problem is to infer the direction of causal effects between two variables. Assume there is an alien which just arrived on earth for research and this alien is attracted

by fire. Every time it observes a fire it also observes firemen. Since the alien does not know about the function of a fire brigade it might assume that these firemen are the cause of these fires, which is obviously not the case.
In summary a correlation between two variables X and Y may imply that

- X causes Y or
- Y causes X or
- there is a third common factor, which causes both X and Y or
- X and Y cause each other in a feedback loop or
- we observe a pure coincidence.

What does that mean for our observational study to detect the causal relationship between smoking and lung cancer of 30 year old women over 10 years? Imagine there is a gene which makes smoking highly attractive for some women and now imagine that the same gene directly causes lung cancer. Then we have an unobserved third common cause in our observational study and would maybe draw a completely different conclusion, if we include this cause into our study. In the early 1950ties Professor Bradford Hill and Dr. Richard Doll came up with some studies which described a connection between smoking and developing lung cancer (Doll and Hill, 1950, 1952, 1954, 1956). Their observational studies showed that there are more smokers among the patients with lung cancer than among other patients. Thus, they concluded that smoking plays an important role in developing lung cancer. However, the famous statistician Ronald Fisher strongly criticized these studies, since correlation never implies causation and there might be other reasons like a third common cause, which might explain the observed correlation (Fisher, 1958a,b,d,c). And indeed, even if there are chemical mechanisms known today, which link smoking and the risk of developing lung cancer (Hecht, 2002), there are also studies which identified a genetic disposition for developing lung cancer (Amos et al., 2008; Hung et al., 2008; Thorgeirsson et al., 2008). However, today there is a general consensus among scientists and physicians, that smoking is a major risk factor for smoking (Torre et al., 2016; Cheng et al., 2016; Swanton and Govindan, 2016; Malhotra et al., 2016). And Ronald Fisher not only was a passionate smoker, but also worked as a consultant for the tobacco industry. But he showed, if causal statements are derived from pure correlation, this offers weak points of the analysis.
Obviously, it is a hard task to uncover causal connections, but sometimes interventional studies cannot be conducted due to ethical aspects even if we know this would be the better choice. In genetic studies there is another reason, why we sometimes have to refer

to an observational instead of an interventional study. Imagine we want to find out the causal effects of 10.000 genes on *MYC*. This would result in 10.000 interventional studies (one for each gene), which is impossible to handle with reasonable effort in reasonable time. An observational study would need much less effort, but the only information we get is a correlation, which is not causally interpretable.
During this introductory section I will present in detail under which assumptions we can infer causal effects from observational data. I summarize how we can estimate at least an incomplete causal structure from observational data and how we can estimate causal effects from that graph without doing any interventional experiments.

1.2.2 Estimation of causal effects from observational data when the causal structure is known

First we need to define a mathematical representation of the causal structure. We assume our causal relationships between genes to be a directed graph consisting of nodes $\mathbf{X} = \mathbf{X_1}, ..., \mathbf{X_p}$ and directed edges $\mathbf{E} = \mathbf{E_1}, ..., \mathbf{E_s}$. The directed graph does not contain cycle, and we call the graph directed acyclic graph (DAG). Every gene is represented by one node in the network. Statistically, every node is a random variable whose values indicate the expression levels of a gene. In this section we assume that the causal structure of our problem is known. We call all nodes with incoming edges to a node X parents of X ($pa(X)$). If a node X has two parents Y and Z, which are not connected to each other, we call this pattern a v-structure and X is called a collider (Pearl, 2009) (Figure 1.3).

Edges represent conditional dependencies; nodes that are not connected represent variables that are conditionally independent. Moreover, we interpret the edges causally. If there is a directed edge $X \rightarrow Y$, we assume that an experimental perturbation of the expression level of X will affect Y, but not vice versa. This causal interpretation is ensured by the Causal Markov condition first described by Kiiveri and Speed (1982). The Causal Markov condition is defined as follows:

Definition 1.2.1 (Causal Markov condition by Spirtes, Glymour and Scheines, section 3.4.1, p. 53 (Spirtes et al., 2000)) *Let G be a causal graph with a set of vertices V and let P be a probability distribution over the vertices in V generated by the causal structure represented in G. G and P satisfy the Causal Markov Condition if and only if for every X in V, X is independent of V \ (Descendants(X) ∪ pa(X)) given pa(X).*

In other words: every observation of a gene X is statistically independent of all other genes which are no descendents of X in the causal graph given the parents of X ($X \perp\!\!\!\perp Y | pa(X)$, with $Y = V \setminus (Descendants(X) \cup pa(X))$).

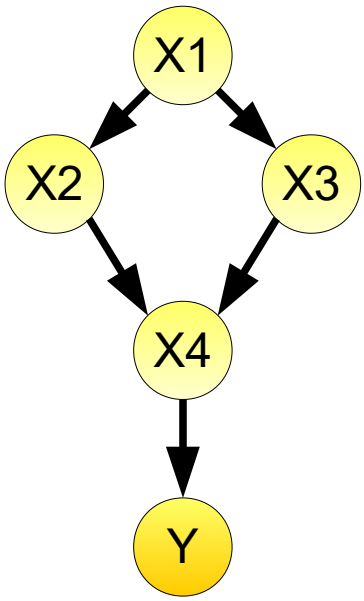


Figure 1.3: Example of a DAG. The DAG consists of 5 nodes and 5 edges. The node X_1 is the parent of X_2 and X_3, X_2 and X_3 are the parents of X_4 and X_4 is the parent of Y. X_2, X_3 and X_4 form a v-structure and X_4 is the collider of this v-structure.

To determine all statistical independencies of a DAG and, thus, to characterize a set of distributions which is compatible with that DAG we can apply the d-separation criterion:

Definition 1.2.2 (d-Separation by Judea Pearl, definition 1.2.3, p. 16, (Pearl, 2009)) *A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if*

i p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or

ii p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set of nodes Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

The converse of the Causal Markov Condition, the causal faithfulness condition ensures that all conditional independencies observed are due to the Causal Markov condition:

Definition 1.2.3 (Causal Faithfulness condition by Spirtes, Glymour and Scheines, section 3.4.3, p. 56 (Spirtes et al., 2000)) *Let G be a causal graph and P a probability distribution generated by G. G and P satisfy the Causal Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov condition applied to G.*

A violation of this property is possible if for example the observations of two different genes are connected by two different paths, which cancel each other out exactly. Thus, we observe an independence (the causal effect is zero) even though the nodes are causally connected.
For now we assume that we know the causal structure. Given this causal structure and the observations of the genes, we want to determine the causal effect of a gene X on another gene Y. More precisely we want to perturb the gene X and we want to observe what happens to gene Y due to this intervention. However, we do not want to do this in the wet lab. We want to perform a purely computational virtual intervention experiment. Goldszmidt and Pearl (1992) introduced a new operator to probability theory to allow for such an external intervention called the do-operator. The do-operator $do(X = x^*)$ sets the variable X to a fixed value x^* and as a result X does not depend on its former parents anymore. But how do we calculate the causal effect of a variable X on a variable Y using this do-operator? Pearl (1993) introduced a graphical criterion called the back-door criterion which provides the link between his do-operator and standard statistical calculus.

Definition 1.2.4 (back-door criterion by Judea Pearl, definition 3.3.1, p. 79, (Pearl, 2009)) *A set of variables Z satisfies the back-door criterion relative to an ordered pair (X, Y) in a DAG G, if*

(i) no node in Z is a descendant of X.

(ii) Z blocks (see def. 1.2.2) every path between X and Y that contains an arrow into X.

Pearls back-door adjustment enables the calculation of the causal effect:

Definition 1.2.5 (back-door adjustment by Judea Pearl, theorem 3.3.2, p. 79, (Pearl, 2009)) *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effects of X on Y is identifiable and is given by the formula $P(Y|X = do(X = x^*)) = \sum_Z P(Y|X, Z)P(Z)$.*

Further, it can be shown, that the true parents of a node X do always fulfill the Back-door criterion relative to every ordered pair (X, Y) (Pearl, 2009). We see

$$P(Y|do(X = x^*)) = \begin{cases} P(Y) & : Y \in pa_X \\ \int P(Y|X = x^*, pa_X)P(pa_X)\,\mathrm{d}pa_X & : Y \notin pa_X. \end{cases} \tag{1.1}$$

Following Rosenbaum and Rubin (1983), Maathuis et al. (2009) defined the difference $E(Y|do(X = x^*)) - E(Y|do(X = x^{**}))$ as "causal effect".

$$\begin{aligned} E(Y|do(X = x^*)) &= \int Y \cdot P(Y|do(X = x^*))\,\mathrm{d}y && :(1.2) \\ &= \begin{cases} \int Y \cdot P(Y)\,\mathrm{d}y & : Y \in pa_X \\ \int\int Y \cdot P(Y|X = x^*, pa_X)P(pa_X)\,\mathrm{d}pa_X\,\mathrm{d}y & : Y \notin pa_X. \end{cases} && (1.3) \\ &= \begin{cases} E(Y) & : Y \in pa_X \\ \int\int Y \cdot P(Y|X = x^*, pa_X)\,\mathrm{d}y P(pa_X)\,\mathrm{d}pa_X & : Y \notin pa_X. \end{cases} && (1.4) \\ &= \begin{cases} E(Y) & : Y \in pa_X \\ \int E(Y|X = x^*, pa_X)P(pa_X)\,\mathrm{d}pa_X & : Y \notin pa_X. \end{cases} && (1.5) \\ &= \begin{cases} E(Y) & : Y \in pa_X \\ \int E(Y|X = x^*, pa_X)P(pa_X)\,\mathrm{d}pa_X & : Y \notin pa_X. \end{cases} && (1.6) \end{aligned}$$

We further assume our variables to be multivariate Gaussian and we conclude

$$E(Y|X = x^*, pa_X) = \beta_0 + \beta_x X + \beta_{pa_X}^T pa_X. \tag{1.7}$$

In summary β_X from

$$Y = \beta_0 + \beta_X X + \beta_{pa_X}^T pa_X + \epsilon : Y \notin pa_X, \tag{1.8}$$

is a consistent estimator of the causal effect of X on Y (Maathuis et al., 2009). If $Y \in pa_X$ the causal effect of X on Y is zero. The parents of X, pa_X, adjust the regression for possible confounders that affect X and Y simultaneously, which could create spurious correlations between X and Y. Note that there might be other sets of variables $\mathbf{Z}$ that also fulfill the back-door criterion and can be used in the regression above instead of the set pa_X.

This basis might help to solve the "smoking - lung cancer problem". The surgeons from Doll and Hill (1950, 1952, 1954, 1956) claim that there is a direct effect between smoking and lung cancer, which is shown by the graph in Figure 1.4 (a), while Fisher would answer that there might be a third common cause, e.g. a genetic factor influencing both, the urge to smoke and a higher probability of getting lung cancer as shown in Figure 1.4 (b). If the genetic factor influencing both could be observed, the causal effect can be calculated by applying the back-door criterion. The back-door criterion aims to close all back-door paths (Figure 1.4 (b), red arrows), while leaving the front-door paths (Figure 1.4 (b), green arrows) open. The causal effect of smoking on getting lung cancer differs depending on the presence or absence of the genetic factor. Closing the back-door paths means to take out the variation due to the genetic factor and to compare people with the same genetic factor to each other which is equivalent with holding the genetic factor constant. Thus, the causal effect observed does not depend on the influence of the genetic factor on smoking anymore, since we control for this variable. However, if the genetic factor is unobserved, the back-door criterion cannot be applied and the problem is more complicated. Pearl (2009) offers another graphical criterion which still helps solving the problem. He introduces another variable "tar deposit in the lung" and applies the front-door criterion to calculate the causal effect.

1.2.3 Causal structure learning

In the previous section we assumed that the causal DAG of the problem is known. A DAG fully specifies the conditional dependencies of all nodes (Pearl, 1988), but not vice versa. Several DAGs with the same skeleton of undirected edges and the same v-structures (Pearl, 2009) encode the same conditional dependency structure (Verma and Pearl, 1990). These equivalence classes of DAGs can be represented by completed partially directed acyclic graphs (CPDAGs) that consist of the joint skeleton and the directed edges which

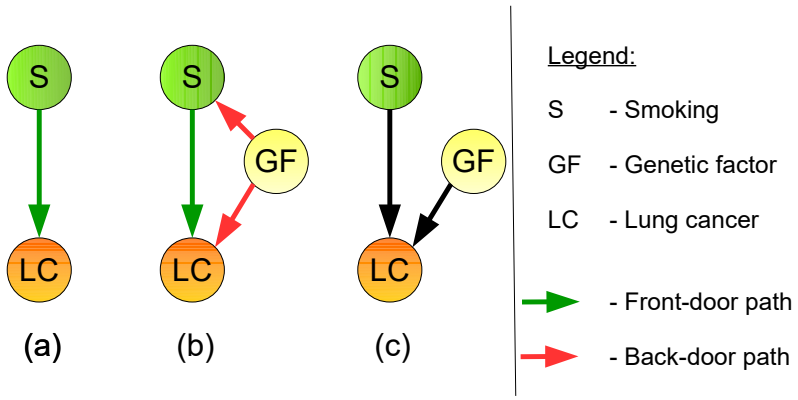


Figure 1.4: The back-door criterion for the "smoking - lung cancer problem". (a) Model which assumes a direct causal effect between smoking and developing lung cancer. (b) Model which assumes a third common cause, a genetic factor, which influences both the urge to smoke and the probability of getting lung cancer. Green arrows show the front-door paths, red arrows show the back-door paths. (c) Application of the back-door criterion to subfigure (b) with the aim to examine whether smoking causes lung cancer.

are common to all DAGs in the equivalence class (Chickering, 2002a). In many scenarios, especially in biology, neither the DAG nor the CPDAG are known. For example, the causal gene regulatory networks which explain the emergence of cancer are widely unknown. Thus, we need to estimate the causal structure from the data we observe. However, causal structure learning is NP hard (Chickering et al., 2004). For a set of n vertices we can find $2^{n(n-1)}$ possible directed networks (Harary and Palmer, 1973), e.g. for 5 nodes there are 1.048.576 possible directed graphs.

Fortunately there is a huge amount of methods, which provide solutions. Structure learning algorithms are divided into 3 main groups: (i)"search and score", (ii)"constraint-based" and (iii)"hybrid" algorithms (Daly et al., 2011). Search and score methods require a score to determine how well the (estimated) graph fits the data. This scoring function is optimized by a search algorithm. Several scoring functions might be used, for example the Likelihood (Cooper and Herskovits, 1992), the Bayesian Information Criterion (BIC) (Schwarz et al., 1978) or the Akaike Information Criterion (Akaike, 1974). Examples for score and search algorithms are the Markov chain Monte Carlo model composition algorithm (Madigan et al., 1995), the Sparse Candidate algorithm (Friedman et al., 1999), the optimal reinsertion algorithm (Moore and Wong, 2003), Genetic algorithms (Larrañaga et al., 1996a,b; de Campos and Huete, 2000; Wong and Leung, 2004; Gao et al., 2007) or the Greedy Search algorithm (Chickering, 2002b).

Even if there are some search and score methods for causal inference (Heckerman, 1995), the majority of causal structure learning methods are constraint based methods (Daly et al., 2011). Constraint based algorithms use conditional independencies (CIs) to ex-

plore the network structure. These CIs are explored by statistical independence tests on triples X, Y and Z, where X and Y are single variables and Z describes a subset of nodes in the network. If X and Y are independent given the set Z ($X \perp\!\!\!\perp Y|Z$), then the edge from X to Y is removed from the graph. In general constraint based methods are much faster than search and score methods. Testing whether $X \perp\!\!\!\perp Y|Z$ requires looking at the variables X, Y and Z while changing an edge in a score based method means that we need to calculate a score over all variables in the network. Spirtes and Glymour (1991) developed one of the most common constraint based algorithms, the PC algorithm. It consists of two steps: the first step, which estimates the skeleton of the network and the second step, where the v-structures are calculated and some further orientation rules are applied to orient the edges. The parameter α of the PC algorithm influences the sparseness of the estimated network, such that larger values of α increase the number of edges in the network. However, the PC algorithm cannot direct all edges since different networks with identical skeletons can encode the same conditional independence assumptions and thus can not be distinguished on observational data only (Verma and Pearl, 1990). The result of the PC algorithm is the equivalence class of graphs represented by a CPDAG. Kalisch and Bühlmann (2007) showed that the PC algorithm is statistical consistent and feasible for high-dimensional data up to thousands of variables as long as the underlying DAG is sparse. The PC algorithm is order-dependent, which means, that the result of the algorithm depends on the ordering of the variables, since this ordering determines the statistical tests to be made. But Colombo and Maathuis (2014) proposed some modifications of the PC algorithm to overcome of this order-dependence in parts.
Hybrid methods combine the concepts of "search and score" and "constraint based" methods. The Max-Min Hill-Climbing (MMHC) algorithm from Tsamardinos et al. (2006) is a widely known representative of this class of structure learning algorithms. The first part of the MMHC algorithm uses the constraint-based MMPC algorithm (Tsamardinos et al., 2003) to reconstruct the skeleton of the graph, while in the second step a "search and score" strategy orients the edges based on the skeleton. An algorithm based on the MMHC algorithm proposed by Nägele et al. (2007) allows the estimation of networks with ten thousands of variables. This algorithm focuses on the Markov Blankets of each node. A Markov Blanket is a special subgraph around a certain node, that contains all other nodes of the graph, which are necessary to predict the behavior of that node and its parents (Pearl, 1988). This variation of the MMHC algorithm estimates the Markov Blankets around each variable and combines the results. The key benefit of this substructure learning is the possibility to parallelize the estimation of Markov Blankets.

1.2.4 Estimation of causal effects from observational data with unknown causal structure

Section 1.2.2 describes how to estimate causal effects from observational data, when the causal structure is known, and section 1.2.3 discusses the possibilities to estimate a causal structure from observational data. In this section, I introduce "Intervention calculus when the Directed acyclic graph is Absent" (IDA, Maathuis et al. (2009, 2010)), which combines these two ideas. IDA consists of two steps: First the PC algorithm estimates a CPDAG from the observational data (Kalisch and Bühlmann, 2007) and then Pearl's do calculus (Pearl, 2009) is used to estimate the causal effects from observational data using the estimated CPDAG in the second step. Since Maathuis et al. (2009) assume Gaussianity the causal effect is estimated using equation 1.8. But there are undirected edges in a CPDAG and this complicates the estimation of causal effects. However, the approach of Maathuis et al. (2009) allows to estimate lower bounds of the effect size. The IDA algorithm first enumerates all possible parent sets of a node X and then calculates causal effects of X on the targets of interest for each of the possible parent sets by equation 1.8. Thus, IDA estimates not necessarily exactly one, but a multiset of causal effects for each ordered pair X and Y (see Figure 1.5 for an example). The minimum absolute value of this multiset provides a lower bound of the absolute effect size. Maathuis et al. (2010) show in applications that IDA outperforms regression-based methods in terms of number of true positives versus number of false positives for the top 5000 predicted effects on the transcriptome of yeast gene deletion strains from a large dataset of expression profiles of wild type yeast. However, in simulations with large networks and medium sized datasets, which are typical for many biological applications, networks often cannot be reconstructed correctly. Meinshausen and Bühlmann (2010) suggest stability selection, a subsampling strategy that is wrapped around IDA as a remedy. K subsets of the data are drawn and for each of these subsets IDA estimates the lower bound of the effects. Finally, the effects are ranked by how often they appear in the top q effects. This procedure is further improved by the CStaR algorithm of Stekhoven et al. (2012) which repeats stability selection for several values of q and the median rank over different values of q is used to calculate the final rank of a causal effect. The CStaR algorithm outperforms plain IDA with respect to true positive selections versus false positive selections (Stekhoven et al., 2012).

1.3 Outline

The thesis is organized in two parts:

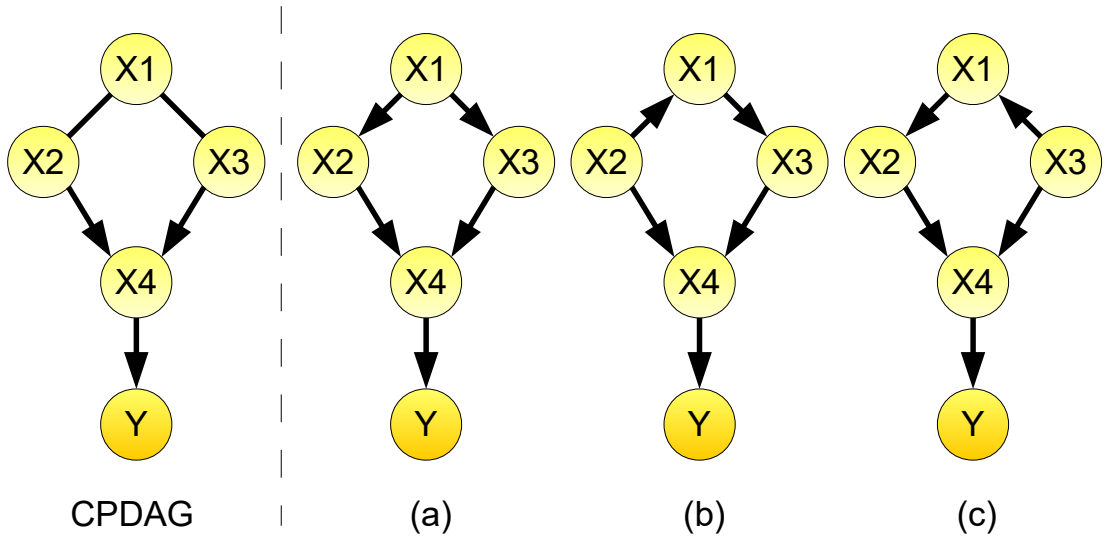


Figure 1.5: A CPDAG and its accompanying DAGs. There are three DAGs ((a)-(c)) belonging to the CPDAG on the left. If we are interested in the causal effect of X1 on Y, IDA calculates the multiset of three causal effects of X1 on Y according to the three DAGs corresponding to the CPDAG. In DAG (a) X1 has no parents and the causal effect of X1 on Y is β_{X1} from $Y = \beta_0 + \beta_{X1}X1 + \epsilon$. The parent of X1 in DAG (b) is X2 and thus the causal effect of X1 on Y is β_{X1} from $Y = \beta_0 + \beta_{X1}X1 + \beta_{X2}X2 + \epsilon$ and similarly the causal effect of X1 on Y is β_{X1} from $Y = \beta_0 + \beta_{X1}X1 + \beta_{X3}X3 + \epsilon$ for DAG (c). The minimum absolute value of these thee causal effects is a lower bound for the effect size.

Part1: In this methodological part I derive methods to work on the following three issues:
First, *MYC* is an transcriptional amplifier and we have to account for that during experimental design. There are already methods described in Section 1.1.3 that deal with that problem in transcriptomics, but these methods do not account for variability in cell lysis during RNA extraction. Thus, I introduce a new calibration method for gene expression data measured under transcriptional amplification. This method uses spike-in cells from *Drosophila melanogaster* to calibrate samples from different amounts of total RNA (Taruttis et al., 2017). In comparison to the usage of the ERCC spike-in set described in Section 1.1.3 the proposed method not only accounts for cell lysis effects but is also cheaper and easier to handle than the ERCC spike-ins.
Second, causal inference from observational data is a hard task. From the introducing Section 1.2 we know that there are already methods for the estimation of causal effects from observational data and that a subsampling approach is highly recommended. However the described methods only estimate a lower bound of the causal effects and do not use the causal effects of all multi sets. The method I propose is called accumulation IDA (aIDA) and uses the mode of the distribution of causal effects to generate the causal effect from the effects estimated from the subsampling runs. This advancement provides an improvement of the already existing methods.
Third, an assumption of these causal inference methods is that the data underlying graph is sparse. Since *MYC* is a transcriptional amplifier the graph underlying a dataset which includes the gene *MYC* is very likely not sparse, since *MYC* influences nearly every gene in the dataset. The spike-in calibration methods help to uncover this dense structure in a transcriptomic dataset. Computationally, the calibration of the data using a spike-in method will result in highly correlated data. So far it is unclear how to deal with this high dependencies since they violate the assumption that the underlying graph is sparse. I try to improve the estimation of causal effects from highly correlated observational data by replacing the PC algorithm with a version of the MMHC algorithm.

Part II: Part II applies the algorithms and methods developed in Part I to a data set, which is appropriate to examine the causal connections around *MYC*. The data set consists of transcriptomic and metabolomic data to unravel causal connections between genes and metabolites. I calibrate the RNA-seq data using the whole spike-in method presented in section 2.4, which results in highly correlated data. Thus, I estimate the causal effects of the genes on the metabolites and the causal effects between genes using the MMHC-aIDA algorithm described in section 4.4, since it shows a better performance on highly correlated data. Whereas aIDA (section 3.4) estimates the causal effects of

the metabolites on *MYC*. Some of the discovered causal relations are already published, others are indications for future experiments.

This thesis shows how causal inference methods enable insights into the interactions between the gene *MYC* with other genes and the metabolome in the context of lymphoma. The underlying data is purely observational, no experimental intervention is needed. Both, the application of my new calibration method and the improved algorithms for causal inference from observational data enable new insights into the complex functions of *MYC* in lymphoma cells.

Part I

Methods

CHAPTER 2

External calibration with Drosophila whole-cell spike-ins delivers absolute mRNA fold changes from human RNA-Seq data

2.1 Section introduction: Global changes of RNA amount between conditions require special normalization techniques

Standard RNA-seq microarray and qRT-PCR protocols start with a fixed size aliquot of RNA. Nevertheless, technical variation in data generation results in variable library sizes or total array intensities, which are addressed by computational normalization methods. However these purely computational normalization methods assume that the total amount of RNA between different conditions to be constant. Every increase of a gene must be compensated by the decrease of other genes, when a fixed size aliquot of RNA is used as the reference point for the gene expression measurements. Even if usually not reported, these measurements reflect percentages of the total transcriptome. This might lead to a distorted view on cellular processes whenever this assumption is violated. In fact using a fixed amount of RNA across samples can introduce artificial dependencies between gene expression levels. Imagine two experimental conditions A and B. If a gene constitutes 0.1% of the transcriptome in cells A and 0.05% of the transcriptome in cells B, a fold change $\frac{B}{A}$ of 0.5 is estimated. However, imagine the amount of RNA in B is four times higher than in A. Now, the estimated fold change $\frac{B}{A}$ is 2. In the relative scenario we assume a constant amount of RNA across samples and observe a downregulation while

Condition	Number of mRNAs of gene X	Fold change of absolute numbers B vs. A	Total number of mRNAs	Proportion of gene X	Fold change of proportions B vs. A
A	100	2	100 000	0.1%	$\frac{1}{2}$
B	200		400 000	0.05%	

Table 2.1: Example for fold change calculation under transcriptional amplification.

in the absolute scenario we take the change of RNA amount across samples into account and find an upregulation between the conditions B and A (Table 2.1 shows an example.). Indeed, we observe changes in RNA amount across samples: Buettner et al. (2015) detected differences in RNA content of cells in different stages of the cell cycle and pointed out that these differences affect interventional RNA-seq studies. Moreover transcriptional amplification is caused by the induction of transcription factor MYC (Lin et al., 2012; Nie et al., 2012) or lipo-poly-saccharid (LPS) (Sabò et al., 2014). But also heat shock and serum starvation trigger global changes of the RNA amount (van de Peppel et al., 2003).

This argument changes the reference point of gene expression data from a fixed size aliquot of RNA to a fixed number of cells. The resulting gene expression values do not add up to a fixed number. They are no percentages anymore. Nevertheless, relative and absolute gene expression measurements are equal as long as the total amount of RNA per cell stays almost constant across all samples.

To enable the change of the reference point from a fixed size aliquot to a fixed number of cells an external standard is added to the sample at an early stage of the protocol, before the fixed amount of RNA is taken from the samples. After that the standard protocols and normalization techniques are applied and the data can be rescaled to a constant number of cells. Lin et al. (2012) and Lovén et al. (2012) already used synthetic RNA spike-ins for the normalization of microarray and RNA-seq gene expression data to the number of cells. An alternative approach uses the amount of extracted mRNA quantified by the total polyA+ content (Aanes et al., 2014). However, there is no approach so far which controls cell lysis and RNA extraction which are crucial steps during the protocol. To overcome this problem in the context of dynamic expression analysis studies Sun et al. (2012) developed a whole cell spike-in method. Here, the cells of the organism under study are mixed with cells from a spike-in organism. These mixed populations are lysed an hybridized to custom microarrays containing probes from both organisms. This idea can be transferred to RNA-seq as long as the reads from the spike-in organism can be reliably separated from the organism under study. Continuing the work of Sun et al. (2012) and Lovén et al. (2012) we show that *Drosophila melanogaster* is a suitable spike

in organism for human RNA-seq profiling studies. And we show that the data calibrated by the *Drosophila melanogaster* spike-in cells leads to better estimates of absolute fold changes and outperforms synthetic RNA spike-ins.

