Integrative network modeling of large multidimensional cancer datasets

Teresia Kling

Department of Molecular and Clinical Medicine
Institute of Medicine
Sahlgrenska Academy at University of Gothenburg

UNIVERSITY OF GOTHENBURG

2015

Cover illustration: By Teresia Kling. RNA and DNA details from Wikimedia Commons. Network from cancerlandscapes.org.

Integrative network modeling of large multidimensional cancer datasets

teresia.kling@gu.se

ISBN: 978-91-628-9557-0 (print)
ISBN: 978-91-628-9558-7 (pdf)
http://hdl.handle.net/2077/39547

Printed by Ineko AB, Gothenburg, Sweden 2015

To my family

ABSTRACT

Our ability to conduct detailed molecular investigations on tissue samples have, during the past decade, enabled the formation of databases containing measurements from thousands of cancer tumors. To harness the potential of the amassing data sets, we introduce new modeling techniques and generalise existing methods for large-scale integration of cancer data. These methods aim to construct network models that link genetic, epigenetic, transcriptional and phenotypic events, by combining genome-wide measurements of multiple kinds.

In paper I we constructed a modeling framework, EPoC, for creating causal networks between gene copy number levels and mRNA expression, and applied it to data from the brain tumor glioblastoma. Some of the predicted regulators were tested in four glioblastoma-derived cell lines and confirmed that the network model could be used to find unknown regulators of cell growth in glioblastoma.

In paper II we used data integrative network modeling to identify novel genomic, epigenetic and transcriptional regulators of glioblastoma subtypes. In addition to confirming known regulators of gliomagenesis, the model also predicted that Annexin A2 (ANXA2) promoter methylation and mRNA expression were linked to the signature target genes of the clinically aggressive mesenchymal molecular subtype. Our findings were validated by knockdown of ANXA2 in glioblastoma-derived cell cultures.

Paper III presents an extension of sparse inverse covariance selection (SICS), which is adapted and optimized for modeling of genetic, epigenetic, and transcriptional data across multiple cancer types. To evaluate the potential of the method, we applied it to data from eight cancers available in The Cancer Genome Atlas and published the model online at cancerlandscapes.org for anyone to explore. The derived multi-cancer model detected known interactions and contained interesting predictions, including functionally coupled network structures shared between cancers.

In summary, we use network modeling of cancer to identify possible drug targets, drivers of molecular subclasses, and reveal similarities and differences between cancer types. The developed tools for network construction can assist in further investigation of the cancer genome, potentially including other data sources and additional cancer diagnoses.

Keywords: network modeling, data integration, glioblastoma, pan-cancer analysis, The Cancer Genome Atlas

SAMMANFATTNING PÅ SVENSKA

Under det senaste decenniet har stora nationella och internationella projekt genomförts, som samlat in mätningar från tusentals cancertumörer. Syftet är att kartlägga genetiska och molekylära förändringar i cancerceller jämfört med frisk vävnad. Genom dessa mätningar försöker man bl.a. hitta mutationer (förändringar i DNA-sekvensen), kopieantalsförändringar (hela eller delar av kromosomer som försvunnit eller blivit kopierade till fler av misstag) och så kallade epigenetiska förändringar som påverkar hur DNA avläses och uttrycks. Man mäter också nivåer av transkriberad mRNA, dvs den enkelsträngade molekyl som är mellansteg i översättningen från DNA till protein. Dessutom håller man reda på kliniska fakta om patientern, som ålder, kön och hur länge de överlevt med sin tumör.

För att kunna utnyttja potentialen hos denna mycket stora datamängd behövs avancerade statistiska modeller som klarar av att hantera och koppla samman data av olika typer och från olika källor. I denna avhandling generaliserar vi existerande metoder för storskalig databearbetning och konstruerar nätverksmodeller som kopplar ihop olika typer av molekylär cancerdata. Nätverksmodeller består av noder som symboliserar datavariabler. Noderna är sammankopplade av länkar som representerar att noderna kan associeras till varandra, baserat på mätdata. Syftet är att skapa en visuellt överblickbar modell över kopplingar mellan ett stort antal variabler, och för att påskynda identifiering av viktiga samband.

De två första artiklarna inriktar sig på två olika tillämpningar av nätverksmodeller på hjärntumören glioblastom. Artikel I fokuserar på sambandet mellan kopieantalsförändringar i DNA och nivåer av mRNA. Artikel II involverar också fler datatyper och koncentrerar sig på deras inverkan på en specifik undergrupp till glioblastom. Artikel III introducerar modeller som kan användas till att hantera data från flera cancertyper samtidigt, och tillämpar metoden på data från åtta cancertyper som finns i den publika databasen The Cancer Genome Atlas.

Sammanfattningsvis visar avhandlingen att statistiska nätverksmodeller kan användas som verktyg för att finna möjliga måltavlor för nya mediciner, identifiera potentiella cancerdrivande mekanismer och visa på likheter och skillnader mellan cancertyper. De utvecklade metoderna för nätverkskonstruktion kan framöver användas för ytterligare forskning kring cancergenomik, förhoppningsvis genom att också involvera fler datatyper och cancerdiagnoser.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Jörnsten, R., Abenius, T., **Kling, T.**, Schmidt, L., Johansson, E., Nordling, T. E. M., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., Nelander, S. *Network modeling of the transcriptional effects of copy number aberrations in glioblastoma.* Molecular systems