2.2 Sample preparation and data analysis

2.2.1 Experimental design

To show that *Drosophila melanogaster* is a suitable spike-in organism for human cells and that calibration on *Drosophila melanogaster* spike-in genes allows to observe gene expression changes that refer to the absolute number of cells as reference point, we designed a systematic dilution experiment to simulate variable amounts of total RNA in fixed number of cells. 500 000 (condition A), 1 million (condition B), and 2 million (condition C) P493-6 cells were mixed with 100.000 vital *Drosophila melanogaster* S2 cells in triplicates. To compare this approach to the method of Lovén et al. (2012) we also added external control RNAs devised by the External RNA control consortium (ERCC) (Baker et al., 2005) to endogenous RNA. This ERCC kit consists of 92 poly-adenylated transcripts from Bacillus subtilis with lengths between 250 and 2000 nucleotides. They can serve two different purposes: (i) they monitor technical variance from RNA extraction via library generation to raw data analysis. The earlier in the protocol they are added, the more steps are monitored. Moreover, rescaling the data to constant-spike in intensities corrects for some of the technical variance. (ii) They can be used to change the reference point from a fixed size aliquot of RNA to a fixed number of cells. This requires adding them before the first fixed size aliquot of RNA is taken. Figure 2.1 summarizes the setup of our dilution experiment. In a second experiment we show that global gene expression changes may occur under different perturbations of the human cell line P493-6 and are observable only when the gene expression data is calibrated to the number of cells in the sample. The P493-6 cell line allows for the ectopic induction of *MYC* by a tetracycline-controlled transcriptional activation. Hence, these cells can be grown in a "*MYC*-high" (no tetracycline added, *MYC* is expressed) and in a "*MYC*-low" (tetracycline added, ectopic *MYC* is repressed) state. One million P493-6 cells in "*MYC*-high" and "*MYC*-low" state, respectively, were spiked with 100.000 *Drosophila melanogaster* S2 cells in 10 replicates for each *MYC* level and the polyA+ fraction of the transcriptomes was sequenced. Further the "*MYC*-low" cells were treated with α-IgM F(ab)2 fragments referred to as "BCR", with sCD40L referred to as "CD40L" and CpG (two replicates for each stimulation). We refer to this experiment as the stimulation experiment. For each sample the general experimental setup was the following: The mixed cells were lysed,

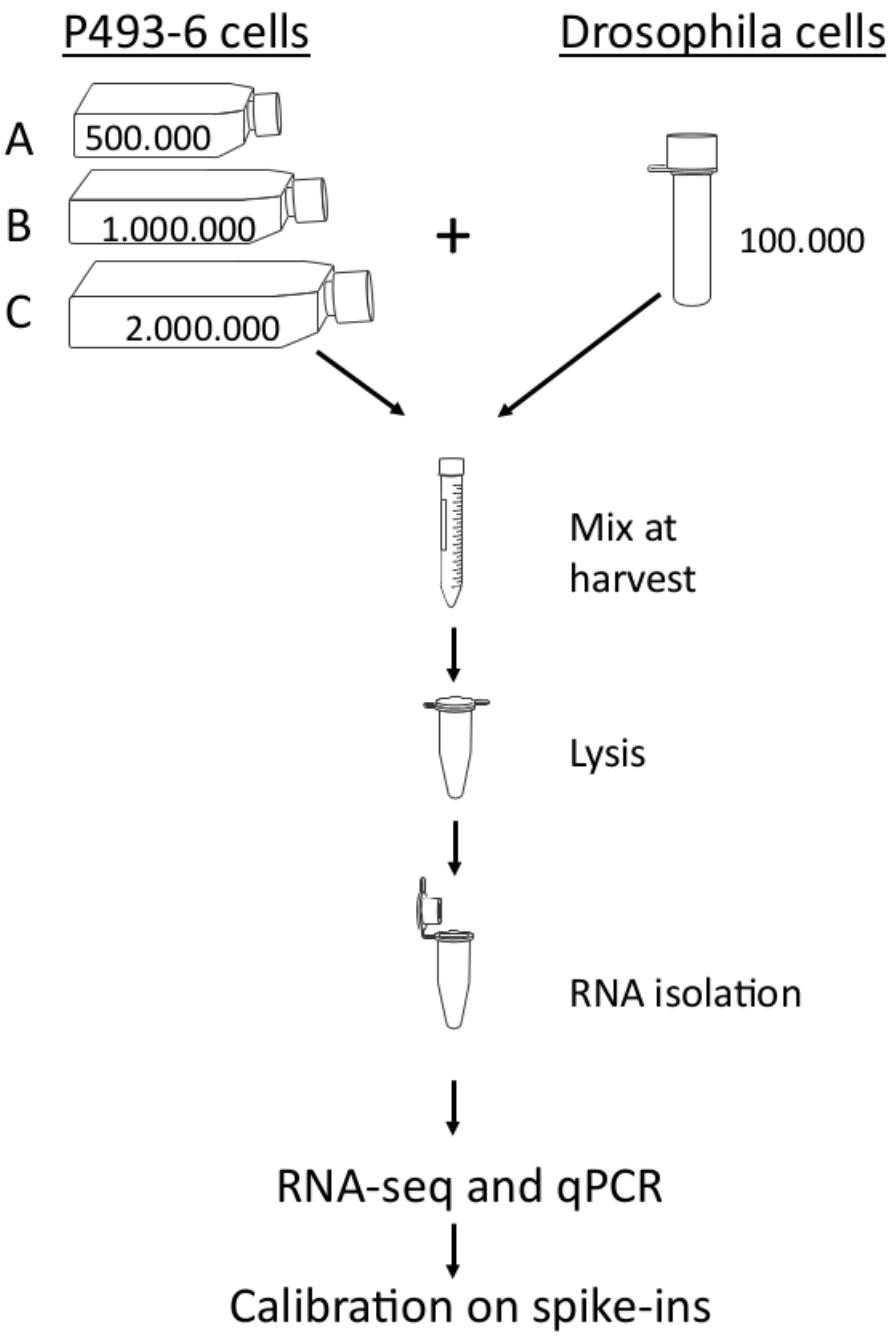


Figure 2.1: Experimental setup. Condition A contained 500 000 P493-6 cells, condition B one million P493-6 cells and condition C two million P493-6 cells. Samples were spiked with 100 000 S2 *Drosophila melanogaster* cells, mixed, lysed, and total RNA extracted. RNA was subjected to RNA sequencing and qPCR. Each condition A, B and C, consists of three replicate samples.

RNA was extracted and the ERCC spike-in RNA was added to allow for the comparison between these synthetic spike-insand *Drosophila melanogaster* spike-in cells. Gene expression of the samples was measured by RNA sequencing. Maren Feist (Department of Haematology and Medical Oncology of the University Medical Center Göttingen) carried out all wet lab experiments. For details on sample preparation see Appendix A.

2.2.2 Preparation of a custom genome

After RNA sequencing the reads need to be assigned to their genomes. For this purpose I constructed a concatenated genome consisting of the human and *Drosophila melanogaster* reference genomes and the ERCC sequences. The human reference was GRCh38 from ensembl, release 77 (Cunningham et al., 2015), the *Drosophila melanogaster* reference was ensembl BDGP5, release 77 and the ERCC sequences were as provided by https://tools.lifetechnologies.com/content/sfs/manuals/cms_095047.txt. This approach ensures that each read is mapped once either to the human or the *Drosophila melanogaster* genome or to the ERCC sequences. Thus, I get a clear assignment of each read to a species.
I mapped all sequence libraries against this concatenated genome using TopHat version 2.0.13 (Kim et al., 2013) indicating an unstranded sequencing protocol (–library-type fr-unstranded) and default settings for the remaining parameters.

2.2.3 Normalization and differential gene expression analysis

I assigned read counts to ensembl gene identifiers using featureCounts version 1.4.5 (Liao et al., 2014). For all datasets, I selected all human genes with more than 100 counts for each sample for normalization and differential gene expression analysis of the respective dataset. Of the *Drosophila melanogaster* genes and ERCC transcripts, the ones with a sum of at least 100 counts over all samples were kept. In the dilution dataset, additionally, all genes with zero counts in at least one sample were removed such that different calibration methods could be run on the same dataset. This resulted in 10028 human ensembl genes, 6070 Drosophila ensembl genes and 43 ERCC transcripts for the dilution dataset. In the dataset consisting of different stimulations of P493-6 cells in "*MYC*-low" conditions and unstimulated "*MYC*-low" and controls, controls, 8477 human ensembl genes, 7050 Drosophila ensembl genes and 52 ERCC transcripts are analyzed.
First, I normalized the data by library size factors calculated with DESeq2 (Love et al., 2014), which refers to a fixed size aliquot of RNA as reference point. As I described in Section 2.1 and shown in Table 2.1 I suggest to change the reference point from this fixed size aliquot to an equal number of cells. For the calibration on the *Drosophila*

melanogaster spike-in cells I calculated the library size factors on the counts of the *Drosophila melanogaster* genes by DESeq2, and applied these to the human gene counts of each sample. For the calibration based on ERCC synthetic gene counts I obtained the scaling factors on the ERCC 'gene' counts as described above, and applied these factors to the human sample data. For all three scenarios I estimated the log2 fold changes with DESeq2 (Love et al., 2014). All fold changes that meet the count cut off above are included in the figures, irrespective of the significance of the fold change.

2.3 Results

2.3.1 *Drosophila melanogaster* cells are suitable spike-in cells for human RNA-seq studies

We decided to use *Drosophila melanogaster* as spike-in organism since this allows the use of the same lysis protocols like the human cells under study. Thus, we mixed these spike-in cells with human P493-6 cells, lysed them, sequenced the RNA and I mapped the resulting libraries to the concatenated genome. Reads that uniquely map to the *Drosophila melanogaster* part of the genome are used to adjust the data to a fixed amount of cells.

To validate this approach we selected two libraries from *Drosophila melanogaster* S2 cells (accession numbers SRR569914 and SRR424185) and two libraries from human P493-6 cells (accession numbers: SRR567561 and SRR567562) from the SRA database (Leinonen et al., 2010). All libraries have been sequenced in single end mode with 50 bp reads for the *Drosophila melanogaster* S2 libraries and 40 bp reads for the P493-6 libraries. To test how many false negative mapped reads are expected, I aligned the human libraries to the human genome alone and to the concatenated genomes of *Homo sapiens* and *Drosophila melanogaster*. The number of multi-mapped reads increased from 85,298,194 to 85,324,091 and from 84,178,617 to 84,205,272, respectively, which corresponds to an increase of 0.02%. Thus, human reads can reliably be identified as human, after mapping against the concatenated genome. Then I mapped the *Drosophila melanogaster* reads against the concatenated genome with the result that less then 0.01% aligned against the human part. This shows that *Drosophila melanogaster* can be used to calibrate the data to total cell number. More results from cross mapping studies are shown in Tables 2.2 and 2.3. In summary the cross mappings between *Drosophila melanogaster* and human genomes are negligible and with that we find that *Drosophila melanogaster* is a suitable spike-in organism for the human P493-6 cells.

[1] i.e. reads uniquely mapped to Drosophila

Library	Reference	Uniquely mapped reads	Multi-mapped reads	Gain of multi-mapped reads		Gain of uniquely mapped reads [1]	
				absolute	relative (%)	absolute	relative[2] (%)
SRR567561	custom	106981050	14366462	1180	0.008	8052123	5.792
	Human	98928927	14365282				
SRR567562	custom	107646456	17351828	2923	0.017	1479379	1.008
	Human	106167077	17348905				

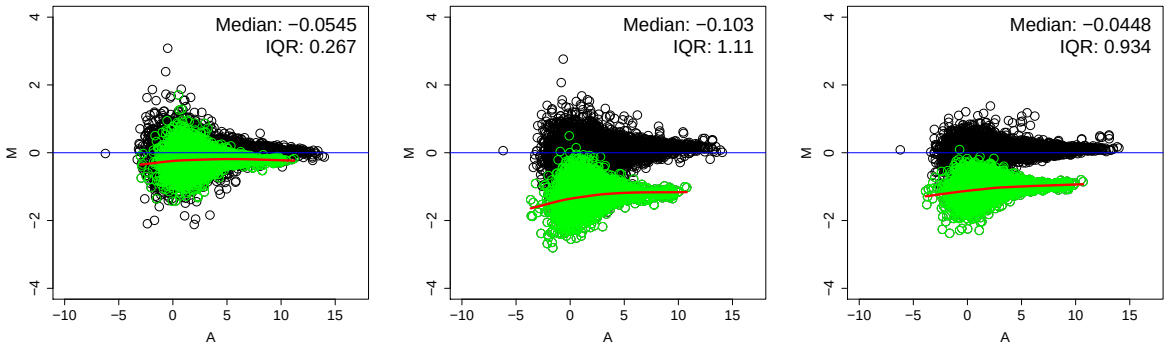
Table 2.2: Multi-mapped reads introduced by adding the *Drosophila melanogaster* genome to the human reference genome.

Library	Total number of reads	Reference	Reads assigned to Human genes	Reads assigned to Drosophila genes	True Positive Rate	True Negative Rate
SRR569914	48 million	custom	416	35742573	-	99.9988%
SRR424185	37 million	custom	72	27303679	-	99.9997%
SRR567561	140 million	custom	80357899	464	99.9994%	-
SRR567562	147 million	custom	87443273	780	99.9991%	-

Table 2.3: Summary of counts assigned to human and *Drosophila melanogaster* genes (by featureCounts) of human and Drosophila libraries mapped to the concatenated human-Drosophila reference genome ('custom') and corresponding True Positive and True Negative Rates.

2.3.2 The calibration by whole cell spike-ins can be done in multiple ways

In our dilution experiment (Section 2.2.1) the spike-in genes are distributed across the whole spectrum of expression levels (Figure 2.2). I compare three ways of spike-in normal-



(a) B versus A **(b)** C versus A **(c)** C versus B

Figure 2.2: MA plots for the 3 conditions. M (e.g.: $\log_2(\frac{C}{B})$) versus A ($\frac{1}{2}\log_2(BC)$) for the 3 calculated $\log_2$ fold changes. The expression levels of the *Drosophila melanogaster* genes (green) cover the whole spectrum human genes (black).

ization: First, I decided to aggregate the data of the dilution experiment by computing size factors on the spike-in genes and I apply these factors to the human genes using DESeq2 (Love et al., 2014). If not stated differently, I will use this approach during the whole chapter for data calibration.
Alternatively a loess regression on the spike-in genes as suggested by Lovén et al. (2012) also adjusts the data to RNA amount per cell. And third, the upper quartile of the spike-in distribution adjusts the data to the total number of cells.
I applied all three options for both, the ERCC kit and and *Drosophila melanogaster* cells as external standards, but did not observe significant differences between the three different computational methods (Figure 2.3). However, Figure 2.3 shows that the calibration by *Drosophila melanogaster* cells outperforms the calibration by the ERCC kit.

2.3.3 Spike-in adjusted data provides estimates of differential expression that are calibrated to the total number of cells

With the help of spike-in cells and ERCC kit, respectively, I calibrate our measurements to a constant amount of cells. To validate that hypothesis I use the dilution experiment (Section 2.2.1). By design the amount of total RNA increases linearly across the 3

[2]relative to uniquely mapped reads of human refernce genome

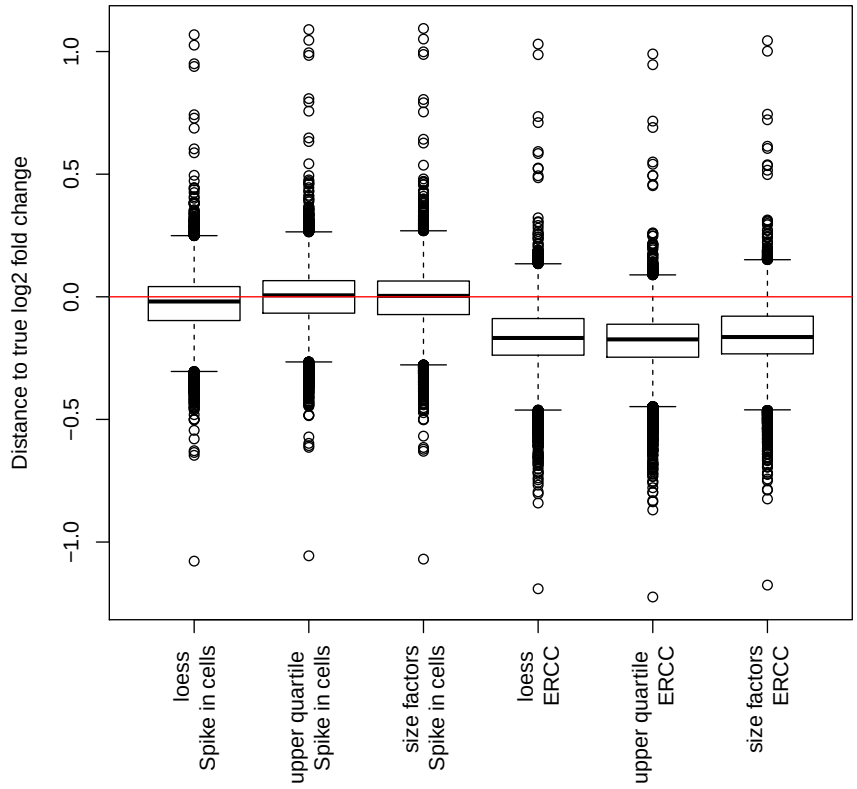
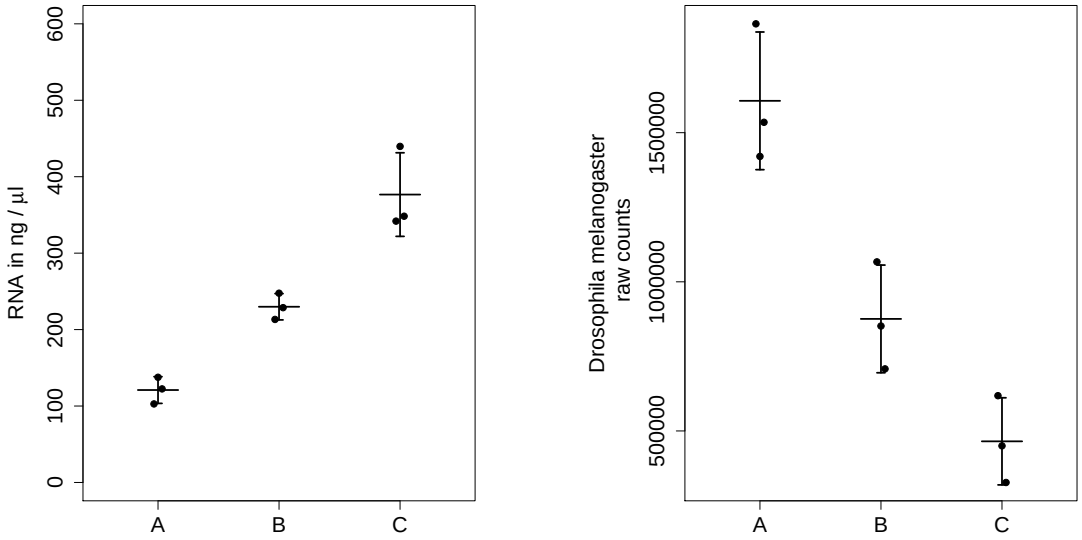


Figure 2.3: Comparison of computational spike-in calibration methods. Three different computational methods calibrate the spike-in data. The "loess" method performs a loess regression on the spike-in genes, "upper quartile" uses the upper quartile of the spike-in genes to scale the data and the "size factors" method calculates the DESeq2 size factors using the spike-in genes. All methods have been applied to the dilution dataset and the distance between the expected and the estimated $\log_2$ fold change was calculated. All methods perform equally well. Further I tested two kinds of spike-ins the *Drosophila melanogaster* spike-in cells and the ERCC spike-in kit. *Drosophila melanogaster* spike-in cells outperform the ERCC kit for all three different methods.

conditions A, B and C (Figure 2.4 (a)). I analyze this dataset (i) relative to a fixed amount of RNA, (ii) to a constant number of cells across samples using the ERCC spike-ins and (iii) to a constant number of cells across samples using the *Drosophila melanogaster* spike-in cells. For (i) we use the library size corrected data for differential gene expression analysis between the 3 conditions A, B and C. For (iii) the percentage of *Drosophila melanogaster* decreases linearly with the total increasing amount of RNA (Figure 2.4). Now, all reads in a sample are scaled such that the number of *Drosophila melanogaster*



(a) RNA amount **(b)** *Drosophila melanogaster* raw counts

Figure 2.4: Comparison of RNA amount and *Drosophila melanogaster* raw counts for the three conditions. While the RNA amount increases linearly across the three conditions (a) of the dilution dataset, the *Drosophila melanogaster* raw counts decrease (b).

reads stays constant across samples. I follow the same strategy for the ERCC spike-in kit.

Figure 2.5 shows the distances between the expected log2 fold changes and the estimated log2 fold changes for the three different calibration strategies. The data calibrated by the fixed aliquot of RNA as reference point is not able to reproduce the expected fold changes, while both methods which calibrate to a reference point which is related to the total number of cells reproduce the expected fold changes well.

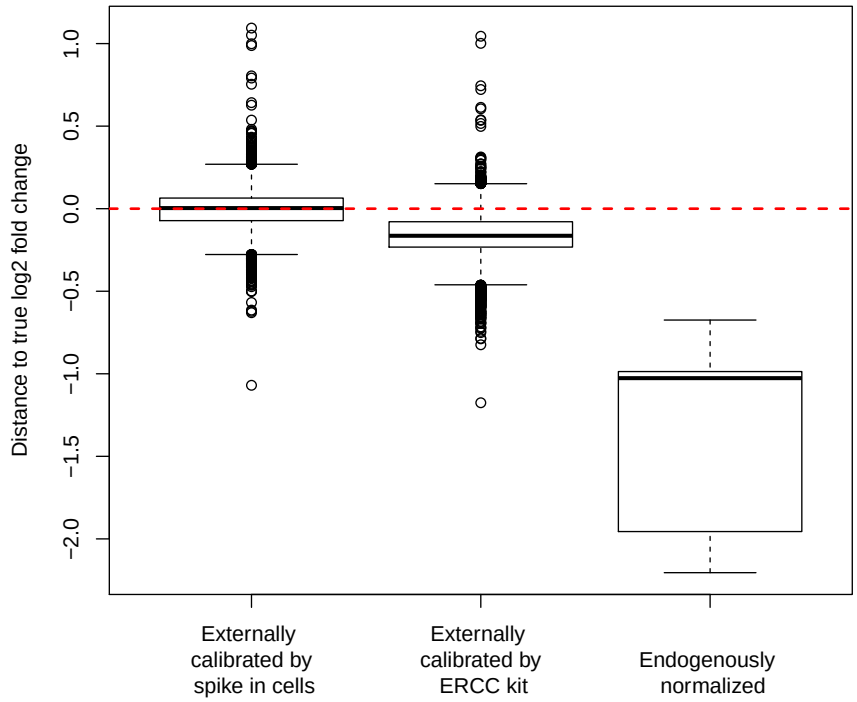


Figure 2.5: Comparison of data calibrated by *Drosophila melanogaster* spike-in cells or ERCC kit, respectively, to endogenously normalized data. Distances to true $\log_2$ ratios after calibrating the human gene expression data on *Drosophila melanogaster* spike-in data (left), after calibrating on ERCC synthetic RNA spike-in data (center), and after endogenous normalization (relative to a fixed weight human RNA aliquot, right). The *Drosophila melanogaster* spike-in protocol outperforms the calibration by ERCC kit and the endogenous normalization.

2.3.4 Whole cell spike-in calibration outperforms other calibration methods

In addition to the above mentioned external spike-ins there are other methods to calibrate the data. Since the whole spike-ins are a measurement of endogenous RNA in the samples, one could also use a direct quantification of RNA. Thus, for the dilution experiment we quantify the amount of total RNA per sample and apply a simple scaling approach to the RNA-seq data for calibration. This approach is similar but not identical with the method proposed by Aanes et al. (2014).
Computationally the RUVg tool provides an alternative approach for adjusting profiles to an external standard (Risso et al., 2014). RUVg calibrates data relative to multiple spike-in measurements, but also claims to detect general transcriptional amplification on a purely computational basis. Figure 2.6 shows the deviations between expected log2 fold changes and log2 fold changes estimated by the three different approaches. Since RUVg alone does not help to calibrate the data, we also scale the data before adjustment. With that approach we are able to reproduce the expected fold changes. However, in our dilution experiment the RUVg algorithm does not improve over our scaling approach.

2.3.5 Whole cell spike-in calibration affirms *MYC* driven general transcriptional amplification in human P493-6 B-cells

Lin et al. (2012) and Nie et al. (2012) observed that in the human B cell line P493-6 induction of the transcription factor *MYC* amplifies the transcription of almost all actively transcribed genes in this cell line. I analyzed *MYC* driven general transcriptional amplification in our spike-in assay and confirmed the results of Lovén et al. (2012). I calibrated the human fraction of the RNA-seq data from the stimulation experiment by the two different reference points: (i) to the number of human cells, by scaling the data to constant accumulated *Drosophila melanogaster* read counts across samples, and (ii) to a fixed size aliquot by scaling the data to constant library size of the human fraction. After that I estimated the log2 fold changes between the "*MYC*-high" and the "*MYC*-low" state of the cells. Figure 2.7(a) shows, that the results deviate substantially between the two different reference points. Using the spike-in calibration, the median log2 fold change is approximately two reflecting imbalanced changes, while endogenous calibration led to log2 fold changes scattered around zero. Further we measured the amount of total RNA for the 10 replicates for each condition and observed a fold change of 1.8 (Figure 2.7(b)), which is in accordance with our spike-in calibration approach.

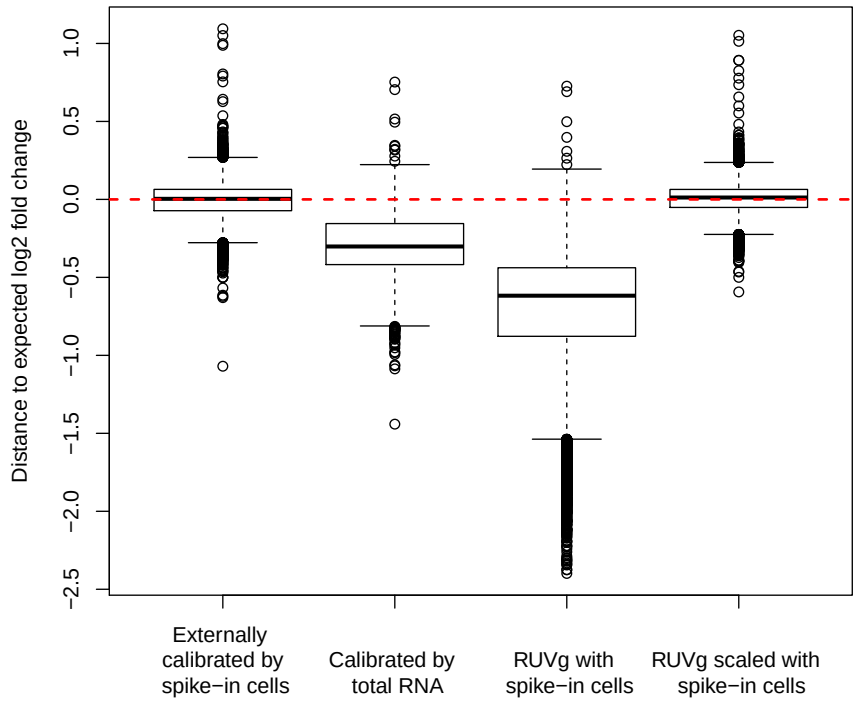
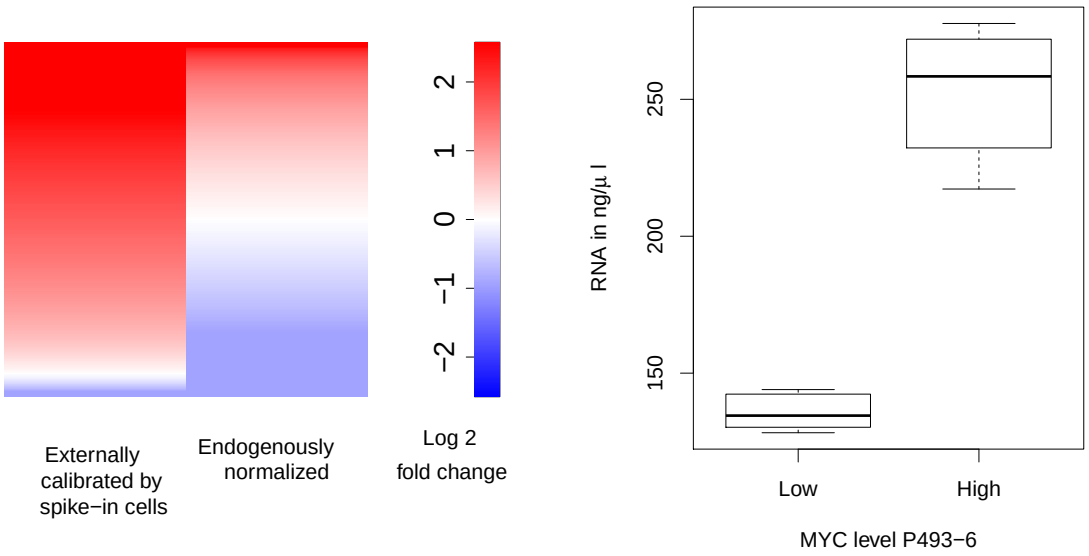


Figure 2.6: Comparison of *Drosophila melanogaster* spike-in cell protocol to different external calibration methods. Distances to the the expected $\log_2$ fold changes on the dilution dataset using the *Drosophila melanogaster* spike-in cells, total RNA, RUV$_g$ using *Drosophila melanogaster* genes, and RUV$_g$ after scaling the data to the upper quartile of the *Drosophila melanogaster* spike-in genes. The *Drosophila melanogaster* spike-in protocol outperforms the calibration by total RNA and RUV$_g$ using the same *Drosophila melanogaster* genes, while RUV$_g$ after scaling the data to the upper quartile of the *Drosophila melanogaster* spike-in genes performs equally well.



(a) **(b)**

Figure 2.7: Gene expression of P493-6 cells for two levels of *MYC*. (a) The $\log_2$ fold changes of P493-6 cells between the two different levels (n=10) differ substantially between the two different reference points. While we observe a balanced gene expression pattern for the endogenously calibrated data (relative to a fixed size aliquot of RNA), we find a upregulation of most of the genes when applying the *Drosophila melanogaster* spike-in protocol. (b) Induction of *MYC* in one million P493-6 cells leads to an increase of RNA amount. (n=10)

2.3.6 General transcriptional amplification in B cells is not limited to *MYC* regulation

Human P493-6 "*MYC*-low" cells have been treated with different stimuli, namely activation of toll like receptor signaling by the dinucleotide CpG, activation of the CD40 pathway by CD40L, and activation of BCR signaling via cross linking the B-cell receptor with an α-IgM F(ab)2 fragments. The estimation of log2 fold changes based on *Drosophila melanogaster* spike-in calibrated data indicates that also CpG and BCR stimulations induce general transcriptional amplification, while normalization based on human genes only shows a balanced pattern. Unlike CpG and BCR stimulation, CD40L shows balanced gene expression changes for both, *Drosophila melanogaster* spike-in cell calibration and normalization based on human genes (Figure 2.8).

We find that global transcriptional amplification is not an artificial observation, but a real physiologic process in B cells and only external calibration enables the observation of this behavior. Further we find that if there is no transcriptional amplification, as in the case of CD40L stimulation, the *Drosophila melanogaster* spike-in protocol is able to detect this.

2.4 Discussion and conclusions

I described an RNA-seq profiling protocol that uses *Drosophila melanogaster* spike-in cells for the calibration of human gene expression data. The *Drosophila melanogaster* spike-in cells allow for an unbiased estimation of fold changes with respect to a fixed number of cells. The RNA needs to be completely extracted, which deserves particular attention during the RNA extraction protocol. Our method outperforms the external calibration by the ERCC spike-in kit and by total RNA quantification. The whole protocol is experimentally simple and inexpensive in comparison to the commercial ERCC kit.

Our spike-in protocol adds an additional reference point, namely the total number cells. If required, the user can calibrate the data in both ways: using library size only based on a fixed size aliquot of RNA as reference point or using *Drosophila melanogaster* spike-ins. However, only with the spike-in corrected data I was able to estimate the expected log2 fold changes in our dilution experiment.

Furthermore, the *Drosophila melanogaster* spike-in cells outperform the ERCC spike-in kit for two reasons. First, the whole cell protocol corrects technical variability in cell lysis and RNA extraction. Since the ERCC kit is added after RNA extraction it cannot monitor these variations. Second, in comparison to the 92 transcripts of the ERCC kit, the *Drosophila melanogaster* spike-in protocol includes thousands of genes, that can be

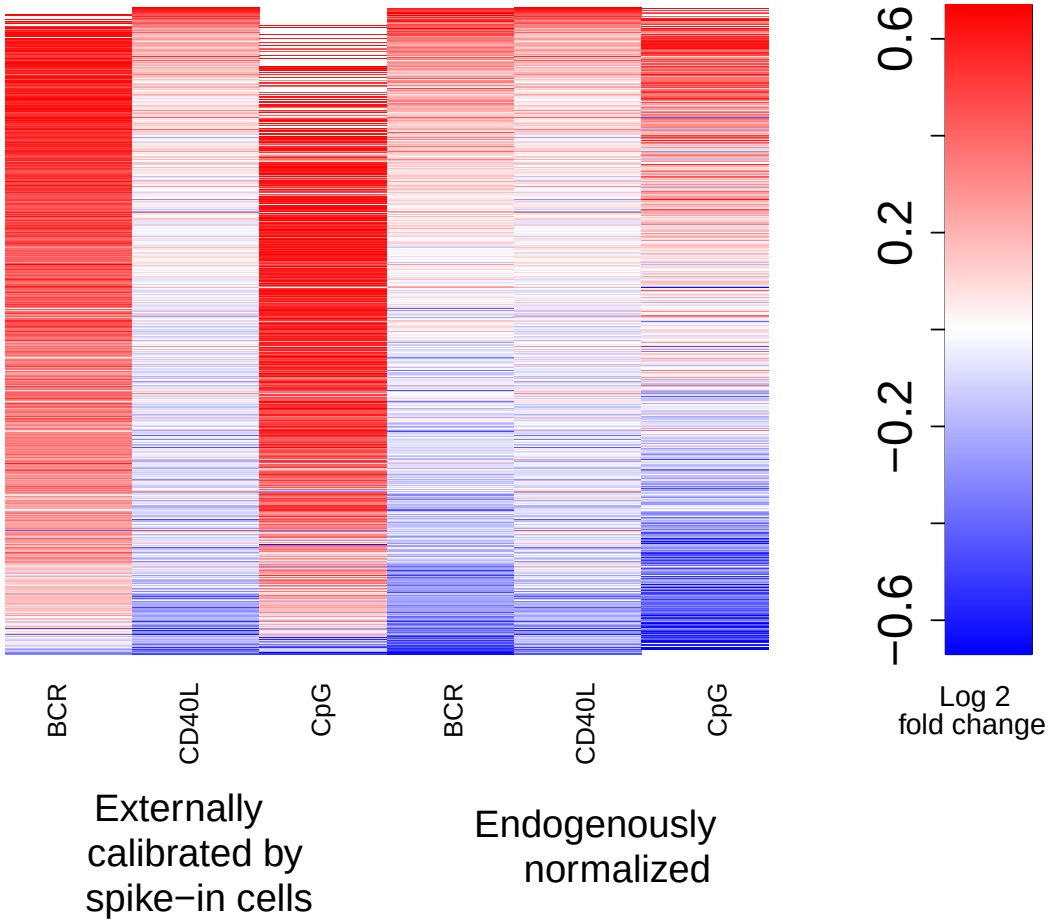


Figure 2.8: Gene expression of "*MYC*-low" P493-6 cells under different stimulations. Data was calibrated by *Drosophila melanogaster* spike-in cells (left) and endogenously normalized (right) under different stimulations (BCR: B cell receptor stimulation, CD40L: CD40 ligand, CpG: stimulation of Toll like receptor-signaling by the dinucleotide). For BCR and CpG we observe an upregulation of nearly all genes when applying the *Drosophila melanogaster* spike-in protocol, while we observe a balanced gene expression pattern for the endogenous normalization.

used for the calibration. As a result, the standard error of the calibration factors is lower in the *Drosophila melanogaster* spike-in protocol.

Furthermore, we observed, that transcriptional amplification is not an artifact, but a common biological phenomenon. The list of up- and downregulated genes varies substantially between the two different reference points. Using a fixed size aliquot of RNA as reference point increasing expression values must be compensated by decreasing others. In presence of transcriptional amplification, the observed downregulation of genes might not result from transcriptional repression but from the compensatory artifact associated with this reference point. Thus, we strongly recommend to use our whole cell *Drosophila melanogaster* spike-in protocol for differential expression analysis.

But also for the reconstruction of causal Bayesian networks we prefer the use of the spike-in protocol. Most methods for causal Bayesian network reconstruction use constraint based methods (Daly et al., 2011), thus they have to detect conditional dependencies between genes. However, the choice of the reference point highly affects the joint distribution of genes. Using a reference point which is associated with a fixed size aliquot of RNA leads to compensatory artifacts in the gene measurements, which has an influence on this joint distribution. This behavior influences the correlation structure between genes and, thus, leads to a biased estimate of the causal Bayesian networks.

CHAPTER 3

aIDA - A statistical approach to virtual cellular experiments

3.1 Motivation

The IDA method was ground-breaking, since Maathuis et al. (2010) were able to estimate the causal effects of 5361 genes in 234 yeast deletion strains from 63 expression profiles of wild type yeast. This was the first high throughput analysis of virtual perturbation experiments. Thus, IDA will influence the research in the field of systems biology significantly. Further improvements in the performance of IDA have been realized by the CStaR algorithm (Stekhoven et al., 2012), which uses stability selection, a resampling method wrapped around IDA that enhances the discovery of the causal effects.
However, both IDA and CStaR suffer from the poor quality of the estimated CPDAGs (Figure 3.7). Due to the small number of observations in comparison to the high number of variables, it is difficult to improve the accuracy of the estimated graphs (Kalisch and Bühlmann, 2007). aIDA provides a more effective way to extract the causal effects from inaccurate networks. The following two steps in the IDA/CStaR procedure offer room for improvement.

1. The selection of a causal effect from an estimated multiset The estimated CPDAG represents the equivalence class of the causal graph and is, thus, only partially directed. For a given pair of nodes X and Y and a given estimated CPDAG, X might be connected to other nodes via undirected edges. Thus, no unique set of parent of X can be determined. In these scenarios for each possible valid set of parents a causal effect is estimated. This leads to a multiset of causal effects. Both, IDA and CStaR, take the

minimum absolute value of these multiset as a lower bound of the causal effect. This approach ensures a guaranteed minimum size of the causal effects. However, whenever there is an undirected edge between the cause and another node of the graph the minimum absolute value will be zero (Section 1.2.2). Thus, the estimated causal effects are biased towards zero. aIDA tries to overcome these biased estimates of causal effects.

2. The method to summarize the estimated causal effects from the subsampling runs CStaR makes use of stability selection to find the most stable causal effects with the highest scores. However, there can be stable medium sized effects which do not appear within the top q causal effects. This behavior can lead to missed causal effects of smaller effect sizes. If an estimated causal effect is valid or not depends on the adjustment set Z in Definition 1.2.5. IDA and all its extensions use the parents derived from the estimated CPDAG as adjustment set. If the parent sets derived from the CPDAG are a valid adjustment set, then at least one of the estimated effects in $M(X \to Y)$ is valid, but it is unclear which one it is.

If the true underlying causal network is known, graph-based criteria (Tian and Pearl, 2002), e.g. the Back-door criterion (Pearl, 2003) help to determine the causal effects. If the true causal network is unknown, aIDA assumes that the distribution of causal effects across subsamples helps to estimate the accurate causal effects. Simulated data from known causal networks helps to study the distributions of valid causal effects versus invalid causal effects of the estimated CPDAG. Two observations support the advantages of using the distribution of causal effects and with that the aIDA approach.

1. The estimated parents often lead to valid estimates of causal effects even if they are not the true parents If the causal network underlying a dataset is estimated correctly, we can estimate valid causal effects from it. But absolutely correct estimated networks are rather unlikely. However they are also not required, since there are mistakes, which do not affect the estimation of causal effects. In other words, as long as the estimated parents fulfill the Back-door criterion relative to an ordered pair X and Y, the estimation of correct causal effects is possible, even if they are not the true parents of X.

To examine how often the Back-door criterion is fulfilled I generated random causal DAGs and used them to generate artificial gene expression data (Details in section C). I generated 50 samples for each random DAG of 1000 nodes. From that dataset I drew 100 subsamples of size n=25 and ran the PC-Algorithm as implemented in (Kalisch et al., 2012) on each subsample, which led to 100 estimated CPDAGs. For all possible ordered pairs of nodes X and Y I estimated the multisets $M(X \to Y)$. In my simulation scenario

the true parents of X are known and, thus, I can count how often the PC-algorithm identified the true parent set. Following Pearl's Back-door criterion the causal effects are identifiable in these cases. However, there are other gene sets Z that fulfill the Back-door criterion since finding the correct parents is sufficient, but not necessary to estimate the true causal effects. Since the true underlying networks are known from the simulation I was able to check, whether the Back-door criterion is fulfilled for the given adjustment set, or not. Figure 3.1 shows that the estimated parent sets are not the true parents in most of the cases. Nevertheless the Back-door criterion is fulfilled for the majority of estimated sets. This leads to valid estimates of the causal effects even if the estimated CPDAGs are highly inaccurate. My simulation study suggests that finding the true parents is difficult, but finding a valid adjustment set is not. For example, in more than 40% of the cause-effects pairs the Back-door criterion was fulfilled in more than 90% of the estimated parent sets over the 100 subsampling runs, conversely there were no subsampling runs that contained at least 90% true causal parent sets.
In addition to the simulations I also examined the reconstruction of biological networks. To this end, I used data from the DREAM3 In-Silico Network Challenge (Marbach et al., 2009, 2010; Prill et al., 2010). Two subgraphs from an *E. coli* transcriptional network and three subgraphs from a *S. cerevisae* transcriptional network with 100 nodes each and various degrees of sparseness (for details see Table 3.1) were used to sample 46 time series for each subgraph. I the study I use the time point zero data. Figure 3.1 is consistent with my findings in the simulated datasets, that the Back-door criterion is fulfilled for the majority of cases, while the detection of true parents remains difficult.

2. Valid estimates of causal effects from subsampling runs generate peaks in histograms over causal effects To illustrate the behavior of the estimated causal effects over multisets and subsampling runs, I used the 10 nodes toy graph in Figure 3.3(a). I sampled data from that graph as described in Section C and took 100 subsamples of that dataset. For each subsample I estimated the multiset of causal effects of node 6

Network	Nodes	Edges	Regulators
Ecoli1	100	125	26
Ecoli2	100	119	19
Yeast1	100	166	60
Yeast2	100	389	71
Yeast3	100	551	81

Table 3.1: Overview of the data sets with 100 nodes from the DREAM3 In-Silico Network Challenge

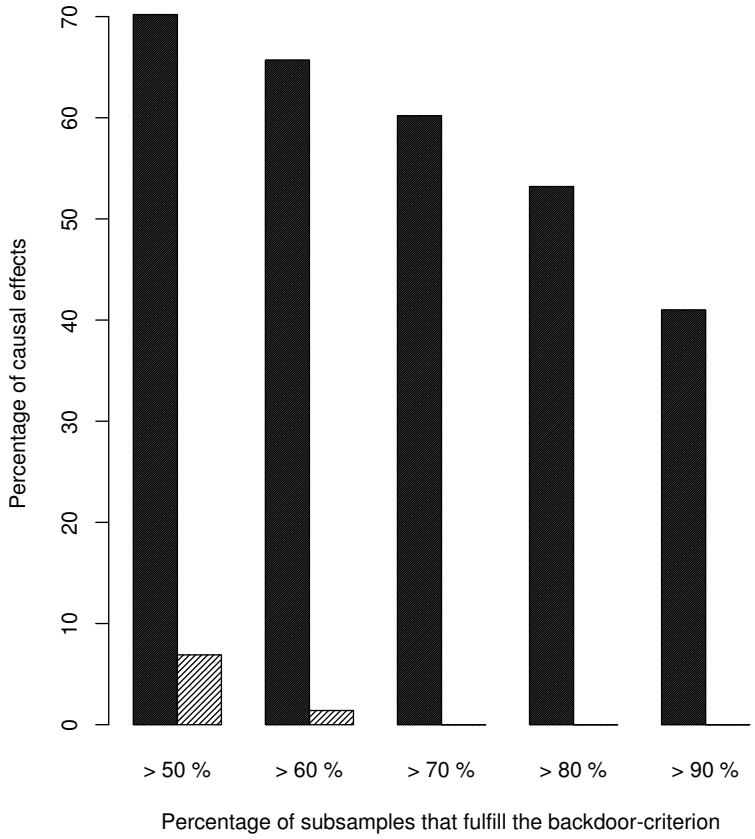


Figure 3.1: Percentage of subsamples that fulfill the Back-door criterion or where the true parents are detected for the artificial data sets The Back-door criterion is fulfilled frequently even if the PC algorithm rarely detects the true parents. 1000 cause-effect pairs were sampled randomly from a network with 1000 nodes. For each of these pairs I counted how often the Back-door criterion was fulfilled by the estimated parent sets within the 100 subsamples (black bars) and how often the parent set was the true parent set (gray bars). The x-axis shows the percentage of estimated causal effects of this 100 subsampling runs where the Back-door criterion was fulfilled. The y-axis displays the proportion of the 1000 sampled cause-effect pairs for which at least x subsampling runs fulfilled the criterion (black) or were the true parent sets (gray). The detection of true parent sets is a very hard task, but meeting the Back-door criterion is not.

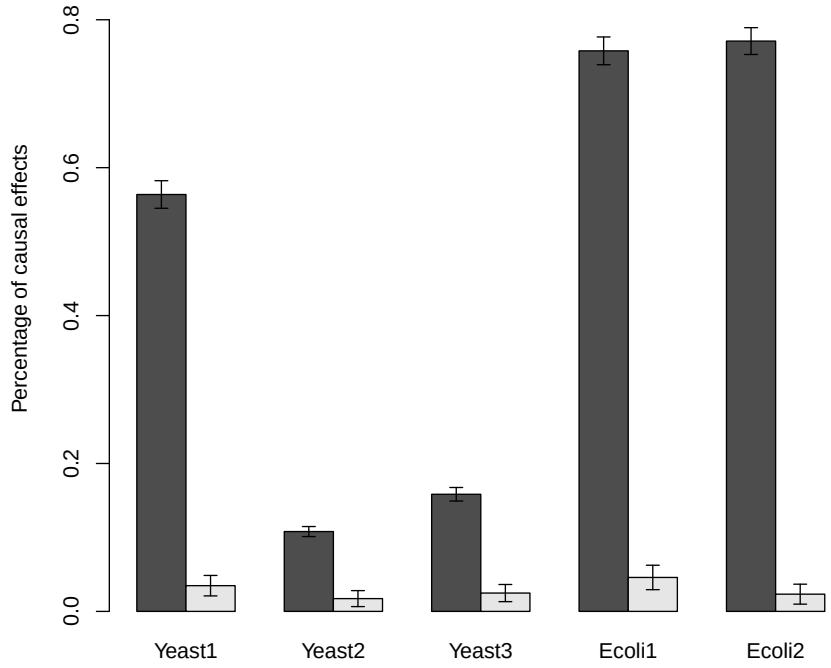
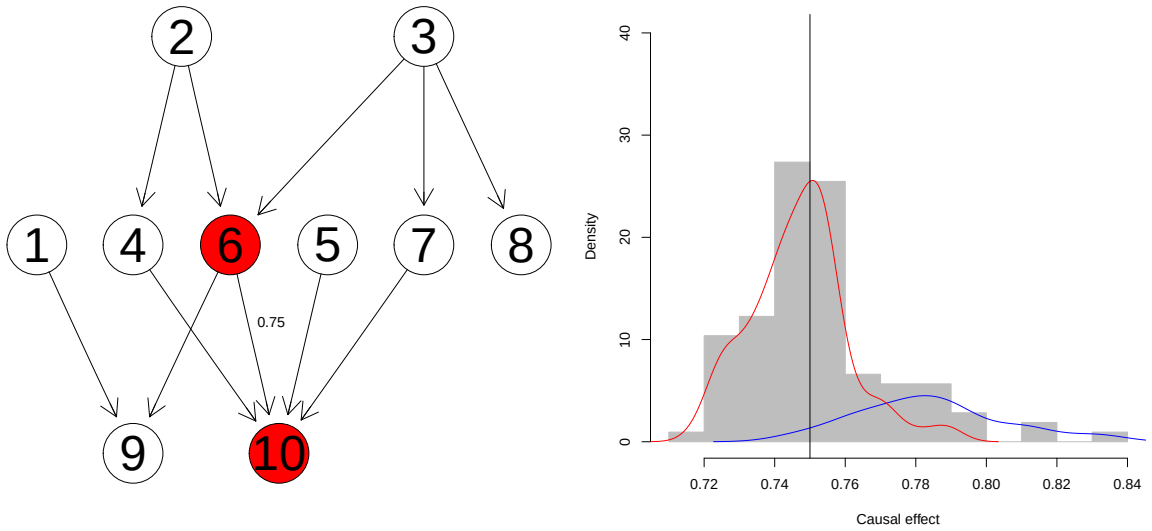


Figure 3.2: Percentage of subsamples that fulfill the Back-door criterion or where the true parents are detected for the DREAM3 challenge data sets Percentage of causal effects where the estimated parents fulfill the Back-door criterion (black) or are the true parents (gray), for the five DREAM3 data sets over the 100 subsampling runs



(a) **(b)**

Figure 3.3: Example for the estimation of the causal effect of node 6 on node 10 for a small simulated data set using aIDA. (a) The causal network of the graph with 10 nodes. The true causal effect of node 6 on node 10 (red) is 0.75. (b) The histogram over the estimated causal effects from the 100 subsampling runs (gray) peaks around the true causal effect (black vertical line). The red curve shows the estimated density derived from the valid adjustment sets (estimated using parents, that fulfill the Back-door criterion and the blue curve from invalid adjustment sets.

on node 10. I pooled these estimates across multisets and subsamples and I showed the distribution of causal effects in the light gray histogram of Figure 3.3(b). Additionally, I estimated a smoothed density of all causal effects estimated from valid adjustment sets (red curve in Figure 3.3(b)). I observed a peak in that estimated density around the true causal effect, depicted by the black vertical line in Figure 3.3(b). This is due to the fact that causal effects were estimated based on valid adjustment sets, and are, thus, unbiased estimates of the true causal effect. The true parents of the cause are not necessarily the only valid adjustment sets. Every other set of nodes, that fulfills the Back-door criterion relative to the cause and the effect leads to unbiased estimates of the true causal effect.
In a second, blue curve I show the density of estimated causal effects that were not derived from valid adjustment sets. This density is not centered around the true causal effect. Since these values are derived from invalid adjustment sets, they can take any value and I do not know anything about their distribution. The Back-door criterion is not the only graphical criterion for valid adjustments (Tian and Pearl, 2002). Therefore, some of the values might still be valid, while others are not.
If the majority of estimated causal effects were valid, I would expect them to have similar values scattered around the true causal effect. Hence, I will observe a peak in the histogram of estimates.
These observations suggest to pool the effects across multisets and subsamples and to take the mode of the smoothed density as an estimate of the true causal effect. This idea forms the basis of my aIDA algorithm.

3.2 The aIDA algorithm

The aIDA algorithm takes a set of expression profiles consisting of p genes observed in n samples as input. The input data is purely observational, that means, that no perturbation experiments have been performed. All samples are assumed to be drawn from the same underlying joint distribution. The output is an ordered set of triples (X, Y, C), were X and Y are genes and C is the estimated causal effect of X on Y. The list of causal effects is sorted by the absolute value of C.

aIDA consists of the following steps (Taruttis et al., 2015):

1. Randomly draw K subsets of samples of size l (i.e.: $l = \frac{2}{3}n, \frac{n}{2}, ...$), resulting in K datasets.

2. For each of these subsets estimate a CPDAG using the PC-algorithm (Kalisch et al., 2012) with sparseness parameter α, resulting in K CPDAGs on the same set

of nodes.

3. For every ordered pair of genes $(X \to Y)$ estimate the multisets $M(X \to Y)$ of causal effects and pool them across all subsamples.

4. Generate one histogram of estimated effects per gene pair (Accumulation step). Smooth these histograms by a Gaussian kernel, detect the mode in the smoothed histogram and use it as an estimate for the causal effect C of X on Y.

5. Collect all causal effects in a $p \times p$ matrix. Sort the effects by the absolute value of C, and output this sorted list.

Step 1-3 are equal to the CStaR algorithm. But while CStaR takes the minimum absolute value as a lower bound of each multiset, aIDA uses all values from the multiset. Since the accumulation idea is added to the IDA concept, I call this algorithm accumulation based IDA (aIDA).

3.3 Results

In this section I will demonstrate that aIDA outperforms CStaR with respect to partial area under curve up to 100 false positives (pAUC(FP=100)) for both simulated datasets and two gene expression microarray datasets from *S. cerevisae*. Both aIDA and CStaR make use of the same implementation of the PC algorithm (Kalisch et al., 2012). For the evaluation of the algorithms I need a ground truth. For the simulated datasets the true causal effects are known from the simulation. In case of the *S. cerevisae* gene expression datasets the target set of causal effects is calculated as described in Maathuis et al. (2010) from an additional set of interventional data (see Section D.1 for details). Note that, since the target set is derived from noisy experimental data, it is not the set of true causal effects, but I expect an enrichment of causal interactions within that set.

3.3.1 Parameter calibration

The most critical parameter to calibrate, for both aIDA and CStaR is the sparseness parameter α of the PC algorithm. An increasing α leads to denser CPDAGs and, with that, to larger multisets. As a consequence, this leads to more regression coefficients, which might increase the standard error of the estimated causal effect. In a first approach I tested the recommended value for α from Maathuis et al. (2010) and Stekhoven et al. (2012). However I observed that for both simulated and real gene expression data, the estimated CPDAGs are too sparse in comparison to the graphs estimated in Maathuis

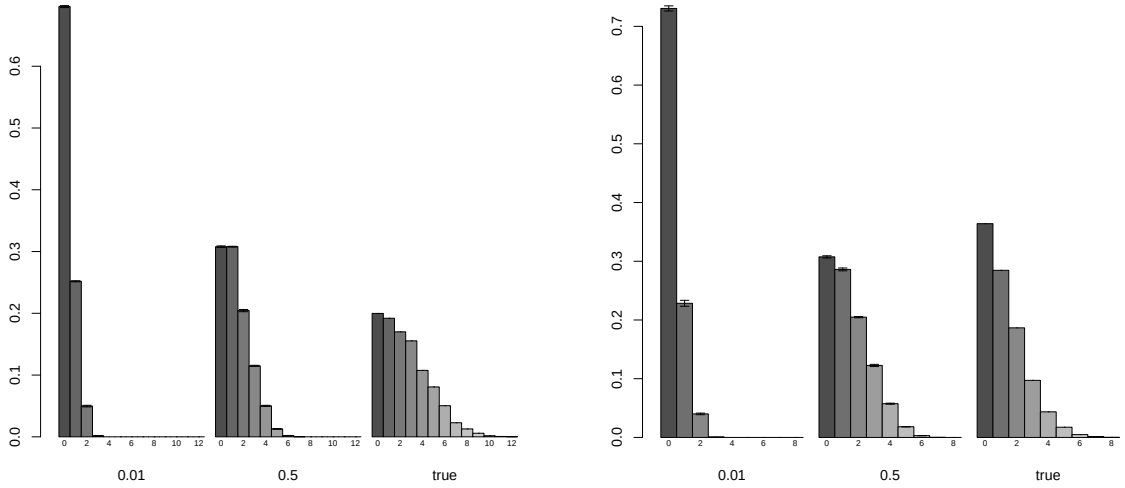
et al. (2010) and Stekhoven et al. (2012). These findings are underpinned by a correction of the PC algorithm to ensure an order-independent skeleton estimation which leads to sparser graphs (Colombo and Maathuis, 2014). I calibrated α for the simulated dataset such that I obtained CPDAGs with a density similar to the density of the true underlying CPDAGs. I found that $\alpha = 0.5$ represents a good choice of α (Figure 3.4). This also leads to a better performance in comparison to smaller values of α (Figure 3.7). Figure 3.5 compares the network densities to a density derived from Balaji et al. (2006) for the *S. cerevisae* datasets. For both values $\alpha = 0.01$ and $\alpha = 0.5$ the PC algorithm underestimates the density of the network. Thus, the data suggests an even higher value of α than 0.5. But due to runtime constraints I did not further increase α. To ensure a fair comparison α and the 100 subsampled datasets are the same for both aIDA and CStaR. And with that, both algorithms rely on the same set of estimated CPDAGs. All other CStaR parameters were set to the values recommended by Stekhoven et al. (2012) (see Section E for details). The bandwidth of the Gaussian Kernel was set to the default value of the density() function from basic R, since calibration did not improve the estimation of causal effects.

3.3.2 Performance on simulated datasets

I compared the performance of aIDA to the performance of CStaR on simulated datasets generated from known causal Gaussian Bayesian networks. A comparison to plain IDA is not necessary, since it has already been shown, that CStaR outperforms IDA with respect to true positive selections versus false positive selections (Stekhoven et al., 2012). For this comparison I simulated random graphs of size 100 and 1000 nodes, respectively. The edge weights were sampled from a uniform distribution on the interval (0.1,1). These weights represent the size of the direct causal effects between the nodes.

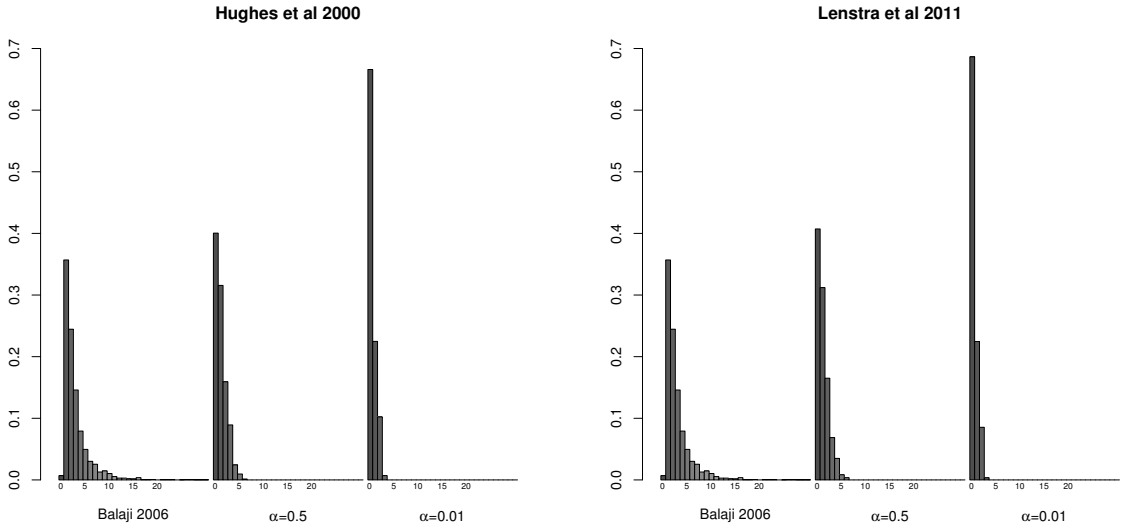
Performance on simulated datasets with 100 nodes First I simulated two sets of 10 datasets from small DAGs with 100 nodes and 50 or 1000 observations, respectively. I applied aIDA and CStaR on this 20 datasets using 100 subsamples of size $\frac{n}{2}$ and $\alpha = 0.1$. To measure the performance I calculated the partial areas under receiver operating characteristics (ROC) curves (pAUC) up to 100 false positives shown in Figure 3.6. The error bars correspond to standard errors across the 10 different datasets in each group. The barplots illustrate, that aIDA outperforms CStaR with respect to pAUC up to 100 false positives.

Performance on simulated datasets with 1000 nodes Causal biological networks are normally larger than 100 nodes. Since the PC algorithm becomes impractically slow



(a) dense **(b)** sparse

Figure 3.4: Distribution of the number of parents for different values of α and the true underlying DAG for the two sets of simulated datasets with 1000 nodes and 50 samples. The distribution of parents estimated using higher values for the parameter α is more similar to the true distribution of the number of parents.



(a) **(b)**

Figure 3.5: Distribution of the number of parents for different values of α and the true underlying DAG for the two *S. cerevisae* datasets. The number of parents estimated using two different values for α are compared to a the number of parents derived from the transcriptional regulatory network published by Balaji et al. (2006). The higher value α=0.5 leads to e better estimation of network density.

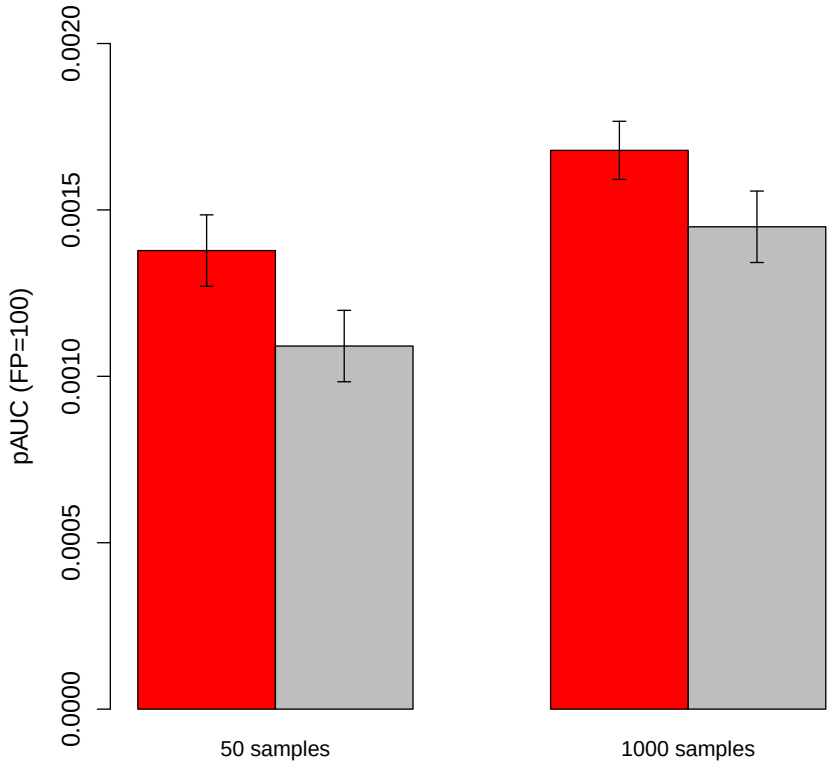


Figure 3.6: Comparison of the partial area under the ROC curve up to 100 false positives for 10 simulated datasets with 100 nodes, and n = 50 and n = 1000 samples. Red bars show values for aIDA, gray bars show values for CStaR. The error bars indicate standard errors across the 10 datasets. aIDA outperforms CStaR for both, 50 and 1000 samples with respect to partial area under the ROC curve up to 100 false positives.

for datasets with more than 1000 nodes and a realistic choice of α, I tested the performance of aIDA with random networks of 1000 nodes. To take several densities of the networks into account I created 5 sparse random networks with approximately 1250 edges and 5 more dense graphs with approximately 2500 edges. From each random DAG I simulated a dataset of 50 samples. Thereafter aIDA and CStaR were applied to the datasets to reestimate the causal effects from the purely observational data. The barplots in Figure 3.7 show the pAUC up to 100 false positives for the 5 datasets derived from the sparse and dense networks generated as described above. The tuning parameter of the PC algorithm was set to 4 different values. I achieve the best performance over all values of α for $\alpha = 0.5$ using the aIDA approach.

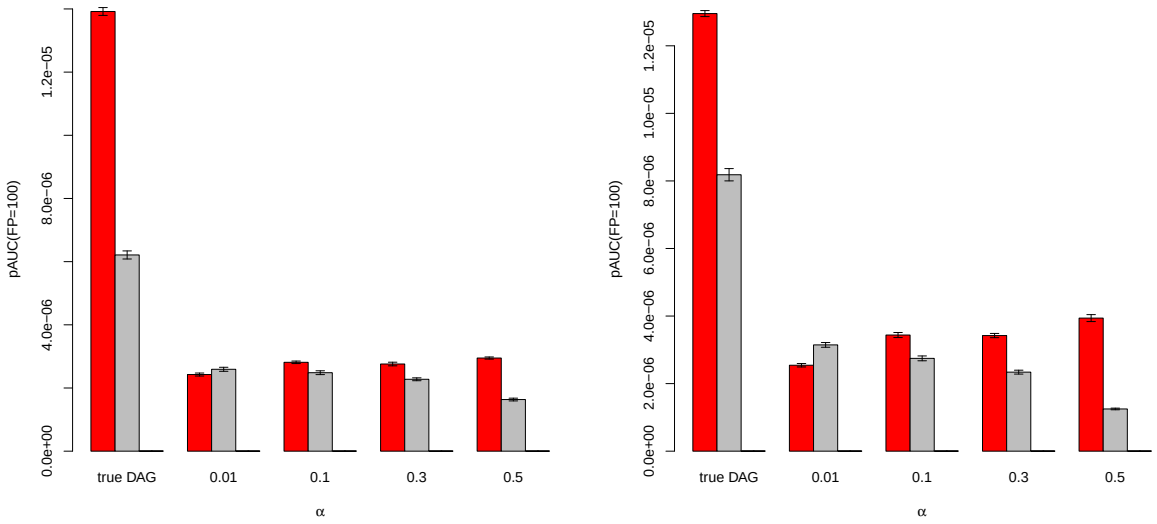
3.3.3 Application to gene expression data of *S.cerevisiae* deletion strains

To examine the performance of aIDA in comparison to CStaR on real world datasets I applied, both, aIDA and CStaR to two large scale yeast gene expression datasets. Both datasets consist of both purely observational data and expression data from deletion strains (Baudin et al., 1993; Wach, 1996) to define a gold standard target set. I estimated the causal effects on the purely observational gene expression data by aIDA and CStaR, respectively and validated my results using the associated interventional experiments.

Application to data from Hughes et al. (2000) The first dataset has already been analyzed by IDA (Maathuis et al., 2010) and CStaR (Stekhoven et al., 2012) to evaluate their performance. The observational data generated by Hughes et al. (2000) consists of 63 wild type samples and the interventional data of 276 deletion mutants. After preprocessing as described previously by Maathuis et al. (2010); Stekhoven et al. (2012), 234 single gene deletion strains remain as interventional dataset (see Section C for details). All data was measured on two-color cDNA microarrays with probes for 5361 genes.

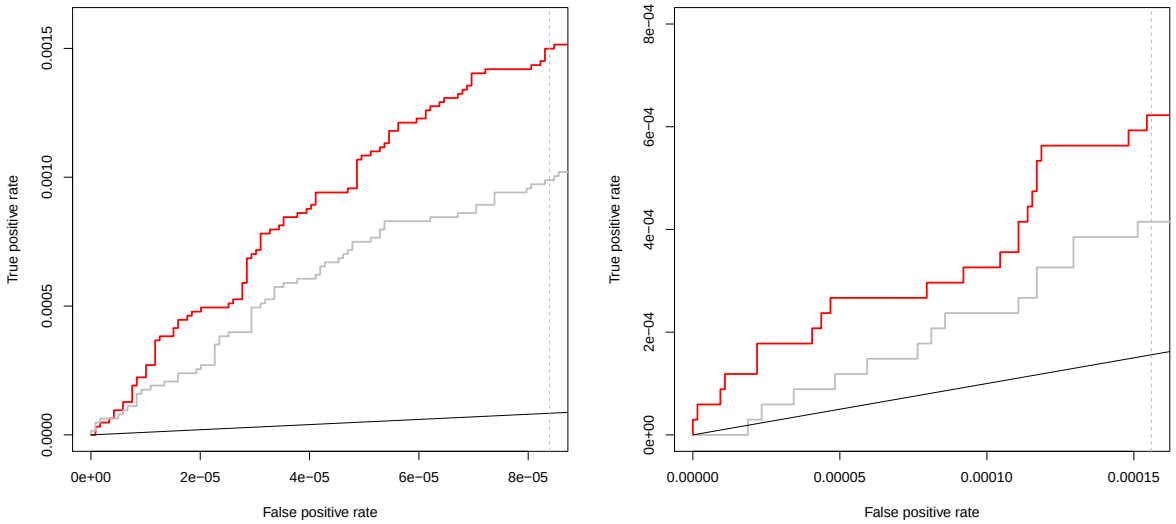
Both aIDA and CStaR were applied to the same 100 subsamples of the 63 observations. I estimated the multisets of causal effects $M(X \to Y)$ for all genes from the 234 gene deletion strains X and all measured 5361 genes Y. The interventional data acts as basis for a target set, which is used for the classification of these predictions. I calculated ROC curves up to 100 false positives for both algorithms, CStaR and aIDA. The curves in Figure 3.8a illustrate that aIDA again outperforms aIDA under the given measurement.

Application to data from Lenstra et al. (2011) To add a second biological dataset to my examinations I used the dataset from Lenstra et al. (2011). I am the first one who



(a) dense **(b)** sparse

Figure 3.7: Comparison of the partial area under the ROC curve up to 100 false positives for the two sets of simulated datasets with 1000 nodes using different values of α and using the true underlying networks. Red bars show values for aIDA, gray bars show values for CStaR. The error bars indicate standard errors across the 5 datasets. aIDA outperforms CStaR for $\alpha \in$(0.1,0.3,0.5) and yields the best results for α =0.5 on both, sparse and dense datasets. Further aIDA shows a better performance, when the true underlying network is known.



(a) Hughes et al. (2000)

(b) Lenstra et al. (2011)

Figure 3.8: ROC curves for the two *S. cerevisae* datasets up to 100 FP. The red curve represents the ROC curve for aIDA, the gray curve shows the ROC curve for CStaR, and the black line refers to random guessing. aIDA improves over CStaR for both datasets. aIDA performs better than random guessing, while CStaR in case of the dataset from Lenstra et al. (2011) at the beginning of the ranked list does not.

used this dataset for causal discoveries (Taruttis et al., 2015). Lenstra et al. (2011) examined the interactions between chromatin and gene expression in *S.cerevisiae* by analyzing mutants of chromatin machinery components. After preprocessing (Section C) the interventional data consists of 138 gene expression profiles from single-gene deletion mutants and 67 observations from wild types. All data was measured on two-color cDNA microarrays with probes for 4890 genes. I found that on this dataset causal discovery is possible. Again, aIDA outperforms CStaR for the ROC curves up to 100 false positives (Figure 3.8b).

3.4 Discussion and conclusions

I introduced aIDA, a method to estimate causal effects from observational data without any knowledge about the causal network underlying the data. aIDA is well-suited for datasets with many variables but only few observations, which is a common situation in biology, by embedding a subsampling strategy around the IDA approach from Maathuis et al. (2009).

In contrast to previous approaches (Maathuis et al., 2009, 2010; Stekhoven et al., 2012), aIDA uses the whole multisets of causal effects from K subsampling runs. My estimate

of the causal effect is the mode of the density calculated over the whole K multisets of causal effects.

The estimation of causal Bayesian networks assumes that the multivariate Gaussian distribution is faithful to the DAG (Kalisch and Bühlmann, 2007), which means that statistical conditional independence can be inferred from the underlying DAG. But faithfulness in general is not testable (Zhang and Spirtes, 2008) and unfaithfulness of the population cannot be ruled out in biological systems which often try to maintain a stable equilibrium state (Andersen, 2013). Furthermore feedback mechanisms cannot be captured by a DAG. These limitations lead to the fact that IDA and thus Accumulation selection cannot replace wet lab experiments, but can be a good tool for experiment design. Hence, here it is very important to make good predictions at the top of the ordered list of absolute causal effects, because this would be my most promising candidates for future experiments.

I compared aIDA to CStaR (Stekhoven et al., 2012), a method which wrapped a stability selection (Meinshausen and Bühlmann, 2010) around a subsampling strategy. aIDA outperforms CStaR on both simulated and real world data sets from S. cerevisiae with respect to partial area under receiver operating characteristics (ROC) curve up to 100 FP, and thus outperforms plain IDA (Stekhoven et al., 2012).

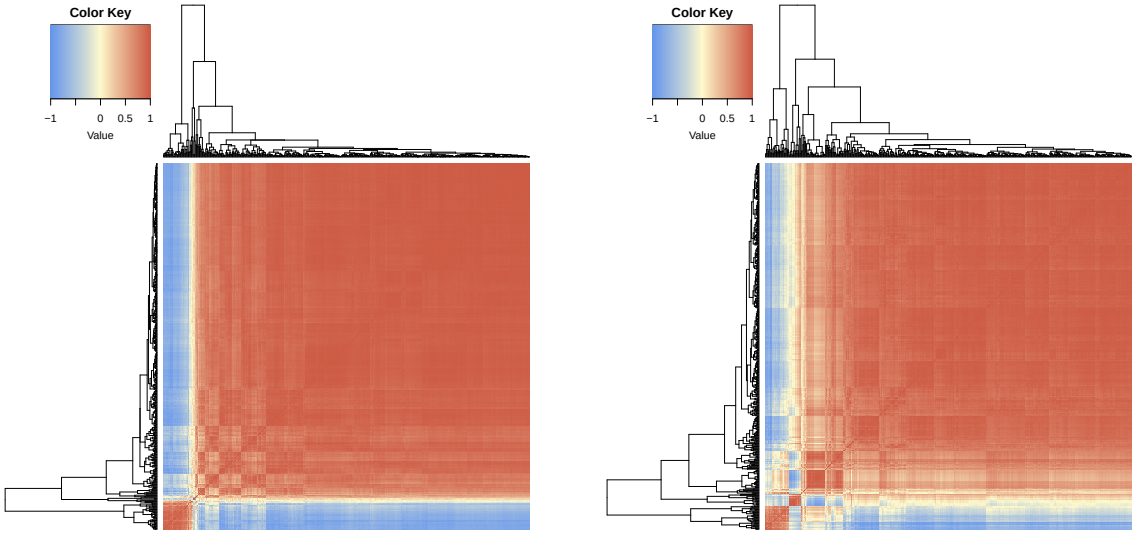
CHAPTER 4

Estimation of causal effects from highly correlated data

4.1 Motivation

The application of spike-in calibration from Section 2.4 to a huge RNA-seq gene expression dataset may result in highly correlated data. Figure 4.1 shows the correlation between the 500 most variable genes for two RNA-seq datasets of the P493-6 cell line. The P493-6 cell line is a B cell lymphoma cell line that allows for an ectopic induction of MYC by a tetracycline-controlled transcriptional activation (Gossen and Bujard, 1992; Polack et al., 1996). The data set consists of 50 samples of P493-6 "MYC-low" cells (tetracycline added, ectopic *MYC* is repressed) and 50 samples of P493-6 "MYC-high" cells (no tetracycline added, *MYC* is expressed). Some samples are treated with combinations of 5 different external stiumli (see Section A for details on data generation). I applied spike-in normalization using DESeq2 size factors (Love et al., 2014) to the raw counts (Taruttis et al., 2017). For a more detailed description on data generation I refer to Section 5. We see that nearly every gene is highly correlated with all other observed genes for both levels of *MYC*. This finding is a consequence of the global role of *MYC* in gene expression (Lin et al., 2012; Nie et al., 2012). Thus, we observe no technical artifact but a real biological process.

The estimation of causal effects from observational data requires the estimation of a causal graph (Pearl, 2003; Maathuis et al., 2009). However, for the data sets shown in Figure 4.1 we expect that the underlying DAG is dense, since every gene has a high probability to be connected to many other genes. This is a violation of the sparseness assumption of the PC algorithm (Kalisch and Bühlmann, 2007) which is a common assumption in

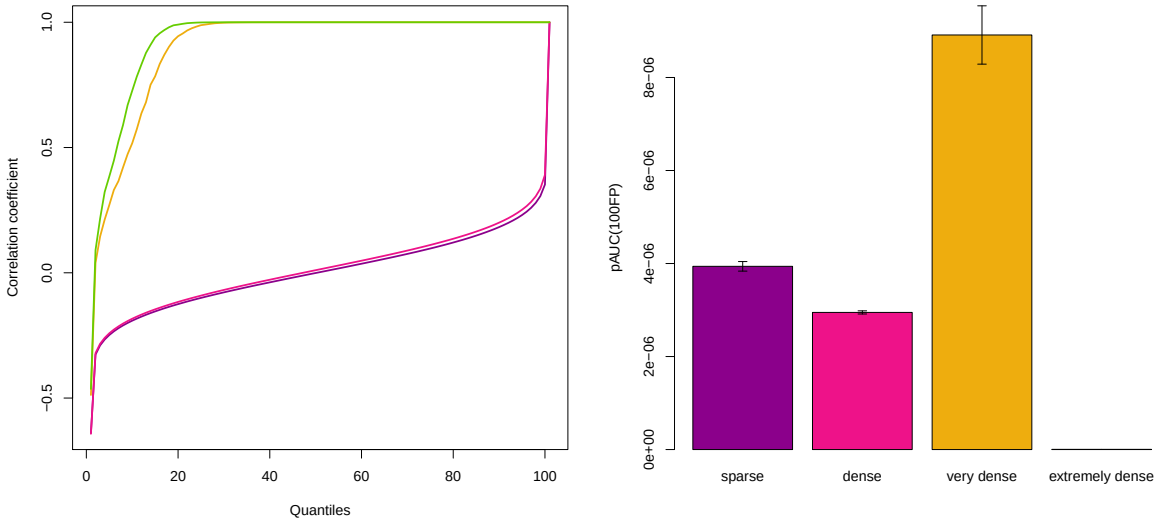


(a) Myc-low **(b)** Myc-high

Figure 4.1: Correlation between the 500 most variable genes for the P493-6 gene expression datasets after spike-in calibration Application of *Drosophila melanogaster* spike-in calibration results in highly correlated genes for both levels of *MYC*.

causal structure learning (Daly et al., 2011). And in fact this violation leads to a poor performance with respect to partial area under receiver operating characteristics curve (pAUCROC) (see Figure 4.2, see Section C for details on data generation).

However, Tsamardinos et al. (2006) claim that their MMHC algorithm requires significantly less tests, so that they perform less test errors and, thus, achieve a better network. I will make use of this advantage by replacing the PC algorithm within the IDA approach by a modification of the MMHC algorithm. Furthermore the performance of the MMHC algorithm does not decrease as fast as the performance of PC algorithm does, when the available sample size is relatively small and MMHC is faster than the PC algorithm (Tsamardinos et al., 2006). These statements suggest the use of MMHC algorithm not only for highly correlated data, but also for sparse networks.
The MMHC algorithm consists of two steps. In the first step, the skeleton is estimated by the Max-Min Parents and Children (MMPC) algorithm (Tsamardinos et al., 2003). After that a Bayesian-scoring hill-climbing algorithm orients the undirected edges. Nägele et al. (2007) described an extension of the MMHC algorithm which makes the algorithm applicable to thousands of variables: Instead of estimating the whole DAG, Nägele et al. (2007) estimate the Markov Blankets around every variable using the MMHC algorithm. This allows for the parallelization of the algorithm and ensures an increase in speed.



(a) **(b)**

Figure 4.2: Relation between performance of aIDA and correlation structure between variables. Datasets consist of 1000 nodes and 50 samples. The sparse (approximately 1250 edges per graph) and the dense (approximately 2500 edges per graph) dataset are the datasets described in Section 3.3.2. The very dense datasets consist of approximately 20000 edges and the extremely dense datasets consist of approximately 50000. Extremely dense graphs result in a worse performance with respect to partial area under ROC curve up to 100 false positives. (a) Quantile plots of correlation coefficients between variables of the sparse (purple curve), dense (pink curve), very dense (orange curve) and extremely dense (green curve) datasets. As expected denser graphs result in higher correlation coefficients between variables. (b) Barplots of the partial area under ROC curve up to 100 false positives for the four sets of datasets (”sparse”, ”dense”, ”very dense”,”extremely dense”). The tuning parameter of the PC algorithm was set to α =0.5, which is more optimal for dense graphs.

After that a "feature partial directed graph" (fPDAG) (Dejori et al., 2005) summarizes the Markov Blankets. With a fPDAG it is possible to describe uncertainties in Bayesian network structure such that to each pair of nodes it assigns a probability for the existence of an edge and its orientation.
However, for the estimation of causal effects the whole CPDAG is not required, but the parents of the causes. Thus, for this purpose it is sufficient to estimate the Markov Blankets around causes without constructing a fPDAG. Driven by these ideas I developed the MMHC-aIDA algorithm, which uses the MMHC algorithm to estimate the Markov Blankets around the causes of interest, instead of estimating the whole CPDAG.

4.2 MMHC-aIDA algorithm

The MMHC-aIDA algorithm takes a set of expression profiles consisting of p genes observed in n samples as input. The input data is purely observational, that means, that no perturbations experiments have been performed. All samples are assumed to be drawn from the same underlying joint distribution. The output is an ordered set of triples (X, Y, C), were X and Y are genes and C is the estimated causal effect of X on Y. I call the number of causes of interest N_x with $1 \leq N_x \leq p$. The list of causal effects is sorted by the absolute value of C.

1. Randomly draw K subsets of samples of size l (i.e.: $l = \frac{2}{3}n, \frac{n}{2}, ...$), resulting in K resampled datasets.

2. For each of these subsets and for each cause X estimate a Markov Blanket around X, resulting in $N_x \times K$ Markov Blankets

3. For every ordered pair of genes $(X \rightarrow Y)$ estimate the multisets $M(X \rightarrow Y)$ of causal effects and pool them across all subsamples.

4. Generate one histogram of estimated effects per gene pair (Accumulation step). Smooth these histograms by a Gaussian kernel, detect the mode in the smoothed histogram and use it as an estimate for the causal effect C of X on Y.

5. Collect all causal effects in a $p \times p$ matrix. Sort the effects by the absolute value of C, and output this sorted list.

Steps 1 and 3-5 are similar to the aIDA algorithm presented in Section 3. MMHC-aIDA only differs in the way how the parents of the causes for the estimation of causal effects are calculated from the observational data (step 2).
Since the MMHC algorithm is used to estimate the Markov Blankets around the causes and with that the parents of the causes X, I call this algorithm MMHC-aIDA.

4.3 Results

4.3.1 Performance on simulated datasets

I tested the performance of MMHC-aIDA with random networks of 1000 nodes. To take several densities of the networks into account I created 5 sparse random networks with approximately 1250 edges, 5 dense graphs with approximately 2500 edges, 5 very dense graphs with approximately 20000 edges and 5 extremely dense graphs with approximately 50000 edges. From each random DAG I simulated a dataset of 50 samples. Thereafter aIDA, CStaR (Stekhoven et al., 2012) and MMHC-aIDA were applied to the 20 datasets (4 levels of sparseness $\times$ 5 datasets for each level of sparseness) to reestimate the causal effects from the purely observational data. The barplots in Figure 4.3 show the pAUC up to 100 false positives for the 5 datasets derived from the sparse and the dense networks generated as described above. The tuning parameter of the PC algorithm was set to $\alpha = 0.5$. From Section 3 we already know that aIDA outperforms CStaR on the partial area under curve up to 100 false positives. And also MMHC-aIDA perform better than CStaR for the two network densities. However, for the sparse and the dense network aIDA also outperforms MMHC-aIDA. What about networks with much higher density and as a consequence of that highly correlated data? The MMHC algorithm claims to improve over the PC algorithm due to less test errors. Does only the change of the network estimation algorithm improve the performance aIDA on more dense networks? Figure 4.4 shows barplots of the pAUC up to 100 false positives for the 5 datasets derived from the very dense and the extremely dense networks. For the very dense networks MMHC-aIDA outperforms aIDA, but for the extremely dense networks the performance of both algorithms drastically breaks down. This underpins the assumption that the MMHC algorithm is more suitable for highly correlated data, and MMHC-aIDA should be the method of choice in those cases.

4.4 Discussion

I introduced MMHC-aIDA for the estimation of causal effects from highly correlated data. High correlations between genes are a result of the *Drosophila melanogaster* spike-in calibration and are a consequence of *MYC*'s global role in gene expression of B cells (Lin et al., 2012; Nie et al., 2012).
MMHC-aIDA did not improve over aIDA for more sparse datasets, but for very dense data sets. The performance for both, MMHC-aIDA and aIDA, breaks down for extremely correlated data. Unfortunately, to my knowledge, no highly correlated biological datasets, that consist of both an observational dataset with at least approximately 50 samples and

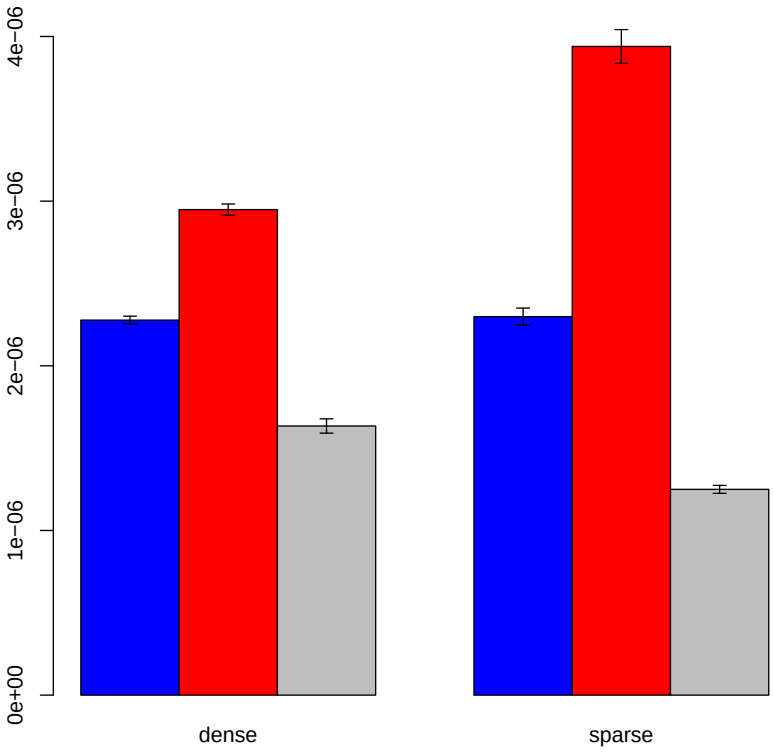


Figure 4.3: Comparison of the partial area under the ROC curve up to 100 false positives for the two sets of simulated sparse and dense datasets with 1000 nodes for α=0.5. Blue bars show values for MMHC-aIDA, red bars show values for aIDA, and gray bars show values for CStaR. The error bars indicate standard errors across the 5 datasets. Both aIDA and MMHC-aIDA outperform CStaR. However aIDA shows the best performance.

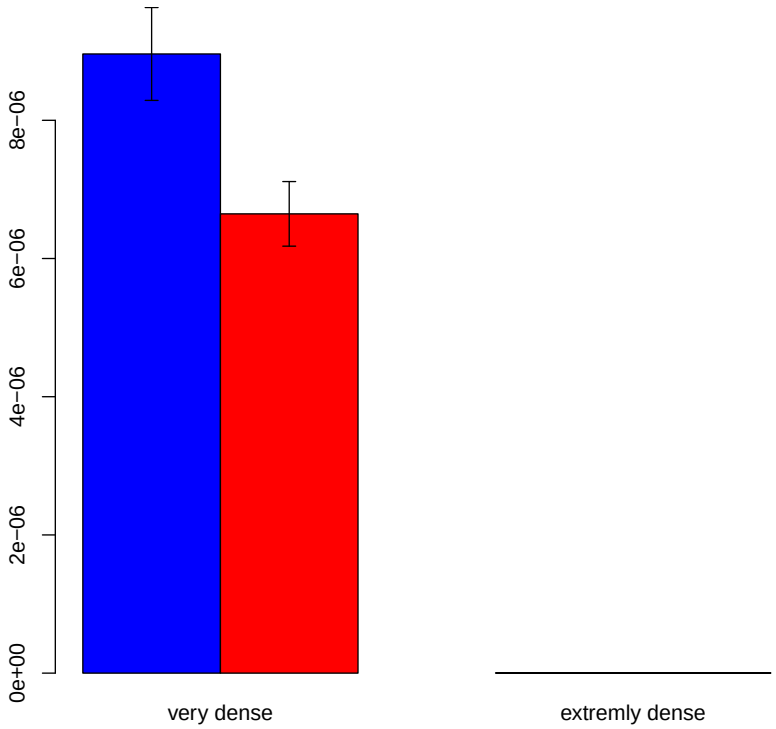


Figure 4.4: Comparison of the partial area under the ROC curve up to 100 false positives for the two sets of simulated very dense and extremely dense datasets with 1000 nodes for α=0.5. Blue bars show values for MMHC-aIDA, red bars show values for aIDA. The error bars indicate standard errors across the 5 datasets. For the very dense data MMHC-aIDA outperforms aIDA, while for the extremely dense data the performance of both algorithms is poor.

a huge set of interventional data exists. Thus, an evaluation of performance on highly correlated biological data is impossible so far.
Estimating causal effects from extremely high correlated data is a difficult task and could not be solved by the MMHC-aIDA algorithm. However, he MMHC-aIDA algorithm performs better on highly correlated data than aIDA does. In summary there is more research is required for that particular problem. A starting point could be the work of (Mandozzi and Bühlmann, 2016a) and (Mandozzi and Bühlmann, 2016b).

Part II

Causal analysis of *MYC*-dependent gene expression and cell metabolism of a B cell lymphoma cell line

CHAPTER 5

Experimental setup and data preparation

5.1 Experimental setup

To examine the causal interactions of the gene *MYC* with the transcriptome and the metabolome under different *MYC* expression levels we selected the P493-6 cell line, a model organism for cell cycle activation by *MYC* in lymphoma cells. The P493-6 cell line allows to examine the causal relationships of *MYC* under the same epigenetic conditions. For this purpose the cells contain a doxycycline depended promoter to switch *MYC* on or off. Untreated P493-6 cells are highly proliferative due to a strong overexpression of *MYC* ("*MYC*-high"). The treatment with doxycycline hampers the expression of *MYC* ("*MYC*-low") by the inducible Tet-Off expression system (Gossen and Bujard, 1992; Polack et al., 1996). However the "*MYC*-low" cells may also proliferate slowly in presence of doxycycline and estradiol via a viral EBNA2-ER fusion protein (Yustein et al., 2010; Sabo et al., 2014).

The causal interactions of *MYC* with the transcriptome and the metabolome of the P493-6 cells are inferred from "*MYC*-high" and "*MYC*-low" state of the cell line. The experimental setup consists of 100 samples. 50 are in "*MYC*-high" and 50 in "*MYC*-low" state. The majority of samples is treated with combinations of five different stimuli in two different dosages to stimulate many different genetic and metabolic pathways. The activated pathways play a central role in B-cell maturation, germinal center reactions and lymphomagenesis. Anti human IgM F(ab)2 fragment (α-IgM) activates the BCR signaling pathway. sCD40L (CD40) treatment results in TNF (tumor necrosis factor) activation. rhIGF-1 (IGF) activates insulin growth factor 1 and rhIL-10 (IL10) activates

interleukin 10. ODN2006 (CpG) positively stimulates the toll like receptor via the dinucleotide CpG. The dosage levels are treatment dependent (see Table A.1 for details).
The experiment was designed in 10 batches (days) with 10 probes per batch. To reduce batch effects, we decided upon a special experimental design. Each batch contains two untreated control samples, one in "*MYC*-high" and the other in "*MYC*-low" state. Further we included one sample for each *MYC* state with a single stimulation in the highest dosage. The remaining six samples per batch are left for random combinations with at least two different stimuli and random dosages (see Table 5.1 for a scheme of the experimental design). The five different stimuli activate pathways that play a central

	Batch	CD40L	BCR	IGF	CpG	IL10	Myc		Batch	CD40L	BCR	IGF	CpG	IL10	Myc
1	1	1.0	0.0	0.0	0.0	0.0	H	51	6	1.0	0.0	0.0	0.0	0.0	H
2	1	0.0	1.0	0.0	0.0	0.0	L	52	6	0.0	1.0	0.0	0.0	0.0	L
3	1	0.0	0.0	0.0	0.0	0.0	L	53	6	0.0	0.0	0.0	0.0	0.0	L
4	1	0.0	0.0	0.0	0.0	0.0	H	54	6	0.0	0.0	0.0	0.0	0.0	H
5	1	0.0	1.0	1.0	1.0	1.0	H	55	6	0.0	0.0	0.0	0.2	1.0	H
6	1	1.0	1.0	1.0	1.0	1.0	H	56	6	0.2	0.2	0.2	1.0	1.0	H
7	1	1.0	1.0	0.0	1.0	0.0	H	57	6	0.2	1.0	1.0	0.0	0.0	H
8	1	1.0	0.0	0.0	1.0	1.0	L	58	6	1.0	1.0	0.0	1.0	1.0	L
9	1	0.2	1.0	1.0	0.0	1.0	L	59	6	1.0	1.0	1.0	1.0	1.0	L
10	1	0.0	0.0	0.0	1.0	1.0	L	60	6	0.2	0.0	1.0	0.2	1.0	L
11	2	0.0	1.0	0.0	0.0	0.0	H	61	7	0.0	1.0	0.0	0.0	0.0	H
12	2	0.0	0.0	1.0	0.0	0.0	L	62	7	0.0	0.0	1.0	0.0	0.0	L
13	2	0.0	0.0	0.0	0.0	0.0	L	63	7	0.0	0.0	0.0	0.0	0.0	L
14	2	0.0	0.0	0.0	0.0	0.0	H	64	7	0.0	0.0	0.0	0.0	0.0	H
15	2	0.0	1.0	0.2	0.2	0.0	H	65	7	0.2	1.0	1.0	0.0	1.0	H
16	2	1.0	0.2	0.2	0.0	0.0	H	66	7	1.0	0.2	0.0	0.0	1.0	H
17	2	1.0	1.0	1.0	1.0	0.2	H	67	7	1.0	1.0	0.0	1.0	1.0	H
18	2	0.2	0.2	0.2	0.2	0.2	L	68	7	0.0	0.2	0.0	0.2	1.0	L
19	2	1.0	0.2	0.0	0.0	1.0	L	69	7	1.0	1.0	1.0	1.0	0.2	L
20	2	0.2	1.0	0.2	0.0	0.2	L	70	7	1.0	1.0	0.0	1.0	0.0	L
21	3	0.0	0.0	1.0	0.0	0.0	H	71	8	0.0	0.0	1.0	0.0	0.0	H
22	3	0.0	0.0	0.0	1.0	0.0	L	72	8	0.0	0.0	0.0	1.0	0.0	L
23	3	0.0	0.0	0.0	0.0	0.0	L	73	8	0.0	0.0	0.0	0.0	0.0	L
24	3	0.0	0.0	0.0	0.0	0.0	H	74	8	0.0	0.0	0.0	0.0	0.0	H
25	3	0.2	0.2	0.2	0.0	0.0	H	75	8	0.0	0.2	0.0	0.2	1.0	H
26	3	0.2	1.0	0.0	0.0	1.0	H	76	8	0.2	0.2	0.2	0.2	0.2	H
27	3	1.0	0.0	0.0	1.0	1.0	H	77	8	0.2	1.0	0.2	0.0	0.2	H
28	3	0.2	1.0	0.2	1.0	0.0	L	78	8	0.0	1.0	1.0	1.0	1.0	L
29	3	0.0	0.0	0.0	0.2	1.0	L	79	8	0.2	0.2	0.0	1.0	1.0	L
30	3	0.2	1.0	0.2	0.2	0.2	L	80	8	0.2	0.2	0.2	1.0	1.0	L
31	4	0.0	0.0	0.0	1.0	0.0	H	81	9	0.0	0.0	0.0	1.0	0.0	H
32	4	0.0	0.0	0.0	0.0	1.0	L	82	9	0.0	0.0	0.0	0.0	1.0	L
33	4	0.0	0.0	0.0	0.0	0.0	L	83	9	0.0	0.0	0.0	0.0	0.0	L
34	4	0.0	0.0	0.0	0.0	0.0	H	84	9	0.0	0.0	0.0	0.0	0.0	H
35	4	0.2	1.0	0.2	0.2	0.2	H	85	9	0.2	1.0	1.0	0.2	0.0	H
36	4	1.0	1.0	0.2	0.2	0.0	H	86	9	0.2	0.2	1.0	0.0	0.2	H
37	4	0.2	0.0	1.0	0.2	0.0	H	87	9	1.0	0.0	1.0	1.0	0.2	H
38	4	1.0	1.0	0.2	0.2	0.0	L	88	9	0.0	1.0	0.2	0.2	0.0	L
39	4	0.2	1.0	1.0	0.2	0.0	L	89	9	0.2	0.2	1.0	0.0	0.2	L
40	4	1.0	0.0	1.0	1.0	0.2	L	90	9	1.0	0.0	1.0	0.2	0.0	L
41	5	0.0	0.0	0.0	0.0	1.0	H	91	10	0.0	0.0	0.0	0.0	1.0	H
42	5	1.0	0.0	0.0	0.0	0.0	L	92	10	1.0	0.0	0.0	0.0	0.0	L
43	5	0.0	0.0	0.0	0.0	0.0	L	93	10	0.0	0.0	0.0	0.0	0.0	L
44	5	0.0	0.0	0.0	0.0	0.0	H	94	10	0.0	0.0	0.0	0.0	0.0	H
45	5	0.0	0.0	0.0	1.0	1.0	H	95	10	1.0	0.0	1.0	0.2	0.0	H
46	5	0.2	0.2	0.0	1.0	1.0	H	96	10	0.2	1.0	0.2	1.0	0.0	H
47	5	0.2	0.0	0.2	1.0	0.0	H	97	10	0.2	0.0	1.0	0.2	1.0	H
48	5	0.2	0.2	0.2	0.0	0.0	L	98	10	0.2	1.0	1.0	0.0	0.0	L
49	5	0.2	1.0	0.0	0.0	1.0	L	99	10	0.2	0.0	1.0	0.2	0.0	L
50	5	0.2	0.0	0.2	1.0	0.0	L	100	10	1.0	0.2	0.2	0.0	0.0	L

Table 5.1: Experimental design of the 100 samples of P493-6 cells. The P493-6 cells were treated with combinations of 5 different stimuli and 2 dosage levels. 1 refers to full dosage, 0 refers to no treatment and 0.2 refers to reduced dosage. H refers to "'*MYC*-high"' cells and L refers to "'*MYC*-low"' cells. For details on data generation see Section A and B.

role in B-cell maturation, germinal center reactions and lymphomagenesis. Anti human IgM F(ab)2 fragment (α-IgM) activates the BCR signaling pathway. sCD40L (CD40)

treatment results in TNF (tumor necrosis factor) activation. rhIGF-1 (IGF) activates insulin growth factor 1 and rhIL-10 (IL10) activates interleukin 10. ODN2006 (CpG) positively stimulates the toll like receptor via the dinucleotide CpG. The dosage levels are treatment dependent (see Table A.1 for details).
For each of the 100 samples we measured both, the transcriptome and the metabolome of the cells. The transcriptome of the P493-6 cells is measured by RNA-Seq analysis. For data normalization we added both the synthetical ERCC spike-in kit and *Drosophila melanogaster* spike-in cells (see Section 2 for details). The raw sequence data is available at NCBI, Bioproject PRJNA312050. Maren Feist (Department of Haematology and Medical Oncology of the University Medical Center Göttingen) carried out all wet lab experiments. RNA-seq was done by Dr. Gabriela Salinas-Riester (Head of Core Microarray and Deep-Sequencing Core Facility, University Medical Center Göttingen) (see Section A for a detailed description of the experimental setup). The Mass spectrometry analysis allows to measure the metabolic processes of the cells. Both, cell pellets and supernatants of the 100 samples are measured to create a consistent view of the metabolic processes. Philipp Schwarzfischer (Group NMR Spectroscopy, Institute of Functional Genomics, University of Regensburg) provided the Mass spectrometry analysis for both cell pellets and supernatants of the 100 samples (see Section B for details).

5.2 Data normalization

5.2.1 Gene expression data

For the normalization of RNA seq data I constructed a genome of human and *Drosophila melanogaster* reference genomes and the ERCC sequences. The human reference was GRCh38 from ensembl, release 77 (Cunningham et al., 2015), the Drosophila melanogaster reference was ensembl BDGP5, release 77 and the ERCC sequences were as provided by https://tools.lifetechnologies.com/content/sfs/manuals/cms 095047.txt. I mapped all sequence libraries against this concatenated genome using TopHat version 2.0.13 (Trapnell et al., 2009) indicating an unstranded sequencing protocol (–library -type fr-unstranded) and default settings for the remaining parameters. I assigned read counts to ensembl gene IDs using featureCounts version 1.4.5 (Liao et al., 2014). To prepare the data for causal analysis I calculated transcripts per million (TPM) values for all ensembl gene IDs (Wagner et al., 2012). TPMs are normalized by a loess regression on the *Drosophila melanogaster* spike-in genes as suggested by Lovén et al. (2012) and Taruttis et al. (2017). The subsequent log2-transformation avoids to blow up values by scaling due to very small variances during data selection. After that I calculated the median for each ensembl ID

over the 50 samples in "MYC-high" and the 50 samples in "MYC-low" state. For each data set I selected the top 50% of genes according to the median, which resulted in 7900 ensembl IDs per dataset. Then I calculated the interquartile range on the remaining ensembl IDs and selected the 5000 genes with the highest interquartile range for the 50 samples in "MYC-high" and the 50 samples in "MYC-low" state (Figures 5.1 and 5.2).

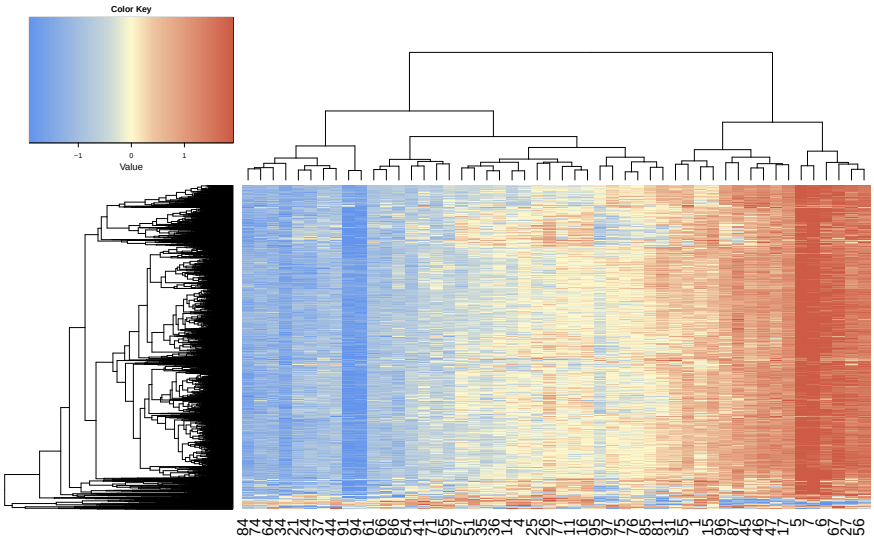


Figure 5.1: Log2-transformed TPM values of the 50 samples and the 5000 selected genes in "*MYC*-high" state On the top 50% genes according to the median the 5000 genes with the highest interquartile range have been selected.

5.2.2 Metabolomics data

The metabolomics cell pellets data requires no further preparation or normalization. The metabolomics supernatant data has been divided by growth factors and measured medium as provided and suggested by Philipp Schwarzfischer (Group NMR Spectroscopy, Institute of Functional Genomics, University of Regensburg). Neither the supernatant data set nor the pellet data set show indications to batch effects (see Figures 5.3 and 5.4). To ensure data comparability to the gene expression data set the metabolomics data was log-transformed and scaled for causal inference analysis.

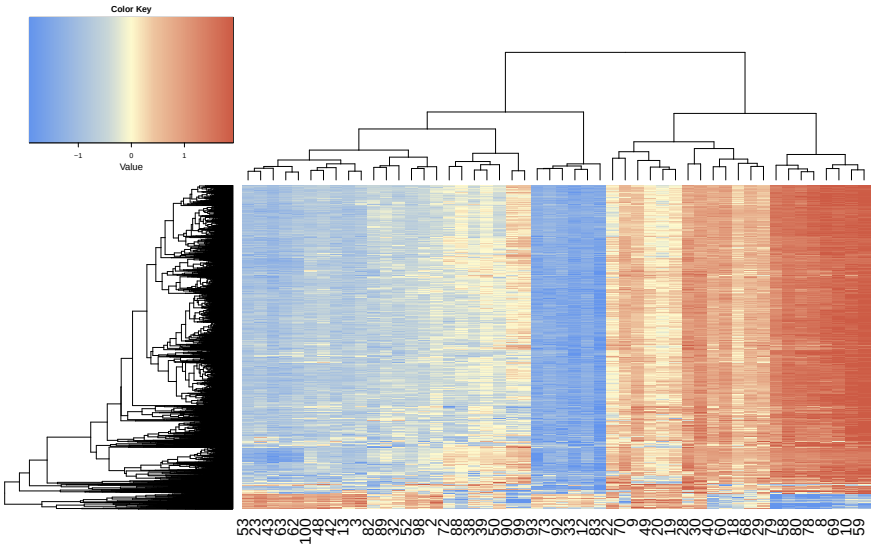


Figure 5.2: Log2-transformed TPM values of the 50 samples and the 5000 selected genes in "*MYC*-low" state On the top 50% genes according to the median the 5000 genes with the highest interquartile range have been selected.

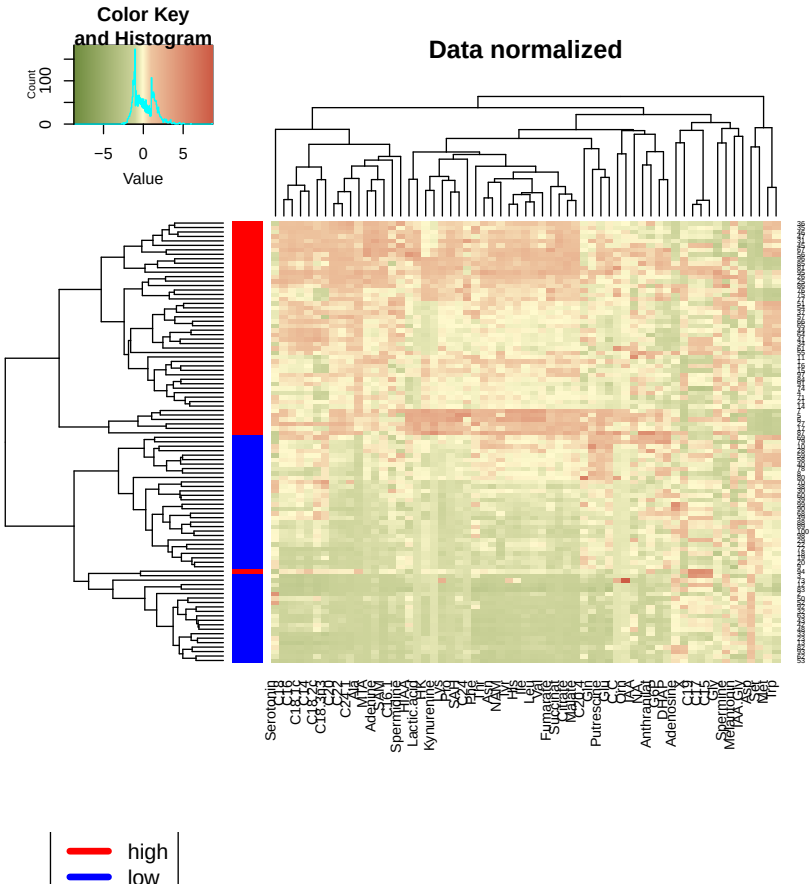


Figure 5.3: Pellet data set received from P493-6 samples The samples in "*MYC*-low" and "*MYC*-high" state are colored in blue and red, respectively. Batch effects are not observed.

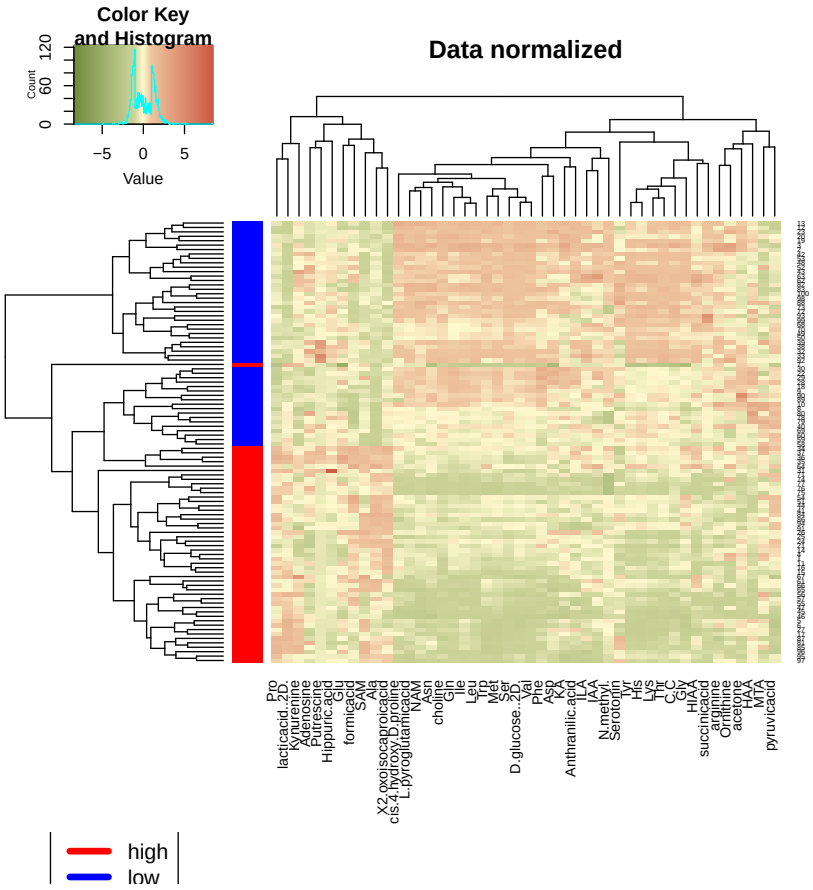


Figure 5.4: Supernatant data set received from P493-6 samples The samples in "*MYC*-low" and "*MYC*-high" state are colored in blue and red, respectively. Batch effects are not observed.

CHAPTER 6

Causal inference analysis

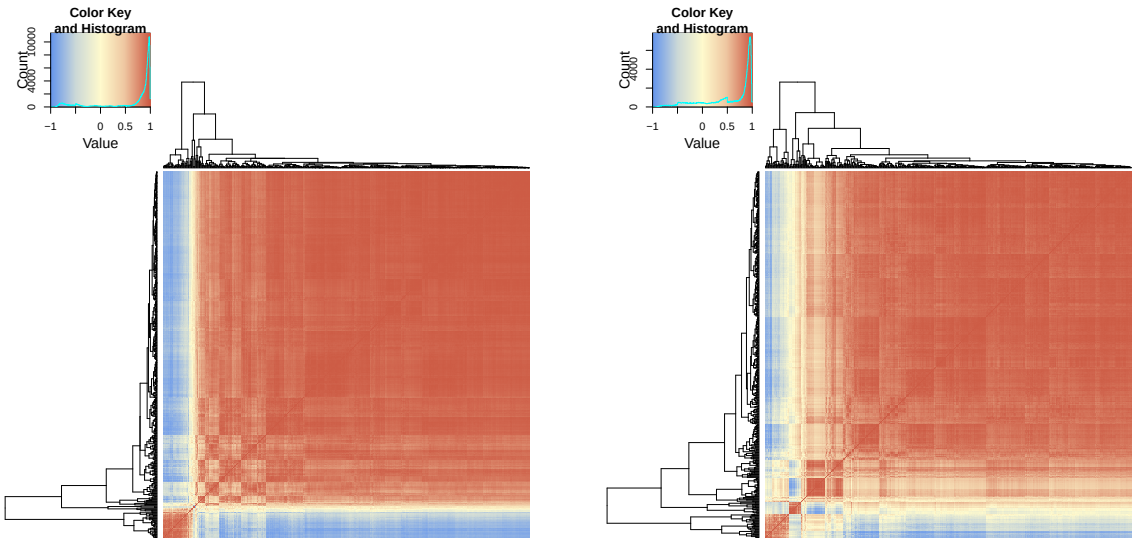
6.1 The Estimation of causal structures depends on the correlation pattern of the samples

6.1.1 Gene expression data

The normalized gene expression data is a highly correlated data set (see Figure 6.1). Following Chapter 4, I applied MMHC-aIDA with tuning parameter $\alpha = 0.5$ instead of plain aIDA to the data set. To underpin both the selection of the MMHC-aIDA algorithm and the selection of the α-parameter I compared the density of the estimated networks to the B cell interactome data provided by Lefebvre et al. (2007). For the comparison I used the interactions from the data set, which have been reported in public databases. In comparison to aIDA the mean number of parents estimated by MMHC-aIDA in the estimated CPDAGs is more reliable (Figure 6.2).

6.1.2 Metabolomics data

I want to determine both the causal effects of the measured metabolites from supernatants and cell pellets on *MYC* and the causal effects of *MYC* on the metabolites measured in supernatants and cell pellets under the two different MYC levels. Since the metabolomics dataset is not highly correlated (Figure 6.3), the aIDA-method (Chapter 3) is the method of choice for the estimation of the causal effects. For the calibration of the α-parameter of the PC algorithm I compared the distribution over the number of parents of the estimated



(a) Myc-low **(b)** Myc-high

Figure 6.1: Correlation between the 5000 most variable genes for the P493-6 gene expression datasets after spike-in calibration Application of *Drosophila melanogaster* spike-in calibration results in highly correlated genes for both levels of *MYC*.

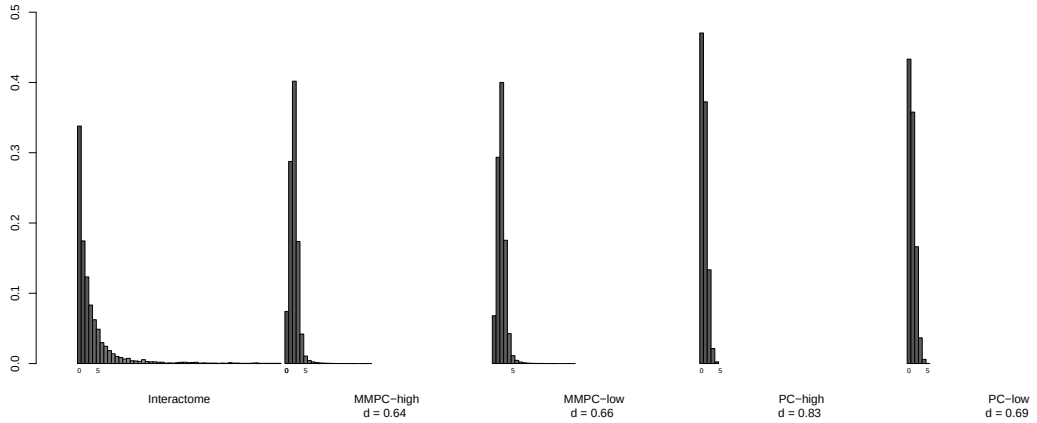
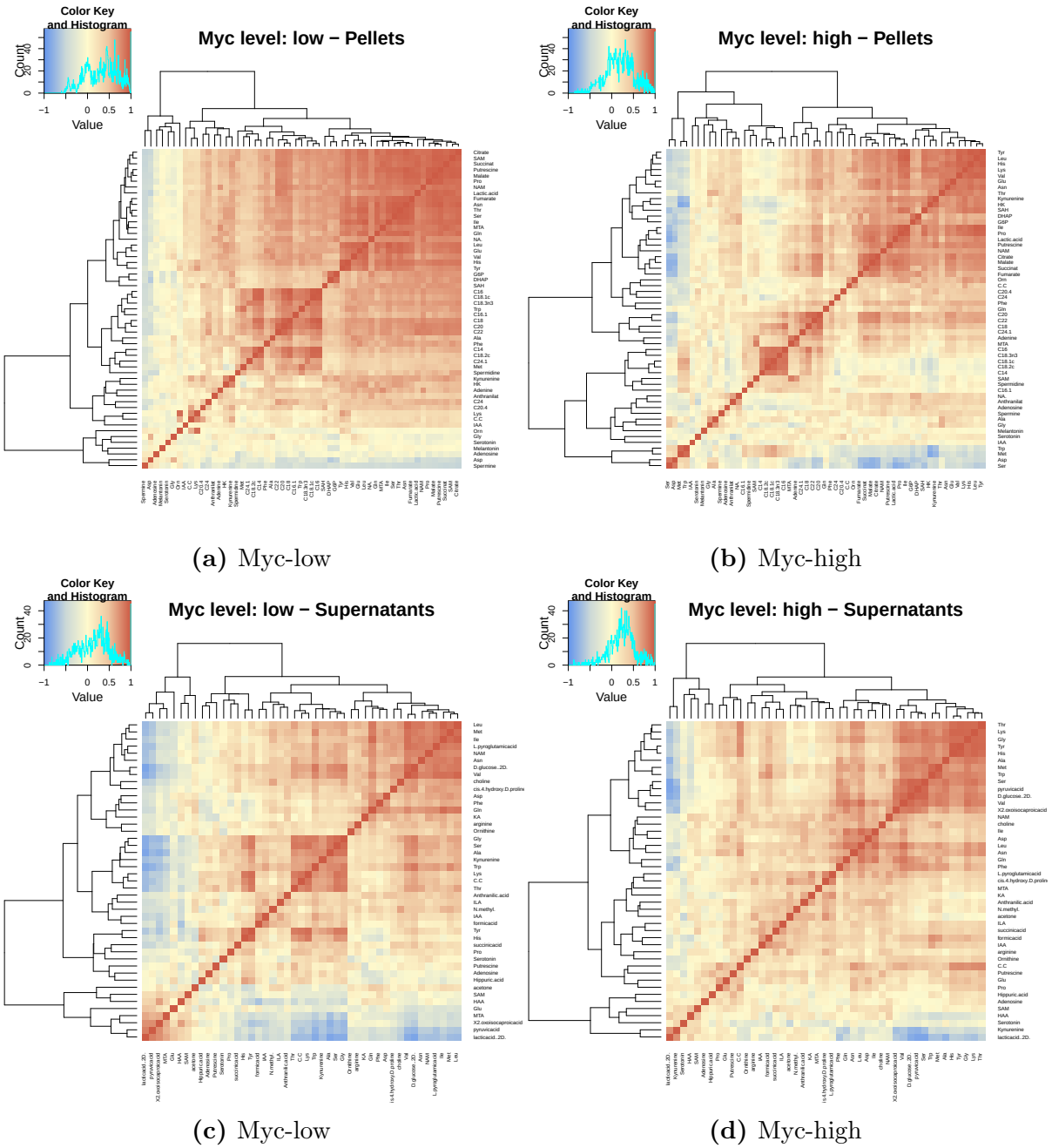


Figure 6.2: The number of parents estimated by PC algorithm and the MMHC algorithm for α=0.5 are compared to a the number of parents derived from the transcriptional B cell interactome network. d denotes the Euclidean distance between the parental distribution of the particular estimated network and the B cell interactome network. The MMHC algorithm results in CPDAGs which show a more similar distribution of parents in comparison to the B cell interactome data set than the networks estimated by PC algorithm.

networks to the distribution of the metabolic pathway network of the KEGG database (Kanehisa et al., 2002). Figure 6.4 shows the distribution of the number of parents for both supernatants and cell pellets under the different *MYC*-levels. The Euclidean distances between the distributions of the different α values and the distribution of the KEGG metabolic pathway network indicate to α=0.1 (see Table 6.1).



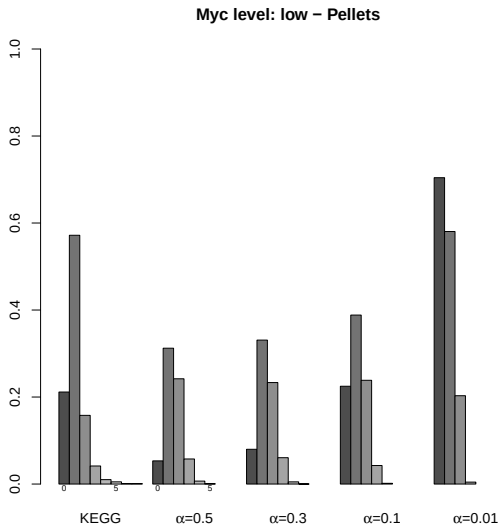
(a) Myc-low **(b)** Myc-high

(c) Myc-low **(d)** Myc-high

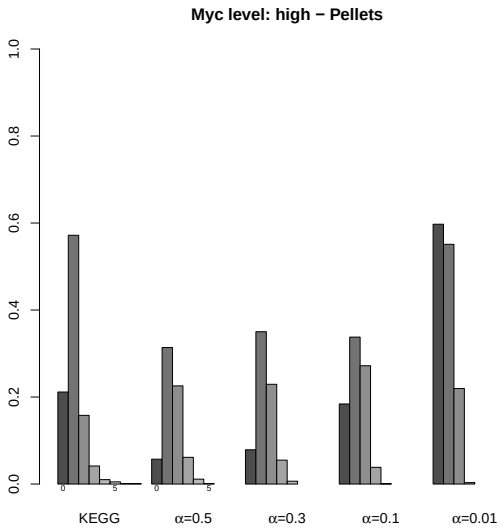
Figure 6.3: Correlation between the measured metabolites of the P493-6 metabolomics datasets after scaling and log-transformation The metabolomics datasets are not highly correlated.

6.2 Causal relationships between the transcriptome and *MYC*

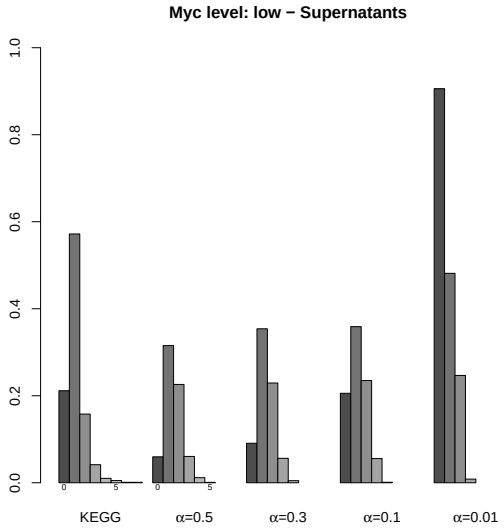
Following Section 6.1 I applied the MMHC algorithm with α=0.5 to the RNA-seqdata set. For each gene in the data set I estimated 100 Markov patterns from 25 random



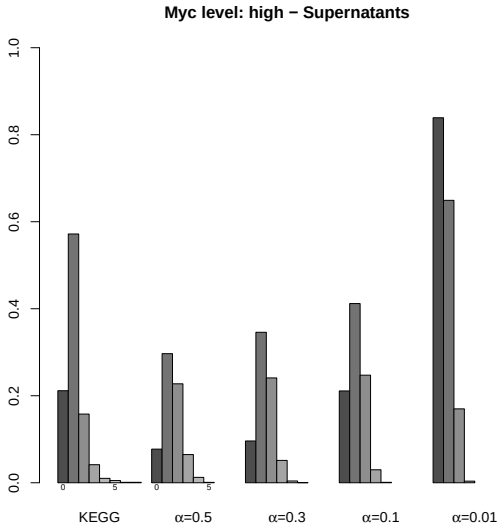
(a) Myc-low



(b) Myc-high



(c) Myc-low



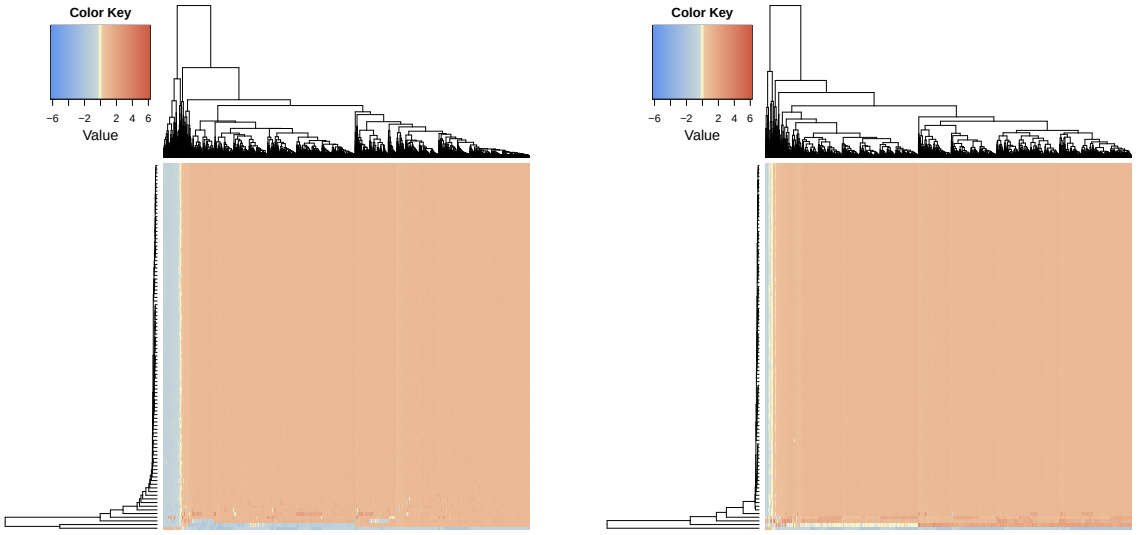
(d) Myc-high

Figure 6.4: Distribution of the number of parents for different values of and the metabolic pathway network of the KEGG database for the pellet and the supernatant datasets under different *MYC* conditions. The distribution of parents estimated using higher values for the parameter α is more similar to the true distribution of the number of parents.

α	Pellets data set		Supernatants data set	
	Myc-high	Myc-low	Myc-high	Myc-low
0.5	0.36	0.36	0.36	0.35
0.3	0.34	0.33	0.31	0.33
0.1	0.33	0.25	0.23	0.29
0.01	0.56	0.70	0.90	1.00

Table 6.1: Euclidean distances between the distribution over the number of parents of the metabolic pathway data set from KEGG and CPDAGs estimated by the PC algorithm for different values of α and the two MYC conditions. For the supernatants and pellets data set and for both MYC states the distance between the parental distribution of the KEGG pathway and the estimated CPDAG with α=0.1 is the smallest. Thus, for further analysis I chose α=0.1 to estimate the causal metabolic networks.

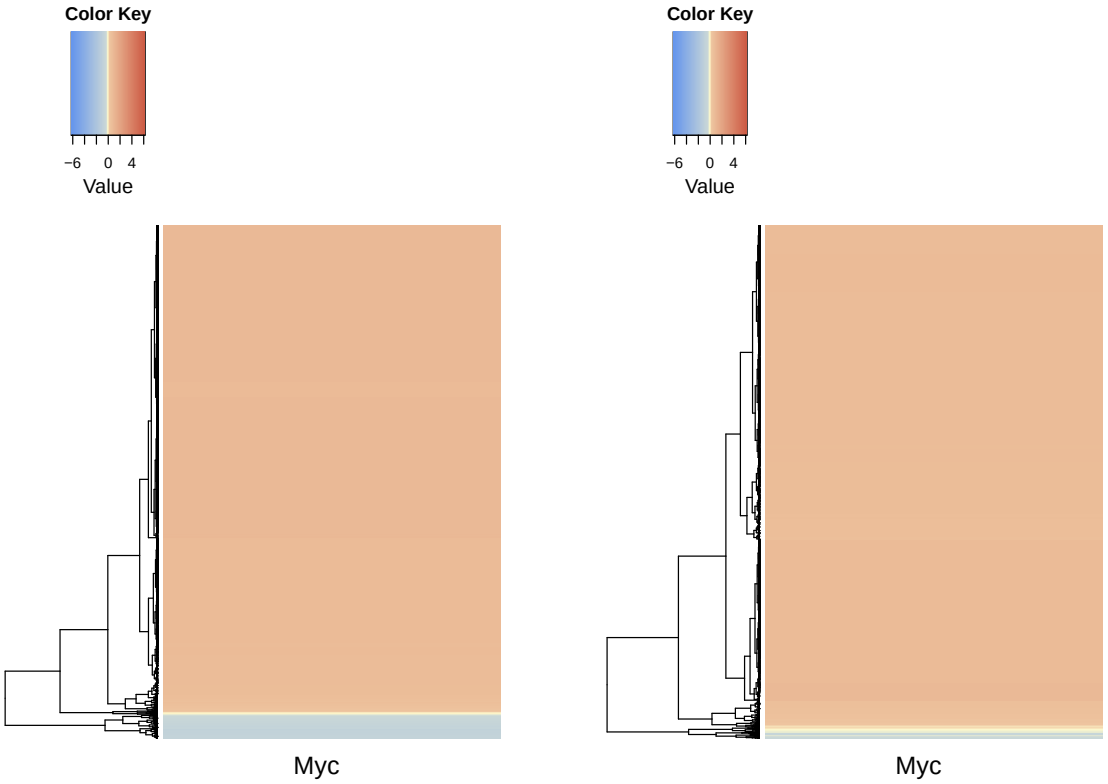
subsamples for each of the two levels of MYC "MYC-low" and "MYC-high". I estimated the multisets of causal effects for all genes from the 5000 genes on all 5000 genes for both the "MYC-low" and "MYC-high" data set. From these multisets I determined the causal effect of one gene on another by MMHC-aIDA (Section 4.4). For each gene I calculated the mean of absolute causal effects of this gene on the 4999 remaining genes over the multi sets estimated from the 100 subsamples. Figure 6.5 shows the top 100 causal regulators for "MYC-low" and the "MYC-high" according to the mean absolute causal effect over the 100 subsamples. Figure 6.6 shows the causal effects of MYC on



(a) Myc-low **(b)** Myc-high

Figure 6.5: Top 100 causal regulators for the "*MYC*-low" and the "*MYC*-high" dataset The causal effects are estimated by the MMHC-aIDA algorithm.

the remaining 4999 genes. *MYC* has positive causal effects on the majority of the genes. This finding is consistent with the hypothesis that *MYC* is a transcriptional amplifier which upregulates the majority genes. The 20 highest absolute causal effects of the 4999



(a) Myc-low **(b)** Myc-high

Figure 6.6: Causal effects of *MYC* on the 4999 remaining genes for the "*MYC*-low" and the "*MYC*-high" dataset The causal effects are estimated by the MMHC-aIDA algorithm.

genes on *MYC* for both data sets are summarized in Table 6.2. The causal inference analysis shows that these 20 genes have the largest effects on *MYC*. The interpretation of the results is the following: If the causal effect of gene X on *MYC* is positive the upregulation of gene X in the virtual intervention experiment leads to an upregulation of *MYC*, while a negative causal effect of gene X on *MYC* means that an upregulation of gene X causes a repression of *MYC*.

Furthermore I summarize the 20 genes with the most negative causal effects on *MYC* for the "*MYC*-high" and the "*MYC*-low" dataset in Table 6.3. These are the most important *MYC* repressor genes in the virtual intervention experiment. Some of these interactions are already known (Figure 6.7). For example *TCF3* upregulates *MYC* in the virtual intervention effects. Mutations of *TCF3* increase the activity of the PI3K

Ensembl gene id	External gene name	Effect ("high")	Ensembl gene id	External gene name	Effect ("low")
ENSG00000234975	FTH1P2	2.27	ENSG00000127589	TUBBP1	2.32
ENSG00000219507	FTH1P8	-1.60	ENSG00000113161	HMGCR	2.31
ENSG00000175886	RPL7AP66	1.59	ENSG00000051341	POLQ	-2.18
ENSG00000080824	HSP90AA1	-1.46	ENSG00000228502	EEF1A1P11	2.00
ENSG00000168827	GFM1	-1.14	ENSG00000131747	TOP2A	-1.80
ENSG00000107331	ABCA2	1.14	ENSG00000219507	FTH1P8	-1.62
ENSG00000188873	RPL10AP2	1.13	ENSG00000214110	LDHAP4	1.58
ENSG00000065183	WDR3	1.10	ENSG00000137310	TCF19	-1.51
ENSG00000160285	LSS	1.00	ENSG00000095139	ARCN1	1.45
ENSG00000214110	LDHAP4	1.00	ENSG00000143228	NUF2	-1.43
ENSG00000134061	CD180	-0.99	ENSG00000179967	PPP1R14BP3	1.36
ENSG00000145911	N4BP3	0.97	ENSG00000092199	HNRNPC	1.07
ENSG00000132153	DHX30	0.95	ENSG00000182774	RPS17	1.03
ENSG00000101938	CHRDL1	0.94	ENSG00000165071	TMEM71	-1.01
ENSG00000105409	ATP1A3	0.94	ENSG00000142937	RPS8	1.01
ENSG00000121057	AKAP1	0.94	ENSG00000132341	RAN	1.00
ENSG00000138617	PARP16	0.94	ENSG00000169251	NMD3	1.00
ENSG00000071564	TCF3	0.94	ENSG00000228205	RP11-778D9.4	1.00
ENSG00000004975	DVL2	0.94	ENSG00000127022	CANX	1.00
ENSG00000143674	MLK4	0.93	ENSG00000166441	RPL27A	0.99

Table 6.2: The 20 highest absolute causal effects on *MYC* for the *MYC*-"high" and the *MYC*-"low" dataset. The causal effects were estimated by the MMPC-algorithm with α =0.5.

pathway in Burkitt's Lymphoma (Sewastianik et al., 2014), which also influences *MYC*-induced proliferation (Walsh et al., 2009). Furthermore, a recent publication of Wei et al. (2020) shows that *TCF3* activates *MYC* in neuroblastoma.

DVL2 increases the expression of *MYC* in the virtual experiment. Smalley et al. (2005) showed that *DVL2* plays an important role in *WNT* signaling and *MYC* is a target gene of this pathway (He et al., 1998). *HSP90AA1* acts as a repressor of *MYC* in the causal inference analysis. Chakravorty et al. (2017) listed *HSP90AA1* as a high potential candidate involved in *MYC* regulation. *GFM1* and *DHX30* are also listed as results of the virtual intervention experiment and are known target genes of *MYC*, which are involved in mitochondrial protein biosynthesis (Morrish and Hockenbery, 2014; Seitz et al., 2011; Zeller et al., 2006; Li et al., 2005). Teater and Melnick (2017) assume a relationship between *MYC* and *KLHL14*, which is identified as potential *MYC* repressor by the causal inference analysis. Further *KLHL14* mutations in ABC DLBCLs are associated with a poor prognosis.

Besides the top 20 causal effects some of the mechanisms described in Section 1.1.2 are observed in the causal inference results: For example the causal effect of *MYC* on *AKT*

Ensembl gene id	External gene name	Effect ("high")	Ensembl gene id	External gene name	Effect ("low")
ENSG00000219507	FTH1P8	-1.60	ENSG00000051341	POLQ	-2.18
ENSG00000080824	HSP90AA1	-1.46	ENSG00000131747	TOP2A	-1.80
ENSG00000168827	GFM1	-1.14	ENSG00000219507	FTH1P8	-1.62
ENSG00000134061	CD180	-0.99	ENSG00000137310	TCF19	-1.51
ENSG00000020181	GPR124	-0.80	ENSG00000143228	NUF2	-1.43
ENSG00000228232	GAPDHP1	-0.73	ENSG00000165071	TMEM71	-1.01
ENSG00000116704	SLC35D1	-0.73	ENSG00000136573	BLK	-0.93
ENSG00000132465	IGJ	-0.62	ENSG00000275395	FCGBP	-0.93
ENSG00000213763	ACTBP2	-0.59	ENSG00000263264	CTB-133G6.1	-0.89
ENSG00000135451	TROAP	-0.58	ENSG00000234184	RP5-887A10.1	-0.89
ENSG00000234184	RP5-887A10.1	-0.57	ENSG00000227507	LTB	-0.88
ENSG00000189057	FAM111B	-0.56	ENSG00000134697	GNL2	-0.88
ENSG00000128218	VPREB3	-0.53	ENSG00000128218	VPREB3	-0.85
ENSG00000197629	MPEG1	-0.53	ENSG00000185862	EVI2B	-0.85
ENSG00000149212	SESN3	-0.51	ENSG00000277448	RP11-538C21.2	-0.85
ENSG00000257221	RP11-689B22.2	-0.51	ENSG00000121807	CCR2	-0.84
ENSG00000133321	RARRES3	-0.50	ENSG00000104894	CD37	-0.84
ENSG00000125046	SSUH2	-0.50	ENSG00000144645	OSBPL10	-0.84
ENSG00000211978	IGHV5-78	-0.50	ENSG00000162892	IL24	-0.83
ENSG00000197705	KLHL14	-0.49	ENSG00000162894	FAIM3	-0.82

Table 6.3: The 20 most negative causal effects on *MYC* for the *MYC*-"high" and the *MYC*-"low" dataset. The causal effects were estimated by the MMPC-algorithm with α =0.5.

and *MTOR* is positive for both *MYC*-state datasets in the virtual experiment. Olive et al. (2009) and Xiao et al. (2008) showed that *MYC* inhibits *PTEN* via the micro-RNA miR-19. *PTEN* is an inhibitor of PI3K and PI3K activates *AKT*. In the "MYC-high" state the causal effect of *MYC* on *RAPTOR* is positive, too (Table 6.4). This positive causal effect may be interpreted as an activation of the PI3K-AKT-mTOR pathway by *MYC* which results in proliferation.

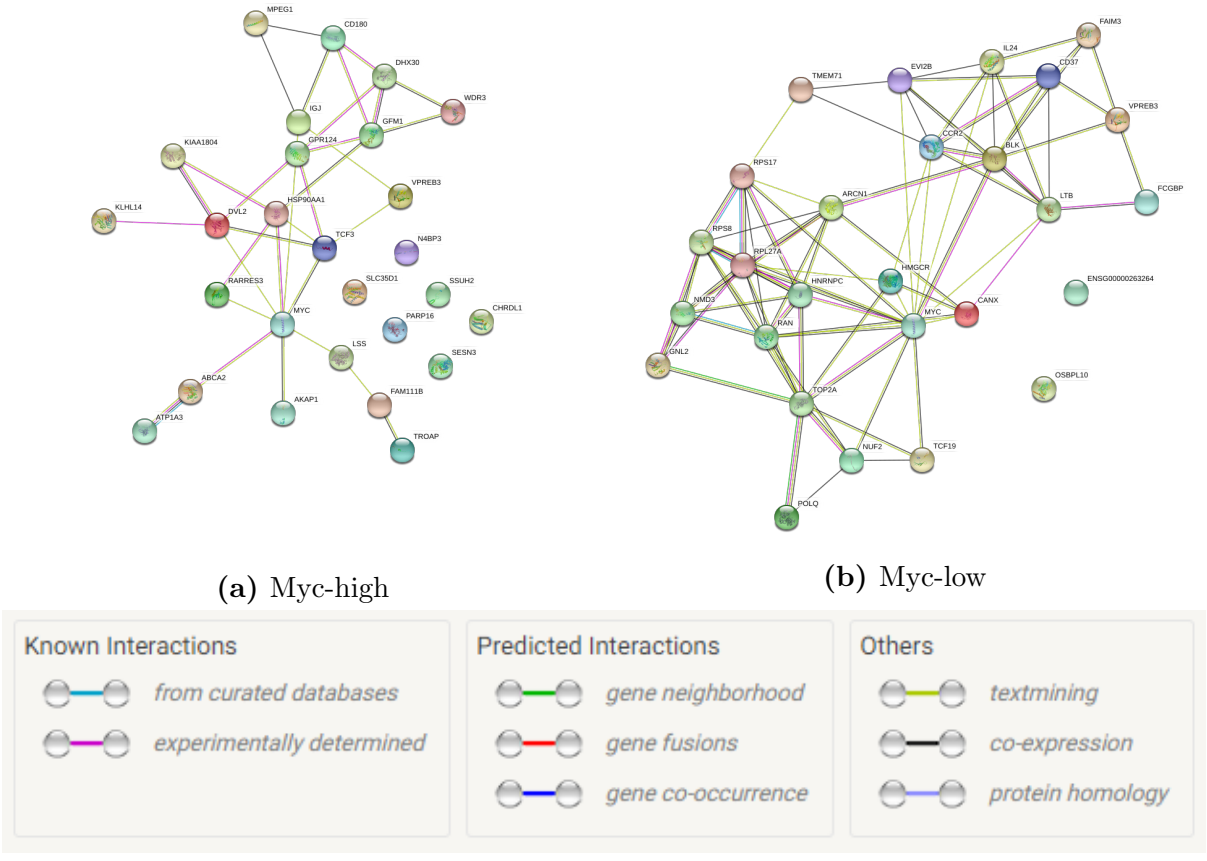
MYC also enables cells to have unlimited replicative potential. One mechanism is the direct activation of *TERT*. Indeed, the causal effect of *MYC* on *TERT* is positive.

VEGF plays an important role in the induction of angiogenesis. From Section 1.1.2 we know that *MYC* upregulates *VEGF*. In the virtual experiment I find that the causal effect of *MYC* on *VEGF* is positive.

In summary the application of the presented methods for causal inference discovers some well known causal connections. This demonstrates that causal inference with MMHC-aIDA is able to recover true causal relationships from purely observational data. Further genes with high causal effects on MYC are high priority candidates for experimental validation.

6.3 Causal relationships between the metabolome and *MYC*

Following Section 6.1 I applied the PC algorithm with α=0.1 to the metabolomics dataset. I estimated 100 CPDAGs from 25 random subsamples for each of the two levels of *MYC* "*MYC*-low" and "*MYC*-high". I estimated the multisets of causal effects of the metabolites on all 5000 genes for both the "*MYC*-low" and "*MYC*-high" datasets and I determined the causal effects of *MYC* on all metabolites using aIDA (Taruttis et al., 2015). Figure 6.8 shows the causal effects of the metabolites on the 5000 genes for "*MYC*-low" and the "*MYC*-high" for the pellet and the supernatant dataset. Figure 6.9 shows the causal effects of the metabolites from the pellet and supernatant datasets on *MYC* and Figure 6.10 shows the causal effects of *MYC* on the metabolites in the pellet and the supernatant datasets. *MYC* has a positive causal effect on the majority of the metabolites



(a) Myc-high

(b) Myc-low

(c) Legend

Figure 6.7: Protein-protein interactions of the 20 genes with the highest absolute and the 20 most negative causal effects on *MYC* . The protein-protein interactions are derived from STRING database (Szklarczyk et al., 2015) with low confidence interaction score threshold.

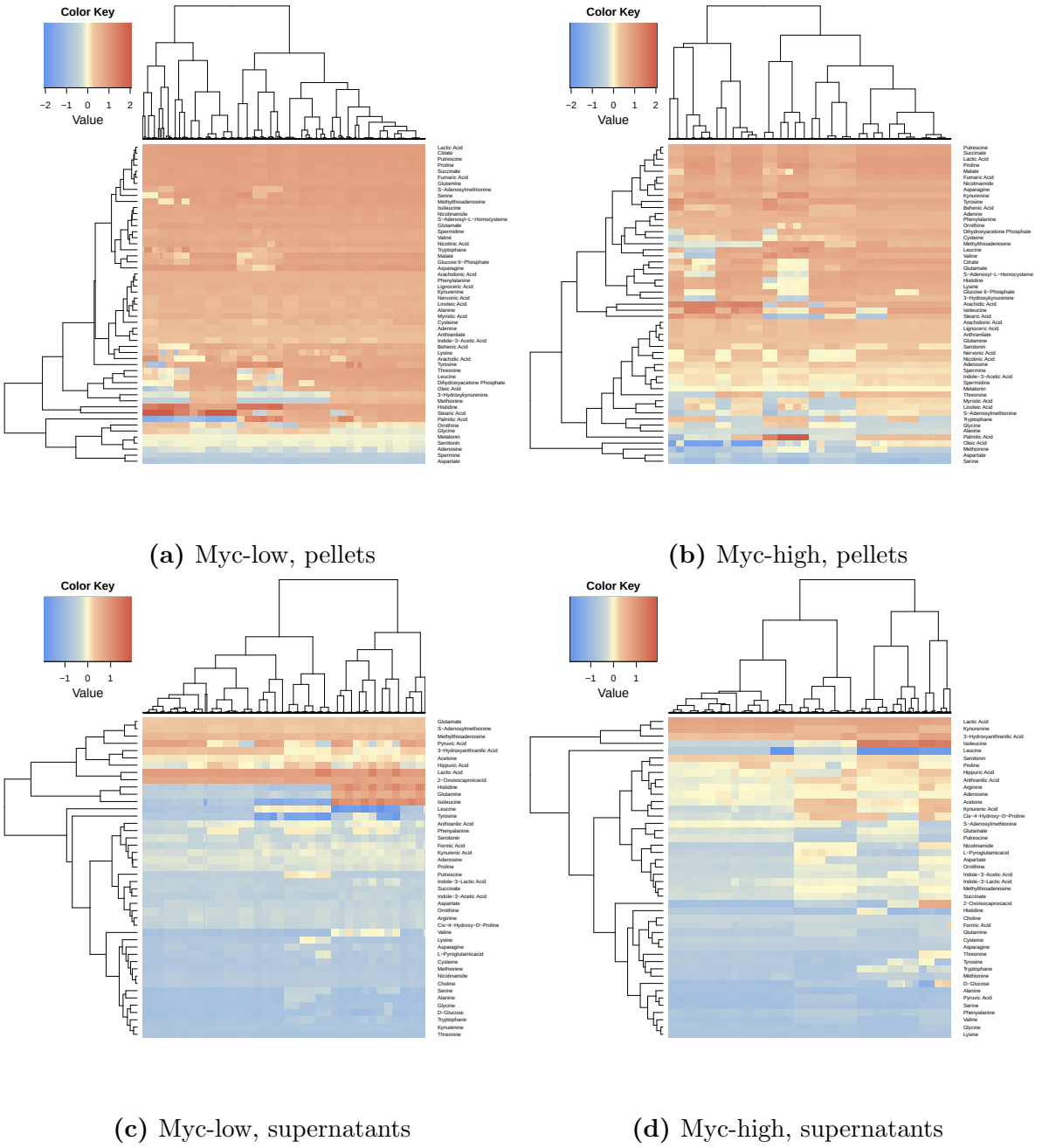
Gene	Ensembl gene name	Effect ("high")	Effect ("low")
AKT	ENSG00000142208	0.83	0.91
MTOR	ENSG00000198793	0.80	0.92
RAPTOR	ENSG00000141564	0.83	-
TERT	ENSG00000164362	0.83	-
VEGF	ENSG00000112715	0.83	-
GLUT1	ENSG00000117394	0.89	-
HK2	ENSG00000159399	0.85	0.86
ENO1	ENSG00000074800	0.86	0.92
PKM2	ENSG00000067225	0.78	0.89
MCT1	ENSG00000155380	0.86	0.91
SHMT1	ENSG00000176974	0.90	-
SHMT2	ENSG00000182199	0.96	-
SLC7A5	ENSG00000103257	0.84	0.88

Table 6.4: Causal effects of *MYC* on selected genes. The causal effects were estimated by the MMPC-algorithm with α =0.5. If there is no causal effect shown, the gene was not selected into the dataset for the causal inference analysis.

in the pellets dataset. The supernatant "*MYC*-high" dataset shows a higher consumption of the metabolites than the supernatant "*MYC*-low" dataset. This finding is consistent with the hypothesis that *MYC* is a transcriptional amplifier which upregulates nearly all genes (Nie et al., 2012; Lin et al., 2012) since *MYC* induces genes involved in the synthesis of lipids, nucleotides and amino acids (Kress et al., 2015). For both, the "*MYC*-high" and the "*MYC*-low" supernatant dataset the causal effect of *MYC* on lactic acid is positive, while the causal effect of *MYC* on glucose is negative. Thus lactate accumulates in the supernatant with increased *MYC* expression while glucose consumption increases with increased *MYC* expression (Figure 6.10). *MYC* induces the glycolytic metabolism also known as Warburg effect (Warburg, 1956; Vander Heiden et al., 2009; Le et al., 2012; Murphy et al., 2013) in the virtual intervention experiment. From the transcriptomic studies in Section 6.2 we further see positive causal effects of *MYC* on the glucose transporter *GLUT1* and the glycolytic enzymes *HK2*, *ENO1* and *PKM2* (Table 6.4), which are known to play a role in cancer metabolism (Osthus et al., 2000; Kim et al., 2004; Israelsen and Vander Heiden, 2015). Further *MYC* has a positive causal effect on *MCT1* (Table 6.4), which is responsible for lactate secretion. Thus the causal effects observed in Section 6.2 are consistent with the observation that *MYC* influences both lactate secretion and glucose consumption.

Indeed and in contrast to the "*MYC*-low" cells, we observe a negative causal effect of *MYC* on serine in "*MYC*-high" cells while the causal effect of *MYC* on glycine is positive (Figure 6.10). Serine is metabolized to glycine by serine hydroxymethyltransferase

(SHMT). The genes *SHMT1* and *SHMT2* are *MYC* targets (Nikiforov et al., 2002), too. Thus, upregulated *MYC* increases the glycine synthesis from serine (Sun et al., 2015; Anderton et al., 2017). The transcriptomic results from Section 6.2 underpin this observation, since *MYC* in the "*MYC*-high" cells has a positive causal effect on *SHMT1* and *SHMT2* (Table 6.4). Furthermore, the serine synthesis pathway is often upregulated in cancer (Dejure and Eilers, 2017).



(a) Myc-low, pellets

(b) Myc-high, pellets

(c) Myc-low, supernatants

(d) Myc-high, supernatants

Figure 6.8: Causal effects of the metabolites on the 5000 genes. The causal effects of the metabolites on the 5000 genes are estimated by aIDA with α=0.1.

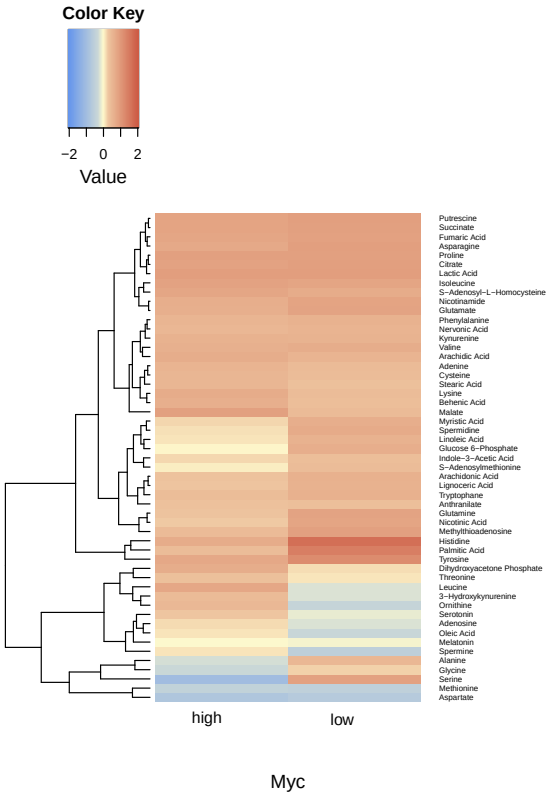
Another mechanism triggered by *MYC* is the uptake of leucine by enhancing the expression of *SLC7A5*, a subunit of the transporter LAT1 (Gao et al., 2009; Hayashi et al., 2012; Qing et al., 2012). The uptake of leucine further increases the expression of mTORC1, which is responsible for proliferation and cancer progression (Nicklin et al., 2009). The results from the causal inference analysis support this observation. *MYC* has a positive causal effect on *SLC7A5* (LAT1, Table 6.4) and with that a positive causal effect on leucine in the pellet dataset, while the causal effect of *MYC* on leucine in the supernatant dataset is negative (Table 6.10). The causal effect of leucine on *MTOR* (causal effect of leucine on *MTOR*: 0.84) and *RAPTOR* (causal effect of leucine on *RAPTOR*: 0.76), both subunits of mTORC1, is positive in the "*MYC*-high" dataset. The gene selection of the "*MYC*-low" dataset did not include the gene *RAPTOR*, but the causal effect of leucine on *MTOR* is positive, too (causal effect of leucine on *MTOR*: 0.73).
Thus, the virtual intervention experiments detect metabolic pathways well-known to be influenced by MYC.

6.4 Discussion

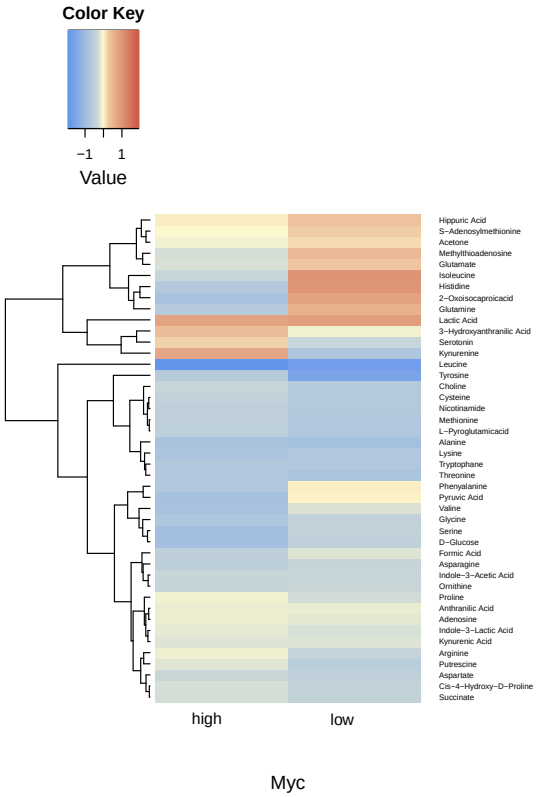
I presented a causal inference analysis in the field of B cell lymphoma research. The analysis focused on the gene *MYC* and its causal relations to the transcriptome and the metabolome. The transcriptome was measured by RNA-seq analysis and the metabolomic dataset has been provided by mass spectrometry analysis. To my knowledge there is no database that includes all causal relations between *MYC* and the metabolome or transcriptome. Hence, the causal DAG is unknown and must be estimated during my analysis. Furthermore, the trancriptomic dataset consists of many more variables than samples. Therefore, as described in the previous chapter, I applied a subsampling strategy to estimate the causal effects. Due to the correlation structure of the data I applied aIDA (Taruttis et al., 2015) to the metabolomics dataset and MMHC-aIDA to the transcriptomic data (Section 4.4). I could show some examples where the results of the causal inference analysis confirm our current knowledge. Especially the list of the most negative causal effects on *MYC* may help to take a look at some promising new candidates for *MYC* repression. However, the results should not be treated as results of real world experiments, since the methods developed and presented in Taruttis et al. (2015) and Section 4.4 cannot replace wet lab experiments. Furthermore, I could not take the whole transcriptome into account for my analysis due to run time restrictions and thus reduced the dataset to the 5000 most promising genes. The metabolomic dataset is not exhaustive as well. Here we were faced with the expensive and time consuming quantification of the metabolites by mass spectrometry analysis and decided for a most useful and feasible

selection of metabolites. NMR-spectrometry would have provided more measurements, but an explicit quantification of each metabolite was not feasible. However, even if these decisions are well justified they violate the assumption that we use all possible causal players in our analysis.

Nevertheless, this was the first causal inference analysis in the context of B cell lymphoma that included a huge part of the transcriptome. I created the causal link to metabolome. I could show that some of the causal connections are already known and explain the carcinogenic effect of *MYC*. The results may also include new causal regulators of *MYC*, but they have to be taken with care and are merely a starting point for new wet lab experiments.

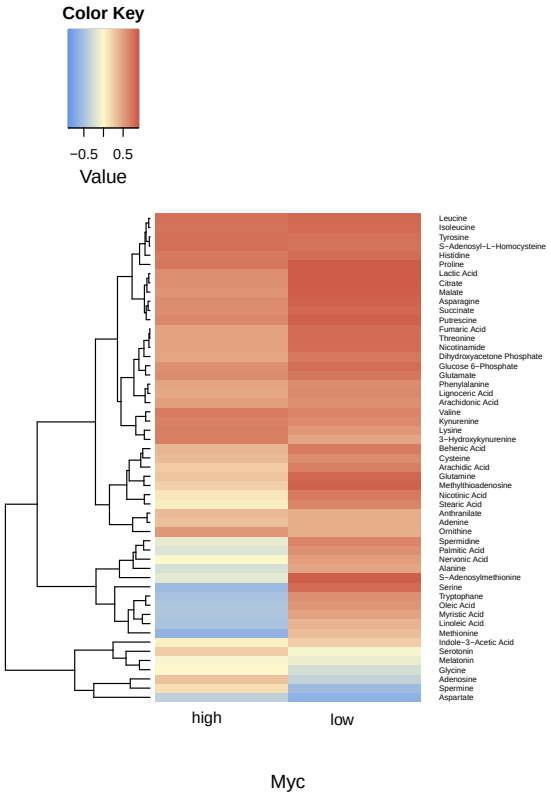


(a) Pellet dataset

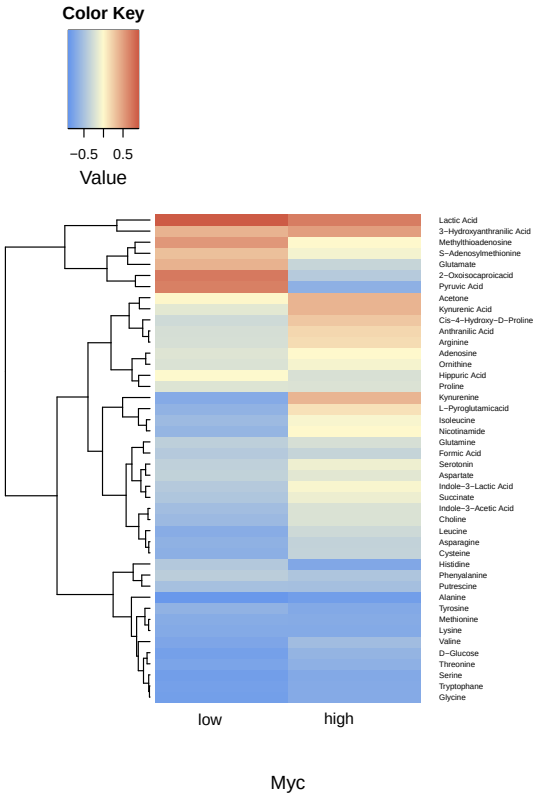


(b) Supernatant dataset

Figure 6.9: Causal effects of the metabolites on *MYC*. The causal effects of the metabolites on MYC estimated by aIDA with α=0.1.



(a) Pellet dataset



(b) Supernatant dataset

Figure 6.10: Causal effects of *MYC* on the metabolites. The causal effects of *MYC* on the metabolites are estimated by MMHC-aIDA with α=0.5.

Part III

Discussion and Outlook

MYC is a hallmark of B cell lymphoma pathogenesis. 30% to 40% of DLBCLs and 70% to 100% of Burkitt lymphomas (Sesques and Johnson, 2017; Johnson et al., 2012; Chisholm et al., 2015; Agarwal et al., 2015; Perry et al., 2013) show increased *MYC* expression. Furthermore, Hanahan and Weinberg (2011) described eight hallmarks of cancer and *MYC* is involved in every single hallmark in the context of lymphomagenesis. In many cases a translocation of *MYC* with an Ig-gene brings the *MYC* oncogene under the control of an Ig-gene promoter. These translocations may occur as a primary or secondary event. While the emergence of translocations involoving *MYC* during B cell developement are well described, *MYC* still conceals some secrets about its function and regulation of gene expression in lymphoma even after more than 30 years of research. Especially the computational analysis of *MYC* positive lymphomas is challenging, since *MYC* is a transcriptional amplifier (Lin et al., 2012; Nie et al., 2012) and common normalization methods cannot deal with a global gene expression change. Thus, the data needs to be adjusted for the transcriptional amplification effect (Lovén et al., 2012). I present a cost-effective and feasible calibration method, which, in contrast to the method of Lovén et al. (2012), also accounts for lysis effects.

However, even if the protocols are conducted with high precision, technical effects during sample preparation and RNA-seqprotocols may differ between human and spike-in cells (Risso et al., 2014). McGee et al. (2019) claim that RNA-seq data is compositional data and that we have to account for that during spike-in normalization methods. They showed that using the ERCC spike-ins together with a compositional approach improves the normalization substantially even if they observe strong variation between ERCC counts in the samples (McGee et al., 2019). Similar effects are expected for the *Drosophila melanogaster* spike-in normalization. The *Drosophila melanogaster* spike-in normalization results in highly correlated data (see Chapter 6.1). It is possible that these correlations to some extend occur due to technical artifacts. Therefore compositional approaches should be considered in future work on *Drosophila melanogaster* and other whole cell spike-in normalization. To sum up, spike-in normalization is the method of choice in RNA-seq normalization when a global gene expression shift between conditions could occur (Evans et al., 2018).

Counting cells before RNA extraction is essential for these methods and counting cells is impossible for some tissue types, for example for solid undissected tissues (Coate and Doyle, 2015). However, whenever possible spike-in normalization offers an additional view of the data, which is important to generate an overall picture of the underlying biological process and spike-ins are mandatory to observe global gene expression changes.

But, even if the spike-in normalization provides improved input data, finding good *MYC* targets by knock down experiments is like searching for a needle in a haystack. There-

fore there is a high need to support the wet lab researchers by computational methods. The causal inference analysis can become a key technology here, but also for many other research questions and in many other fields. Pearl (2009) provides a logical framework for causal inference analysis, while Maathuis et al. (2009) extends the concept of causal inference to causal inference from observational data. Hitherto, a subsampling strategy is recommended for the estimation of causal effects from observational data (Meinshausen and Bühlmann, 2010; Stekhoven et al., 2012). But, neither IDA (Maathuis et al. (2009)) nor CStaR (Stekhoven et al., 2012) make use of the distribution of the causal effects over the multisets and subsampling runs, but estimate a lower bound of the effect size. I developed aIDA which uses the mode of this distribution to estimate the causal effects. And in fact, aIDA outperforms IDA and CStaR on simulated datasets and yeast datasets. However, there are some assumptions which may not be fulfilled and which affect the causal inference analysis: Andersen (2013) points out that the faithfulness assumption can be violated in biological systems, since these systems tend to maintain an equilibrium state. The faithfulness assumption is not testable in general (Zhang and Spirtes, 2008), and thus it remains unknown whether it is violated or not. Furthermore the underlying biological progress may include feedback mechanisms and a DAG cannot represent cyclic relations. Thus unfaithful subgraphs and feedback mechanisms impede the causal structure learning. Other relevant issues have been tackled theoretically. For example Frot et al. (2019) deal with the problem of having hidden variables in the causal graph. In their study, they develop several methods for dealing with that issue and compare their results to some state-of-the-art algorithms on both simulated and small (less than 1000 genes) real world data sets. Further Perković et al. (2017) showed how background knowledge is used to improve the estimation of causal effects from observational data and Hauser and Bühlmann (2015) presented how to use additional interventional data to estimate a DAG. Nandy et al. (2017) show which assumptions have to be made to estimate the effect of joint interventions. The three studies apply their methods to simulated data sets or small (less than 100 variables) real word data sets. Now we have to apply these methods to huge real world data sets. Using prior knowledge either within the graph estimation step or when deriving the causal effects from the multi sets could further improve the estimation of causal effects from observational data.

The application of our calibration method leads to highly correlated datasets. This is a violation of the assumption, that the underlying causal graph needs to be sparse. I show that using the MMHC algorithm instead of the PC algorithm together with my accumulation method outperforms aIDA for highly correlated datasets. However, if the density of the graph becomes too high, both graph learning algorithms break down. Score and search methods show a better performance when the underlying causal network is

dense (Daly et al., 2011). The greedy equivalence search (GES) (Chickering, 2002b) is an important score and search method. Nandy et al. (2018) showed that both, GES and adaptively restricted GES (ARGES) generally outperform the PC algorithm. Furthermore, Scutari et al. (2019) show that another score and search algorithm, the tabu search, outperforms constraint-based methods in terms of accuracy and run time. Future research on causal inference in highly correlated settings should consider these score and search methods.

After deriving these methods to improve causal inference from observational data I present an exemplary study on a B cell lymphoma cell line which includes both, measurements of the transcriptome and metabolome. However, neither the gene expression data nor the metabolomics data are complete observations of the transcriptome and metabolome. The gene expression data needed to be restricted to 5000 genes due to run time and only a part of the metabolome could be quantified. This is a violation of the assumption of causal inference analysis, where all variables must be observed. Furthermore, biologically there is no direct link between metabolome and transcriptome and vice versa. For future research adding proteome data adds the intermediate layer and will improve the analysis.

However, I showed that some examples of well known pathways and biological mechanisms are underpinned by the causal inference analysis. This inspires trust and confidence for future experiments.

In the future, we need to construct real world experiments to show the benefits and the limits of these tools. This includes the preparation of data sets with many samples (> 100 samples) for the prediction of causal effects from this observational data and large scale knock down experiments of theses top estimated causal effects for validation. We have to convince our lab partners, that it is worthwhile expense, since clean data from well suited observational experiments with many samples offer the advantage to derive $p^2 - p$ causal effects from p genes. There is a need for methods that visualize and process this huge amount of causal effects. This thesis can serve as a blueprint for similar approaches to infer causal information from observational data in the field of cancer research. Especially in the context of next generation sequencing the extremely fast decreasing costs of sample preparation (Wetterstrand (2020), Figure 6.11) enables the generation of large datasets suitable for causal inference. However the application of these tools in the wet labs is still in an early stage and this must change.

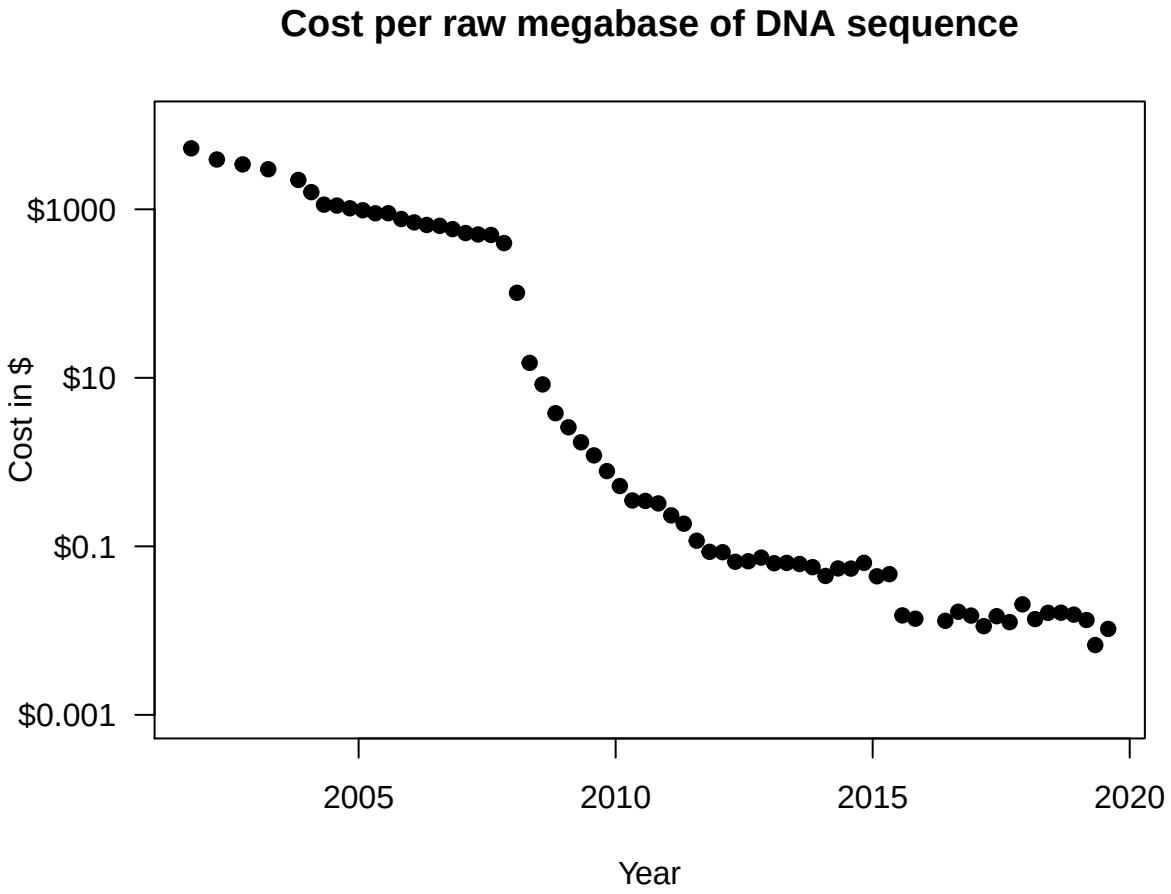


Figure 6.11: Cost per raw megabase of DNA sequence versus time Cost per raw megabase of DNA sequence includes the costs for lab, administration, management, utilities, reagents, consumables sequencing instruments computational activities during sequence production and submission to a public database. The data was provided by Wetterstrand (2020).

List of Figures

List of Tables

Part IV

Appendix

APPENDIX A

Experimental setup of RNA-seq experiment

Maren Feist (Department of Haematology and Medical Oncology of the University Medical Center Göttingen) carried out all wet lab experiments. RNA-seq was done by Dr. Gabriela Salinas-Riester (Head of Core Microarray and Deep-Sequencing Core Facility, University Medical Center Göttingen).

A.1 Cell culture and cell spike-in

Schneider S2 cells from *Drosophila melanogaster* were cultured in Schneider's Drosophila medium (Gibco) supplemented with 10% FCS (Gibco) and penicillin/streptomycin. Cells were grown at room temperature (normal air pressure) and splitted once to twice a week. For spike-in preparation, S2 cells were counted with a haemacytometer and aliquots of $5 \cdot 10^6$ cells were frozen in 1 ml freezing media (45% conditioned S2-media+ 45% FCS +10% DMSO) using a freezing container. Spike-in stocks were stored at -150°C. All experimental samples were spiked with cells from one freezing stock. The human B cell line P493-6, which carries a conditional tetracycline/doxycycline-regulated *MYC* gene, a kind gift from Georg Bornkamm (Munich), were cultured in RPMI medium (Lonza) supplemented with 10% tetracycline-free FCS (Lonza) and penicillin/streptomycin at 37°C and 5% CO2. For suppression of MYC, cells were treated with 1ng/ml doxycycline for 16h. For the Myc high condition, cells were kept in RPMI medium without doxycycline. For the comparison of Myc high and Myc low levels, 10 biological replicates are available for each group. In the dilution experiment, each group consists of three biological replicates. Further the P493-6 cells were treated with α-IgM F(ab)2 fragments (α-IgM),

sCD40L (CD40), rhIGF-1 (IGF), rhIL-10 (IL10) and ODN2006 (CpG) in two different dosages. In total there are 100 samples of P493-6 cells which are treated with that stimuli as described in Table 5.1, where 1 refers to full dosage, 0 refers to no treatment and 0.2 refers to reduced dosage. The concentrations which refer to full or reduced dosage are stimuli depended (Table A.1). For the stimulation experiments, P493-6 cells were treated

Stimulant	Full dosage	reduced dosage
α-IgM	1.3 μg/ml	26 ng/ml
CD40	100 ng/ml	20 ng/ml
IGF	100 ng/ml	20 ng/ml
IL10	25 ng/ml	5 ng/ml
CpG	0.5 μM	0.1 μM

Table A.1: Dosages for the treatment of the P493-6 cells with different stimuli.

with α-IgM F(ab)2 fragments (130ng/ml, Jackson Immunity) , sCD40L (100ng/ml, AutogenBioclear), , CpG (0.5μM, OD2006, Invivogen) for 24h with full dosage. For all experiments, cells were seeded in fresh media at a density of $1 \cdot 10^6/ml$ 24h before harvesting of cells. For harvest of RNA, P493-6 cells were counted with a haemacytometer and a sample with indicated cell number was transferred into a centrifugation tube. For each experiment a fresh aliquot of S2 cells was thawed for 2min in a 37°C water bath. Keeping the spike-in cells in suspension, 20μl (=100.000 cells) were added directly to each transferred P493-6 cell suspension. Together, cells were centrifuged for 5min at 900g and washed once with cold PBS. Dry pellets were stored at -80°C.

A.2 RNA Isolation and ERCC spike-in

Cell pellets were lysed and total RNA extracted using the NucleoSpin RNA Isolation Kit (Machery-Nagel) according to the manufacturer's protocol. Total RNA was extracted from columns using 50μl of RNase free water and spiked with 2μl of a 1:100 dilution of ERCC spike-in Mix 1 (Life Technologies). Quantity and quality of the RNA were assessed using a Nanodrop 1000 and Agilent Bioanalyser 2100, respectively.

A.3 RT-qPCR

For real time PCR, RNA was transcribed to cDNA using SuperScript II Reverse Transcriptase (Invitrogen) and random hexamer primers (IBA BioTAGnology). cDNA samples were analyzed by SYBR Green-based real-time PCR using the 7900HT Fast Real-Time PCR System (Applied Biosystems). All δCt values were normalized to Act42A

expression of the *Drosophila* spike-in or the internal housekeeper gene GAPDH for comparison.

A.4 RNA sequencing

RNA sequencing libraries were prepared from 1μg total RNA containing the ERCC spike-in Mix 1 using the TruSeq RNA Sample Preparation Kit v2 (Illumina). Libraries were sequenced in single end mode for 100 cycles on an Illumina HiSeq 2000 with a mean sequencing depth of 39.9 mio. reads per sample in the dilution experiment and 27.8 mio reads per sample in the MYC high versus MYC low experiment.

APPENDIX B

Experimental setup of metabolomics experiment

B.1 Cell culture and extraction of cell pellets and supernatants

The P493-6 cells have been treated with doxycycline and combinations of α-IgM F(ab)2 fragments (α-IgM), sCD40L (CD40), rhIGF-1 (IGF), rhIL-10 (IL10) and ODN2006 (CpG) in two different dosages as described in section A.1. 5×10^6 cells were centrifuged (300 x g, 5 min, 4°C) after 24 hours and supernatants were transferred into a new tube. The cell pellets were washed in phosphate-buffered saline two times and resuspended in cold 80 % methanol. 10 kD ultra centrifugal filters were activated by adding 3 ml of H2O and centrifuged for 30 min at 4 000 g and the supernatants were loaded onto filters and centrifuged (4000 x g, 30 min, 4°C). The resulting filtrate was transferred into a new tube. Finally, supernatant and pellet were stored at −80°C.

B.2 Mass spectrometry

The metabolites were extracted using the methanol method described by Dettmer et al. (2011). The measurement of amino acid, tryptophan derivates, organic acids and MTA metabolites was performed as previously described by Van Der Goot et al. (2012); Zhu et al. (2011); Stevens et al. (2010) by using an internal isotope labeled standard.

APPENDIX C

Data generation and preprocessing

C.1 Simulation of artificial datasets

The artificial datasets were generated using the R package pcalg (Kalisch et al., 2012). The generation of random DAGs with n nodes starts with the first node x. The number of neighboring nodes is drawn from a Binomial distribution Bin(k,p), where k defines the number of nodes with a higher order than the node x and p is the probability of connecting this node to a node with a higher topological order. After that the nodes connected to x are randomly drawn from all nodes with a higher topological order.

Table C.1 summarizes the parameters for the generation of the several artificial datasets with 1000 nodes.

The edge weights are sampled from an uniform distribution from 0.1 to 1 for each edge.

In a second step we used the R package pcalg (Kalisch et al. (2012)) to simulate a dataset from that graph.

The data of a certain node X_i is calculated by the following equation:

$$X_i = w_1 * pa_i^1 + ... + w_k * pa_i^k + E_i, \tag{C.1}$$

, where $pa_i^1, ..., pa_i^k$ are the parents of X_i, $w_1, ..., w_k$ denote the weights of the incoming edges to X_i and E_i defines the error distribution.

The E_i are sampled from a normal distribution $N(0, 0.001)$.

	Sparse	Dense	Very dense	Extreme dense
p	0.0025	0.005	0.04	0.1

Table C.1: Parameter for the generation of the artificial datasets

C.2 Hughes et al. (2000) dataset

The Hughes et al. (2000) dataset consists of 234 single gene knock out samples (interventional dataset for validation) and 63 wild type gene expression profiles of 5261 genes (observational data). Maathuis et al. (2010)) and Stekhoven et al. (2012) already used this dataset for causal inference from observational data. Thus, the data was preprocessed as described in Maathuis et al. (2010).

C.3 Lenstra et al. (2011) dataset

The Lenstra et al. (2011) dataset is available from ArrayExpress. Both, the interventional and the observational part of the dataset, were measured on A-UMCU-10 - UMC Utrecht S. cerevisiae 16K two channel arrays (version 1.3). The interventional data consists of 165 single gene knock outs (E-TABM-984-processed-data-2481698487.txt). All knockouts that did not map to two hybridization names in E-TABM-984.sdrf.txt were removed. After that all knock out genes, which could not be mapped to ensembl IDs (via SGD IDs from A-UMCU-10.adf.txt and biomart (Smedley et al., 2015)) were removed. Finally, we receive 138 gene knock outs. The observational part of the data consists of 67 gene expression profiles from *S.cerevisae* wild type stains. In both datasets only genes which could be mapped to ensembl gene IDs were considered which led to a set of 4890 genes. No further gene selection was performed. The log fold changes between the two channels were computed using the R package limma (Smyth (2004)).
Both datasets were standardized to obtain N(0,1) centered and scaled gene expression data for each measured gene.

APPENDIX D

Definition of the target set of causal effects

D.1 Hughes and Holstege data

The target set of causal effects for the Hughes et al. (2000) and the Lenstra et al. (2011) dataset of gene i on gene j are calculated as described in Maathuis et al. (2010) using the following formula

$$\beta_{ij} = \frac{|a_{i,j} - mean(a_{-i,j})|}{|a_{i,c(i)} - mean(a_{-i,c(i)})|}, \tag{D.1}$$

where $a_{i,j}$ are the entries of the interventional data matrix with the 234 or 136 knock outs in the rows and the 5361 and 4890 genes in the columns and $mean(a_{-i,j})$ is a short cut for the mean of the j-th column of the interventional data matrix without considering the ith row.

The highest absolute 5% of causal effects define our target set, which is the set of causal effects we want to predict. Nevertheless, since biochemical experiments that characterize direct interactions are missing, this target set is not the set of the top true causal effects, but we expect an enrichment of causal interactions within that set.

D.2 Artificial datasets

Since the true underlying DAG is known in that case, we apply the second part of the IDA method, the estimation of causal effects given the data and the DAG, to our sampled data and DAG. These causal effects represent our ground truth. The highest absolute 5% of true causal effects define our target set, which is the set of causal effects we want

to predict.

APPENDIX E

CStaR parameters

For the small simulated datasets with 100 nodes we chose q $\in \{1\%, 1.2\%, 1.4\%, ..., 10\%\}$ of all possible causal effects. For the large simulated datasets with 1000 nodes and the two gene expression datasets from *S.cerevisae* (Hughes et al., 2000; Lenstra et al., 2011) we used q $\in \{0.01\%, 0.03\%, 0.05\%, ..., 1\%\}$ of all possible causal effects. Stekhoven et al. (2012) showed that the results were insensitive to the choice of the range of qs.

Bibliography

Aanes, H., Winata, C., Moen, L. F., Østrup, O., Mathavan, S., Collas, P., Rognes, T., and Aleström, P. (2014). Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels. *PLoS ONE*, 9(2):e89158+.

Adams, J., Harris, A., Pinkert, C., Corcoran, L., Alexander, W., Cory, S., Palmiter, R. D., and Brinster, R. (1985). The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature*, 318(6046):533–538.

Adams, J. M. and Cory, S. (2007). The Bcl-2-regulated apoptosis switch: mechanism and therapeutic potential. *The Journal of cell biology*, 19(5):488–496.

Agarwal, R., Lade, S., Liew, D., Rogers, T.-M., Byrne, D., Feleppa, F., Juneja, S., and Westerman, D. A. (2015). Role of immunohistochemistry in the era of genetic testing in myc-positive aggressive b-cell lymphomas: a study of 209 cases. *Journal of clinical pathology*, pages jclinpath–2015.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Amati, B., Brooks, M. W., Levy, N., Littlewood, T. D., Evan, G. I., and Land, H. (1993). Oncogenic activity of the c-myc protein requires dimerization with max. *Cell*, 72(2):233–245.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., et al. (2008). Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25. 1. *Nature genetics*, 40(5):616.

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5):672–683.

Anderton, B., Camarda, R., Balakrishnan, S., Balakrishnan, A., Kohnz, R. A., Lim, L., Evason, K. J., Momcilovic, O., Kruttwig, K., Huang, Q., et al. (2017). Myc-driven inhibition of the glutamate-cysteine ligase promotes glutathione depletion in liver cancer. *EMBO reports*, 18(4):569–585.

Aranda, R., Dineen, S. M., Craig, R. L., Guerrieri, R. A., and Robertson, J. M. (2009). Comparison and evaluation of rna quantification methods using viral, prokaryotic, and eukaryotic rna over a 10 4 concentration range. *Analytical biochemistry*, 387(1):122–127.

Aubrey, B. J., Strasser, A., and Kelly, G. L. (2016). Tumor-suppressor functions of the tp53 pathway. *Cold Spring Harbor perspectives in medicine*, 6(5):a026062.

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., Causton, H., Conley, M. P., Elespuru, R., Fero, M., Foy, C., Fuscoe, J., Gao, X., Gerhold, D. L., Gilles, P., Goodsaid, F., Guo, X., Hackett, J., Hockett, R. D., Ikonomi, P., Irizarry, R. A., Kawasaki, E. S., Kaysser-Kranich, T., Kerr, K., Kiser, G., Koch, W. H., Lee, K. Y., Liu, C., Liu, Z. L., Lucas, A., Manohar, C. F., Miyada, G., Modrusan, Z., Parkes, H., Puri, R. K., Reid, L., Ryder, T. B., Salit, M., Samaha, R. R., Scherf, U., Sendera, T. J., Setterquist, R. A., Shi, L., Shippy, R., Soriano, J. V., Wagar, E. A., Warrington, J. A., Williams, M., Wilmer, F., Wilson, M., Wolber, P. K., Wu, X., Zadro, R., and , E. R. N. A. C. C. (2005). The External RNA Controls Consortium: a progress report. *Nat Methods*, 2(10):731–734.

Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M., and Aravind, L. (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of molecular biology*, 360(1):213–227.

Basso, K. and Dalla-Favera, R. (2015). Germinal centres and b cell lymphomagenesis. *Nature reviews Immunology*, 15(3):172–184.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993). A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic acids research*, 21(14):3329.

Baudino, T. A., McKay, C., Pendeville-Samain, H., Nilsson, J. A., Maclean, K. H., White, E. L., Davis, A. C., Ihle, J. N., and Cleveland, J. L. (2002). c-myc is essential

for vasculogenesis and angiogenesis during development and tumor progression. *Genes & development*, 16(19):2530–2543.

Birchmeier, W. and Behrens, J. (1994). Cadherin expression in carcinomas: role in the formation of cell junctions and the prevention of invasiveness. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1198(1):11–26.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160.

Canel, M., Serrels, A., Frame, M. C., and Brunton, V. G. (2013). E-cadherin–integrin crosstalk in cancer invasion and metastasis. *J Cell Sci*, 126(2):393–401.

Cantley, L. C. (2002). The phosphoinositide 3-kinase pathway. *Science*, 296(5573):1655–1657.

Casey, S. C., Tong, L., Li, Y., Do, R., Walz, S., Fitzgerald, K. N., Gouw, A. M., Baylot, V., Gütgemann, I., Eilers, M., et al. (2016). Myc regulates the antitumor immune response through cd47 and pd-l1. *Science*, 352(6282):227–231.

Cesare, A. J. and Reddel, R. R. (2010). Alternative lengthening of telomeres: models, mechanisms and implications. *Nature reviews genetics*, 11(5):319–330.

Chakravorty, D., Jana, T., Mandal, S. D., Seth, A., Bhattacharya, A., and Saha, S. (2017). Mycbase: a database of functional sites and biochemical properties of myc in both normal and cancer cells. *BMC bioinformatics*, 18(1):224.

Cheng, T.-Y. D., Cramb, S. M., Baade, P. D., Youlden, D. R., Nwogu, C., and Reid, M. E. (2016). The international epidemiology of lung cancer: latest trends, disparities, and tumor characteristics. *Journal of Thoracic Oncology*, 11(10):1653–1671.

Chickering, D. M. (2002a). Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498.

Chickering, D. M. (2002b). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Chickering, D. M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5(Oct):1287–1330.

Chisholm, K. M., Bangs, C. D., Bacchi, C. E., Molina-Kirsch, H., Cherry, A., and Natkunam, Y. (2015). Expression profiles of myc protein and myc gene rearrangement in lymphomas. *The American journal of surgical pathology*, 39(3):294–303.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133.

Cleary, M. L., Smith, S. D., and Sklar, J. (1986). Cloning and structural analysis of cdnas for bcl-2 and a hybrid bcl-2/immunoglobulin transcript resulting from the t (14; 18) translocation. *Cell*, 47(1):19–28.

Coate, J. E. and Doyle, J. J. (2015). Variation in transcriptome size: are we getting the message? *Chromosoma*, 124(1):27–43.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782.

Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M. J., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C., and Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in burkitt lymphoma cells. *Proceedings of the National Academy of Sciences*, 79(24):7824–7827.

Dalla-Favera, R., Martinotti, S., Gallo, R. C., Erikson, J., and Croce, C. M. (1983). Translocation and rearrangements of the c-myc oncogene locus in human undifferentiated b-cell lymphomas. *Science*, 219(4587):963–967.

Daly, R., Shen, Q., and Aitken, J. S. (2011). Learning Bayesian networks: approaches and issues. *Knowledge Eng. Review*, 26(2):99–157.

D'Cruz, C. M., Gunther, E. J., Boxer, R. B., Hartman, J. L., Sintasath, L., Moody, S. E., Cox, J. D., Ha, S. I., Belka, G. K., Golant, A., et al. (2001). c-myc induces mammary tumorigenesis by means of a preferred pathway involving spontaneous kras2 mutations. *Nature medicine*, 7(2):235–239.

de Campos, L. M. and Huete, J. F. (2000). Approximating causal orderings for bayesian networks using genetic algorithms and simulated annealing. In *Proceedings of the Eighth IPMU Conference*, volume 1, pages 333–340.

De Silva, N. S. and Klein, U. (2015). Dynamics of b cells in germinal centres. *Nature Reviews Immunology*, 15(3):137–148.

Dejori, M. et al. (2005). *Inference Modeling of Gene Regulatory Networks*. PhD thesis, Technische Universität München.

Dejure, F. R. and Eilers, M. (2017). Myc and tumor metabolism: chicken and egg. *The EMBO journal*, 36(23):3409–3420.

Dettmer, K., Nürnberger, N., Kaspar, H., Gruber, M. A., Almstetter, M. F., and Oefner, P. J. (2011). Metabolite extraction from adherently growing mammalian cells for metabolomics studies: optimization of harvesting and extraction protocols. *Analytical and bioanalytical chemistry*, 399(3):1127–1139.

Devic, S. (2016). Warburg effect-a consequence or the cause of carcinogenesis? *Journal of Cancer*, 7(7):817.

Di Lisio, L., Sánchez-Beato, M., Gómez-López, G., Rodríguez, M. E., Montes-Moreno, S., Mollejo, M., Menárguez, J., Martínez, M., Alvés, F. J., Pisano, D. G., et al. (2012). Microrna signatures in b-cell lymphomas. *Blood cancer journal*, 2(2):e57–e57.

Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *British medical journal*, 2(4682):739.

Doll, R. and Hill, A. B. (1952). Study of the aetiology of carcinoma of the lung. *British medical journal*, 2(4797):1271.

Doll, R. and Hill, A. B. (1954). The mortality of doctors in relation to their smoking habits. *British medical journal*, 1(4877):1451.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. *British medical journal*, 2(5001):1071.

Efeyan, A., Zoncu, R., and Sabatini, D. M. (2012). Amino acids and mtorc1: from lysosomes to disease. *Trends in molecular medicine*, 18(9):524–533.

Evan, G. I., Wyllie, A. H., Gilbert, C. S., Littlewood, T. D., Land, H., Brooks, M., Waters, C. M., Penn, L. Z., and Hancock, D. C. (1992). Induction of apoptosis in fibroblasts by c-myc protein. *Cell*, 69(1):119–128.

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19(5):776–792.

Fidler, I. J. (2003). The pathogenesis of cancer metastasis: the'seed and soil'hypothesis revisited. *Nature Reviews Cancer*, 3(6):453–458.

Fisher, R. (1958a). Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science*, 2:151–166.

Fisher, R. (1958b). The nature of probability. *The Centennial Review of Arts & Science*, 2:261–274.

Fisher, R. A. (1958c). Cancer and smoking. *Nature*, 182(4635):596.

Fisher, R. A. (1958d). Lung cancer and cigarettes? *Nature*, 182(4628):108.

Friedman, N., Nachman, I., and Peér, D. (1999). Learning bayesian network structure from massive datasets: the sparse candidate algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc.

Frot, B., Nandy, P., and Maathuis, M. H. (2019). Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):459–487.

Gabay, M., Li, Y., and Felsher, D. W. (2014). Myc activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harbor perspectives in medicine*, 4(6):a014241.

Gaidano, G., Ballerini, P., Gong, J. Z., Inghirami, G., Neri, A., Newcomb, E. W., Magrath, I. T., Knowles, D. M., and Dalla-Favera, R. (1991). p53 mutations in human lymphoid malignancies: association with burkitt lymphoma and chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 88(12):5413–5417.

Gao, P., Tchernyshyov, I., Chang, T.-C., Lee, Y.-S., Kita, K., Ochi, T., Zeller, K. I., De Marzo, A. M., Van Eyk, J. E., Mendell, J. T., et al. (2009). c-myc suppression of mir-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature*, 458(7239):762.

Gao, S., Xiao, Q., Pan, Q., and Li, Q. (2007). Learning dynamic bayesian networks structure based on bayesian optimization algorithm. pages 424–431.

Giuriato, S., Ryeom, S., Fan, A. C., Bachireddy, P., Lynch, R. C., Rioth, M. J., Van Riggelen, J., Kopelman, A. M., Passegué, E., Tang, F., et al. (2006). Sustained regression of tumors upon myc inactivation requires p53 or thrombospondin-1 to reverse the angiogenic switch. *Proceedings of the National Academy of Sciences*, 103(44):16266–16271.

Goldszmidt, M. and Pearl, J. (1992). Rank-based Systems: A Simple Approach to Belief Revision, Belief Update, and Reasoning about Evidence and Actions. pages 661–672.

Gossen, M. and Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences*, 89(12):5547–5551.

Greenberg, R. A., O'Hagan, R. C., Deng, H., Ziao, Q., Hann, S. R., Adams, R. R., Lichtsteiner, S., Chin, L., Morin, G., and DePinho, R. A. (1999). Telomerase reverse transcriptase gene is a direct target of c-myc but is not functionally equivalent in cellular transformation. *Oncogene*, 18(5):1219–1226.

Greider, C. W. and Blackburn, E. H. (1987). The telomere terminal transferase of tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell*, 51(6):887–898.

Gupta, G. P. and Massagué, J. (2006). Cancer metastasis: building a framework. *Cell*, 127(4):679–695.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Harary, F. and Palmer, E. M. (1973). *Graphical enumeration.* Addison-Wesley.

Harley, C. B., Futcher, A. B., and Greider, C. W. (1990). Telomeres shorten during ageing of human fibroblasts. *Nature*, 345(6274):458.

Hauser, A. and Bühlmann, P. (2015). Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 291–318.

Hayashi, K., Jutabha, P., Endou, H., and Anzai, N. (2012). c-myc is crucial for the expression of lat1 in mia paca-2 human pancreatic cancer cells. *Oncology reports*, 28(3):862–866.

He, J., Hu, Y., Hu, M., and Li, B. (2015). Development of pd-1/pd-l1 pathway in tumor immune microenvironment and treatment for non-small cell lung cancer. *Scientific reports*, 5.

He, T.-C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., Da Costa, L. T., Morin, P. J., Vogelstein, B., and Kinzler, K. W. (1998). Identification of c-myc as a target of the apc pathway. *Science*, 281(5382):1509–1512.

Hecht, S. S. (2002). Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention. *The lancet oncology*, 3(8):461–469.

Heckerman, D. (1995). A bayesian approach to learning causal networks. pages 285–295.

Hermeking, H., Eick, D., et al. (1994). Mediation of c-myc-induced apoptosis by p53. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 2091–2091.

Hoffman, B. and Liebermann, D. (2008). Apoptotic signaling by c-myc. *Oncogene*, 27(50):6462–6472.

Hoffman, B. and Liebermann, D. A. (1998). The proto-oncogene c-myc and apoptosis. *Oncogene*, 17(25).

Hsieh, A. L., Walton, Z. E., Altman, B. J., Stine, Z. E., and Dang, C. V. (2015). Myc and metabolism on the path to cancer. In *Seminars in cell & developmental biology*, volume 43, pages 11–21. Elsevier.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., et al. (2008). A susceptibility

locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, 452(7187):633.

Huntsman, D. G. and Caldas, C. (1998). Assignment1 of the E-cadherin gene (CDH1) to chromosome 16q22.1 by radiation hybrid mapping. *Cytogenet. Cell Genet.*, 83(1-2):82–83.

Israelsen, W. J. and Vander Heiden, M. G. (2015). Pyruvate kinase: function, regulation and role in cancer. In *Seminars in cell & developmental biology*, volume 43, pages 43–51. Elsevier.

Jacobs, J. J., Scheijen, B., Voncken, J.-W., Kieboom, K., Berns, A., and van Lohuizen, M. (1999). Bmi-1 collaborates with c-myc in tumorigenesis by inhibiting c-myc-induced apoptosis via ink4a/arf. *Genes & development*, 13(20):2678–2690.

Jäger, U., Böcskör, S., Le, T., Mitterbauer, G., Bolz, I., Chott, A., Kneba, M., Mannhalter, C., and Nadel, B. (2000). Follicular lymphomas' bcl-2/igh junctions contain templated nucleotide insertions: novel insights into the mechanism of t (14; 18) translocation. *Blood*, 95(11):3520–3529.

Janz, S., Potter, M., and Rabkin, C. S. (2003). Lymphoma-and leukemia-associated chromosomal translocations in healthy individuals. *Genes, Chromosomes and Cancer*, 36(3):211–223.

Johnson, N. A., Savage, K. J., Ludkovski, O., Ben-Neriah, S., Woods, R., Steidl, C., Dyer, M. J., Siebert, R., Kuruvilla, J., Klasa, R., et al. (2009). Lymphomas with concurrent bcl2 and myc translocations: the critical factors associated with survival. *Blood*, 114(11):2273–2279.

Johnson, N. A., Slack, G. W., Savage, K. J., Connors, J. M., Ben-Neriah, S., Rogic, S., Scott, D. W., Tan, K. L., Steidl, C., Sehn, L. H., et al. (2012). Concurrent expression of myc and bcl2 in diffuse large b-cell lymphoma treated with rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone. *Journal of clinical oncology*, 30(28):3452–3459.

Jung, D., Giallourakis, C., Mostoslavsky, R., and Alt, F. W. (2006). Mechanism and control of v (d) j recombination at the immunoglobulin heavy chain locus. *Annu. Rev. Immunol.*, 24:541–570.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Kanehisa, M. et al. (2002). The kegg database. In *Novartis Foundation Symposium*, pages 91–100. Wiley Online Library.

Kiiveri, H. and Speed, T. (1982). Structural analysis of multivariate data: A review. *Sociological methodology*, 13:209–289.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36.

Kim, J.-w., Zeller, K. I., Wang, Y., Jegga, A. G., Aronow, B. J., O'Donnell, K. A., and Dang, C. V. (2004). Evaluation of myc e-box phylogenetic footprints in glycolytic genes by chromatin immunoprecipitation assays. *Molecular and cellular biology*, 24(13):5923–5936.

Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. L., Coviello, G. M., Wright, W. E., Weinrich, S. L., and Shay, J. W. (1994). Specific association of human telomerase activity with immortal cells and cancer. *Science*, pages 2011–2015.

Koh, C. M., Khattar, E., Leow, S. C., Liu, C. Y., Muller, J., Ang, W. X., Li, Y., Franzoso, G., Li, S., Guccione, E., et al. (2015). Telomerase regulates myc-driven oncogenesis independent of its reverse transcriptase activity. *The Journal of clinical investigation*, 125(5):2109–2122.

Kress, T. R., Sabò, A., and Amati, B. (2015). Myc: connecting selective transcriptional control to global rna production. *Nature Reviews Cancer*, 15(10):593.

Kretzner, L., Blackwood, E. M., and Eisenman, R. N. (1992). Myc and max proteins possess distinct transcriptional activities. *Nature*, 359(6394):426.

Küppers, R. (2005). Mechanisms of b-cell lymphoma pathogenesis. *Nature Reviews Cancer*, 5(4):251–262.

Küppers, R. and Dalla-Favera, R. (2001). Mechanisms of chromosomal translocations in b cell lymphomas. *Oncogene*, 20(40):5580.

Küppers, R., Klein, U., Hansmann, M.-L., and Rajewsky, K. (1999a). Cellular origin of human B-cell lymphomas. *New England Journal of Medicine*, 341(20):1520–1529.

Küppers, R., Klein, U., Hansmann, M.-L., and Rajewsky, K. (1999b). Cellular origin of human b-cell lymphomas. *New England Journal of Medicine*, 341(20):1520–1529.

Larrañaga, P., Kuijpers, C. M., Murga, R. H., and Yurramendi, Y. (1996a). Learning bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE transactions on systems, man, and cybernetics-part A: systems and humans*, 26(4):487–493.

Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R. H., and Kuijpers, C. M. H. (1996b). Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE transactions on pattern analysis and machine intelligence*, 18(9):912–926.

Le, A., Lane, A. N., Hamaker, M., Bose, S., Gouw, A., Barbi, J., Tsukamoto, T., Rojas, C. J., Slusher, B. S., Zhang, H., et al. (2012). Glucose-independent glutamine metabolism via tca cycling for proliferation and survival in b cells. *Cell metabolism*, 15(1):110–121.

Leber, M. F. and Efferth, T. (2009). Molecular principles of cancer invasion and metastasis (review). *International journal of oncology*, 34(4):881.

Lefebvre, C., Lim, W. K., Basso, K., Dalla Favera, R., and Califano, A. (2007). A context-specific network of protein-dna and protein-protein interactions reveals new regulatory motifs in human b cells. In *Systems Biology and Computational Proteomics*, pages 42–56. Springer.

Leinonen, R., Sugawara, H., and Shumway, M. (2010). The sequence read archive. *Nucleic acids research*, page gkq1019.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, 396(6712):643–649.

Lenstra, T. L., Benschop, J. J., Kim, T., Schulze, J. M., Brabers, N. A., Margaritis, T., van de Pasch, L. A., van Heesch, S. A., Brok, M. O., Koerkamp, M. J. G., et al. (2011). The specificity and topology of chromatin interaction pathways in yeast. *Molecular cell*, 42(4):536–549.

Li, F., Wang, Y., Zeller, K. I., Potter, J. J., Wonsey, D. R., O'Donnell, K. A., Kim, J.-w., Yustein, J. T., Lee, L. A., and Dang, C. V. (2005). Myc stimulates nuclearly encoded mitochondrial genes and mitochondrial biogenesis. *Molecular and cellular biology*, 25(14):6225–6234.

Li, S. and Li, Q. (2014). Cancer stem cells and tumor metastasis (review). *International journal of oncology*, 44(6):1806–1812.

Li, X., Brock, G. N., Rouchka, E. C., Cooper, N. G., Wu, D., O'Toole, T. E., Gill, R. S., Eteleeb, A. M., O'Brien, L., and Rai, S. N. (2017). A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of rna-seq data. *PloS one*, 12(5):e0176185.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Liberti, M. V. and Locasale, J. W. (2016). The warburg effect: how does it benefit cancer cells? *Trends in biochemical sciences*, 41(3):211–218.

Lin, C. Y., Lovén, J., Rahl, P. B., Paranal, R. M., Burge, C. B., Bradner, J. E., Lee, T. I., and Young, R. A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*, 151(1):56–67.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):1.

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I., and Young, R. A. (2012). Revisiting global gene expression analysis. *Cell*, 151(3):476–482.

Ma, L., Young, J., Prabhala, H., Pan, E., Mestdagh, P., Muth, D., Teruya-Feldstein, J., Reinhardt, F., Onder, T. T., Valastyan, S., et al. (2010). mir-9, a myc/mycn-activated microrna, regulates e-cadherin and cancer metastasis. *Nature cell biology*, 12(3):247–256.

Ma, X. M. M. and Blenis, J. (2009). Molecular mechanisms of mTOR-mediated translational control. *Nature reviews. Molecular cell biology*, 10(5):307–318.

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248.

Maathuis, M. H., Kalisch, M., Bühlmann, P., et al. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232.

Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C., and Boffetta, P. (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48(3):889–902.

Mandozzi, J. and Bühlmann, P. (2016a). Hierarchical testing in the high-dimensional setting with correlated variables. *Journal of the American Statistical Association*, 111(513):331–343.

Mandozzi, J. and Bühlmann, P. (2016b). A sequential rejection testing method for high-dimensional regression with correlated variables. *International Journal of Biostatistics*, 12(1):79–95.

Marinkovic, D., Marinkovic, T., Mahr, B., Hess, J., and Wirth, T. (2004). Reversible lymphomagenesis in conditionally c-myc expressing mice. *International journal of cancer*, 110(3):336–342.

Mårtensson, I.-L., Almqvist, N., Grimsholm, O., and Bernardi, A. I. (2010). The pre-b cell receptor checkpoint. *FEBS letters*, 584(12):2572–2579.

McClintock, B. (1941). The stability of broken ends of chromosomes in zea mays. *Genetics*, 26(2):234–282.

McDonnell, T. J., Deane, N., Platt, F. M., Nunez, G., Jaeger, U., McKearn, J. P., and Korsmeyer, S. J. (1989). bcl-2-immunoglobulin transgenic mice demonstrate extended b cell survival and follicular lymphoproliferation. *Cell*, 57(1):79–88.

McGee, W. A., Pimentel, H., Pachter, L., and Wu, J. Y. (2019). Compositional data analysis is necessary for simulating and analyzing rna-seq data.

Mehlen, P. and Puisieux, A. (2006). Metastasis: a question of life or death. *Nature Reviews Cancer*, 6(6):449–458.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meyer, N. and Penn, L. Z. (2008). Reflecting on 25 years with MYC. *Nat. Rev. Cancer*, 8(12):976–990.

Moore, A. and Wong, W.-K. (2003). Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. In *ICML*, volume 3, pages 552–559.

Morrish, F. and Hockenbery, D. (2014). Myc and mitochondrial biogenesis. *Cold Spring Harbor perspectives in medicine*, 4(5):a014225.

Muller, H. (1938). The remaking of chromosomes. *Collecting net*, 13:181–198.

Murphy, T. A., Dang, C. V., and Young, J. D. (2013). Isotopically nonstationary 13c flux analysis of myc-induced metabolic reprogramming in b-cells. *Metabolic engineering*, 15:206–217.

Nägele, A., Dejori, M., and Stetter, M. (2007). Bayesian substructure learning-approximate learning of very large network structures. In *European Conference on Machine Learning*, pages 238–249. Springer.

Nandy, P., Hauser, A., Maathuis, M. H., et al. (2018). High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183.

Nandy, P., Maathuis, M. H., Richardson, T. S., et al. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674.

Nelson, W. G. and Kastan, M. B. (1994). Dna strand breaks: the dna template alterations that trigger p53-dependent dna damage response pathways. *Molecular and cellular biology*, 14(3):1815–1823.

Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61.

Newcomb, E. W. (1995). P53 gene mutations in lymphoid diseases and their possible relevance to drug resistance. *Leukemia & lymphoma*, 17(3-4):211–221.

Nägele, A., Dejori, M., and Stetter, M. (2007). Bayesian substructure learning - approximate learning of very large network structures. In Kok, J. N., Koronacki, J., de Mántaras, R. L., Matwin, S., Mladenic, D., and Skowron, A., editors, *ECML*, volume 4701 of *Lecture Notes in Computer Science*, pages 238–249. Springer.

Nicklin, P., Bergman, P., Zhang, B., Triantafellow, E., Wang, H., Nyfeler, B., Yang, H., Hild, M., Kung, C., Wilson, C., et al. (2009). Bidirectional transport of amino acids regulates mtor and autophagy. *Cell*, 136(3):521–534.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D. R., Tessarollo, L., Casellas, R., et al. (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1):68–79.

Nikiforov, M. A., Chandriani, S., O'Connell, B., Petrenko, O., Kotenko, I., Beavis, A., Sedivy, J. M., and Cole, M. D. (2002). A functional screen for myc-responsive genes

reveals serine hydroxymethyltransferase, a major source of the one-carbon unit for cell metabolism. *Molecular and cellular biology*, 22(16):5793–5800.

Olive, V., Bennett, M. J., Walker, J. C., Ma, C., Jiang, I., Cordon-Cardo, C., Li, Q.-J., Lowe, S. W., Hannon, G. J., and He, L. (2009). mir-19 is a key oncogenic component of mir-17-92. *Genes & development*, 23(24):2839–2849.

Osthus, R. C., Shim, H., Kim, S., Li, Q., Reddy, R., Mukherjee, M., Xu, Y., Wonsey, D., Lee, L. A., and Dang, C. V. (2000). Deregulation of glucose transporter 1 and glycolytic gene expression by c-myc. *Journal of Biological Chemistry*, 275(29):21797–21800.

Pasqualucci, L., Neumeister, P., Goossens, T., Nanjangud, G., Chaganti, R., Küppers, R., and Dalla-Favera, R. (2001). Hypermutation of multiple proto-oncogenes in b-cell diffuse large-cell lymphomas. *Nature*, 412(6844):341–346.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems. *San Mateo, CA: Kaufmann*, 23:33–34.

Pearl, J. (1993). Graphical Models, Causality, And Intervention.

Pearl, J. (2003). Causality: models, reasoning and inference. *Economet. Theor*, 19:675–685.

Pearl, J. (2009). Causality: Models, reasoning, and inference cambridge university press. *New York, NY*.

Perković, E., Kalisch, M., and Maathuis, M. H. (2017). Interpreting and using cpdags with background knowledge. *arXiv preprint arXiv:1707.02171*.

Perry, A. M., Crockett, D., Dave, B. J., Althof, P., Winkler, L., Smith, L. M., Aoun, P., Chan, W. C., Fu, K., Greiner, T. C., et al. (2013). B-cell lymphoma, unclassifiable, with features intermediate between diffuse large b-cell lymphoma and burkitt lymphoma: study of 39 cases. *British journal of haematology*, 162(1):40–49.

Persson, H. and Leder, P. (1984). Nuclear localization and dna binding properties of a protein expressed by human c-myc oncogene. *Science*, 225:718–721.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models.

Polack, A., Hörtnagel, K., Pajic, A., Christoph, B., Baier, B., Falk, M., Mautner, J., Geltinger, C., Bornkamm, G., and Kempkes, B. (1996). c-myc activation renders proliferation of epstein-barr virus (ebv)-transformed cells independent of ebv nuclear

antigen 2 and latent membrane protein 1. *Proceedings of the National Academy of Sciences*, 93(19):10411–10416.

Preudhomme, C., Dervite, I., Wattel, E., Vanrumbeke, M., Flactif, M., Lai, J. L., Hecquet, B., Coppin, M. C., Nelken, B., and Gosselin, B. (1995). Clinical significance of p53 mutations in newly diagnosed burkitt's lymphoma and acute lymphoblastic leukemia: a report of 48 cases. *Journal of Clinical Oncology*, 13(4):812–820.

Qing, G., Li, B., Vu, A., Skuli, N., Walton, Z. E., Liu, X., Mayes, P. A., Wise, D. R., Thompson, C. B., Maris, J. M., et al. (2012). Atf4 regulates myc-mediated neuroblastoma cell death upon glutamine deprivation. *Cancer cell*, 22(5):631–644.

Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature*, 381(6585):751.

Ramiro, A. R., Jankovic, M., Callen, E., Difilippantonio, S., Chen, H.-T., McBride, K. M., Eisenreich, T. R., Chen, J., Dickins, R. A., Lowe, S. W., et al. (2006). Role of genomic instability and p53 in aid-induced c-myc–igh translocations. *Nature*, 440(7080):105–109.

Reed, J. C., Cuddy, M., Slabiak, T., Croce, C. M., and Nowell, P. C. (1988). Oncogenic potential of bcl-2 demonstrated by gene transfer.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896.

Robbiani, D. F. and Nussenzweig, M. C. (2013). Chromosome translocation, b cell lymphoma, and activation-induced cytidine deaminase. *Annual Review of Pathology: Mechanisms of Disease*, 8:79–103.

Roehle, A., Hoefig, K. P., Repsilber, D., Thorns, C., Ziepert, M., Wesche, K. O., Thiere, M., Loeffler, M., Klapper, W., Pfreundschuh, M., et al. (2008). Microrna signatures characterize diffuse large b-cell lymphomas and follicular lymphomas. *British journal of haematology*, 142(5):732–744.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70:41–55.

Sabò, A., Kress, T. R., Pelizzola, M., de Pretis, S., Gorski, M. M., Tesi, A., Morelli, M. J., Bora, P., Doni, M., Verrecchia, A., et al. (2014). Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature*, 511(7510):488–492.

Sabo, A., Kress, T. R., Pelizzola, M., De Pretis, S., Gorski, M. M., Tesi, A., Morelli, M. J., Bora, P., Doni, M., Verrecchia, A., et al. (2014). Selective transcriptional regulation by myc in cellular growth control and lymphomagenesis. *Nature*, 511(7510):488.

Sander, S., Calado, D. P., Srinivasan, L., Köchert, K., Zhang, B., Rosolowski, M., Rodig, S. J., Holzmann, K., Stilgenbauer, S., Siebert, R., Bullinger, L., and Rajewsky, K. (2012). Synergy between pi3k signaling and myc in burkitt lymphomagenesis. *Cancer cell*, 22(2):167–179.

Schmitt, C. A., Fridman, J. S., Yang, M., Lee, S., Baranov, E., Hoffman, R. M., and Lowe, S. W. (2002). A senescence program controlled by p53 and p16 ink4a contributes to the outcome of cancer therapy. *Cell*, 109(3):335–346.

Schmitt, C. A. and Lowe, S. W. (2001). Bcl-2 mediates chemoresistance in matched pairs of primary eμ-myc lymphomas in vivo. *Blood Cells, Molecules, and Diseases*, 27(1):206–216.

Schmitt, C. A., McCurrach, M. E., de Stanchina, E., Wallace-Brodeur, R. R., and Lowe, S. W. (1999). Ink4a/arf mutations accelerate lymphomagenesis and promote chemoresistance by disabling p53. *Genes & development*, 13(20):2670–2677.

Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Shaffer, A. L., Hodson, D. J., Buras, E., et al. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 490(7418):116–120.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scutari, M., Graafland, C. E., and Gutiérrez, J. M. (2019). Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253.

Sedivy, J. M. (1998). Can ends justify the means?: telomeres and the mechanisms of replicative senescence and immortalization in mammalian cells. *Proceedings of the National Academy of Sciences*, 95(16):9078–9081.

Seifert, M., Scholtysik, R., and Küppers, R. (2013). Origin and pathogenesis of b cell lymphomas. *Lymphoma: Methods and Protocols*, pages 1–25.

Seitz, V., Butzhammer, P., Hirsch, B., Hecht, J., Gütgemann, I., Ehlers, A., Lenze, D., Oker, E., Sommerfeld, A., von der Wall, E., et al. (2011). Deep sequencing of myc dna-binding sites in burkitt lymphoma. *PloS one*, 6(11):e26837.

Sesques, P. and Johnson, N. A. (2017). Approach to the diagnosis and treatment of high-grade b-cell lymphomas with myc and bcl2 and/or bcl6 rearrangements. *Blood*, 129(3):280–288.

Sewastianik, T., Prochorec-Sobieszek, M., Chapuy, B., and Juszczyński, P. (2014). Myc deregulation in lymphoid tumors: molecular mechanisms, clinical consequences and therapeutic implications. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1846(2):457–467.

Shaffer, A., Rosenwald, A., and Staudt, L. M. (2002). Lymphoid malignancies: the dark side of b-cell differentiation. *Nature reviews. Immunology*, 2(12):920.

Shaffer III, A. L., Young, R. M., and Staudt, L. M. (2012). Pathogenesis of human b cell lymphomas. *Annual review of immunology*, 30:565–610.

Smalley, M. J., Signoret, N., Robertson, D., Tilley, A., Hann, A., Ewan, K., Ding, Y., Paterson, H., and Dale, T. C. (2005). Dishevelled (dvl-2) activates canonical wnt signalling in the absence of cytoplasmic puncta. *Journal of Cell Science*, 118(22):5279–5289.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J. W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R. J., Dassi, E., Genova, A. D., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R. C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J. M., Goodstein, D., Gray, K., Guerra-Assunção, J. A., Haggarty, B., Han, D.-J., Han, B. W., Harris, T., Harshbarger, J., Hastings, R. K., Hayes, R. D., Hoede, C., Hu, S., Hu, Z.-L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J.-H., Letellier, T., Li, C.-Y., Lio, P., Liu, C.-J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A. M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A. A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D. M., Steinbach, D., Stone, K., Stupka, E., Teague, J. W., Dayem Ullah, A. Z., Wang, J., Ware, D., Wong-Erasmus, M., Youens-Clark, K., Zadissa, A., Zhang, S.-J., and Kasprzyk, A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72.

Spranger, S., Gajewski, T. F., and Kline, J. (2016). Myc-a thorn in the side of cancer immunity. *Cell research*, 26(6):639–640.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M. H., and Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28(21):2819–2823.

Stevens, A. P., Dettmer, K., Kirovski, G., Samejima, K., Hellerbrand, C., Bosserhoff, A. K., and Oefner, P. J. (2010). Quantification of intermediates of the methionine and polyamine metabolism by liquid chromatography–tandem mass spectrometry in cultured tumor cells and liver biopsies. *Journal of chromatography A*, 1217(19):3282–3288.

Sun, L., Song, L., Wan, Q., Wu, G., Li, X., Wang, Y., Wang, J., Liu, Z., Zhong, X., He, X., et al. (2015). cmyc-mediated activation of serine biosynthesis pathway is critical for cancer progression under nutrient deprivation conditions. *Cell research*, 25(4):429.

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., Maier, K. C., Seizl, M., Tresch, A., and Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome research*, 22(7):1350–1359.

Swann, J. B. and Smyth, M. J. (2007). Immune surveillance of tumors. *Journal of Clinical Investigation*, 117(5):1137.

Swanton, C. and Govindan, R. (2016). Clinical implications of genomic discoveries in lung cancer. *New England Journal of Medicine*, 374(19):1864–1873.

Szenthe, K., Nagy, K., Buzas, K., Niller, H. H., Minarovits, J., et al. (2013). Micrornas as targets and tools in b-cell lymphoma therapy. *Journal of Cancer Therapy*, 4(03):466.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database-Issue):447–452.

Talmadge, J. E. and Fidler, I. J. (2010). Aacr centennial series: the biology of cancer metastasis: historical perspective. *Cancer research*, 70(14):5649–5669.

Taruttis, F., Feist, M., Schwarzfischer, P., Gronwald, W., Kube, D., Spang, R., and Engelmann, J. C. (2017). External calibration with drosophila whole-cell spike-ins delivers absolute mrna fold changes from human rna-seq and qpcr data. *Biotechniques*, 62(2):53–61.

Taruttis, F., Spang, R., and Engelmann, J. C. (2015). A statistical approach to virtual cellular experiments: improved causal discovery using accumulation IDA (aIDA). *Bioinformatics*, page btv461.

Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., and Leder, P. (1982). Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human burkitt lymphoma and murine plasmacytoma cells. *Proceedings of the National Academy of Sciences*, 79(24):7837–7841.

Teater, M. and Melnick, A. (2017). Untangling the web of lymphoma somatic mutations. *Cell*, 171(2):270–272.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452(7187):638.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *AAAI/IAAI*, pages 567–573.

Torre, L. A., Siegel, R. L., and Jemal, A. (2016). Lung cancer statistics. In *Lung cancer and personalized medicine*, pages 1–19. Springer.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111.

Tsamardinos, I., Aliferis, C. F., Statnikov, A., D, P., D, P., Statnikov, E., and Brown, L. E. (2003). Scaling-up bayesian network learning to thousands of variables using local learning techniques. Technical report.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

Tsujimoto, Y., Finger, L. R., Yunis, J., Nowell, P. C., and Croce, C. M. (1984). Cloning of the chromosome breakpoint of neoplastic b cells with the t (14; 18) chromosome translocation. *Science*, 226:1097–1100.

Tsujimoto, Y., Gorham, J., Cossman, J., Jaffe, E., and Croce, C. M. (1985). The t (14; 18) chromosome translocations involved in b-cell neoplasms result from mistakes in vdj joining. *Science*, 229:1390–1394.

Tsujimoto, Y., Louie, E., Bashir, M., and Croce, C. (1988). The reciprocal partners of both the t (14; 18) and the t (11; 14) translocations involved in b-cell neoplasms are rearranged by the same mechanism. *Oncogene*, 2(4):347–351.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D., and Holstege, F. C. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO reports*, 4(4):387–393.

Van Der Goot, A. T., Zhu, W., Vázquez-Manrique, R. P., Seinstra, R. I., Dettmer, K., Michels, H., Farina, F., Krijnen, J., Melki, R., Buijsman, R. C., et al. (2012). Delaying aging and the aging-associated decline in protein homeostasis by inhibition of tryptophan degradation. *Proceedings of the National Academy of Sciences*, 109(37):14912–14917.

Vander Heiden, M. G., Cantley, L. C., and Thompson, C. B. (2009). Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033.

Vaux, D. L., Cory, S., and Adams, J. M. (1988). Bcl-2 gene promotes haemopoietic cell survival and cooperates with c-myc to immortalize pre-B cells. *Nature*, 335(6189):440–442.

Vaziri, H., West, M. D., Allsopp, R. C., Davison, T. S., Wu, Y.-S., Arrowsmith, C. H., Poirier, G. G., and Benchimol, S. (1997). Atm-dependent telomere loss in aging human diploid fibroblasts and dna damage lead to the post-translational activation of p53 protein involving poly (adp-ribose) polymerase. *The EMBO journal*, 16(19):6018–6033.

Verhaak, R., Hoadley, K., Purdom, E., Wang, V., Qi, Y., Wilkerson, M., Miller, C., Ding, L., Golub, T., Mesirov, J., et al. (2010). Cancer genome atlas research network integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *UAI*, pages 255–270. Elsevier.

von Eyss, B. and Eilers, M. (2011). Addicted to myc—but why? *Genes & development*, 25(9):895–897.

Wach, A. (1996). PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in S. cerevisiae. *Yeast*, 12(3):259–265.

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, 131(4):281–285.

Walsh, K. J., Fan, S., Patel, A., Jacobs, C. L., Smith, J. L., Liu, Q., Rizzieri, D. A., and Dave, S. (2009). Pi3k inhibitors inhibit lymphoma growth by downregulation of myc-dependent proliferation.

Wang, J., Xie, L. Y., Allan, S., Beach, D., and Hannon, G. J. (1998). Myc activates telomerase. *Genes & development*, 12(12):1769–1774.

Warburg, O. (1956). On the origin of cancer. *Science*, 123(3191):309–314.

Wei, S.-J., Nguyen, T. H., Yang, I.-H., Mook, D. G., Makena, M. R., Verlekar, D., Hindle, A., Martinez, G. M., Yang, S., Shimada, H., et al. (2020). Myc transcription activation mediated by oct4 as a mechanism of resistance to 13-cis ra-mediated differentiation in neuroblastoma. *Cell death & disease*, 11(5):1–20.

Wetterstrand, K. A. (2019 (accessed April 1, 2020)). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. www.genome.gov/sequencingcostsdata.

Wild, C., Weiderpass, E., and Stewart, B. (2020). World cancer report: cancer research for cancer prevention. *Lyon: International Agency for Research on Cancer*.

Willis, T. G. and Dyer, M. J. (2000). The role of immunoglobulin translocations in the pathogenesis of b-cell malignancies. *Blood*, 96(3):808–822.

Wong, M. L. and Leung, K. S. (2004). An efficient data mining method for learning bayesian networks using an evolutionary algorithm-based hybrid approach. *IEEE Transactions on Evolutionary Computation*, 8(4):378–404.

Wu, K.-J., Grandori, C., Amacker, M., Simon-Vermot, N., Polack, A., Lingner, J., and Dalla-Favera, R. (1999). Direct activation of tert transcription by c-myc. *Nature genetics*, 21(2).

Xiao, C., Srinivasan, L., Calado, D. P., Patterson, H. C., Zhang, B., Wang, J., Henderson, J. M., Kutok, J. L., and Rajewsky, K. (2008). Lymphoproliferative disease and autoimmunity in mice with increased mir-17-92 expression in lymphocytes. *Nature immunology*, 9(4):405–414.

Yustein, J. T., Liu, Y.-C., Gao, P., Jie, C., Le, A., Vuica-Ross, M., Chng, W. J., Eberhart, C. G., Bergsagel, P. L., and Dang, C. V. (2010). Induction of ectopic myc target gene jag2 augments hypoxic growth and tumorigenesis in a human b-cell model. *Proceedings of the National Academy of Sciences*, page 200901230.

Zeller, K. I., Zhao, X., Lee, C. W., Chiu, K. P., Yao, F., Yustein, J. T., Ooi, H. S., Orlov, Y. L., Shahab, A., Yong, H. C., et al. (2006). Global mapping of c-myc binding sites and target gene networks in human b cells. *Proceedings of the National Academy of Sciences*, 103(47):17834–17839.

Zhang, J. and Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271.

Zhu, W., Stevens, A. P., Dettmer, K., Gottfried, E., Hoves, S., Kreutz, M., Holler, E., Canelas, A. B., Kema, I., and Oefner, P. J. (2011). Quantitative profiling of tryptophan metabolites in serum, urine, and cell culture supernatants by liquid chromatography–tandem mass spectrometry. *Analytical and bioanalytical chemistry*, 401(10):3249–3261.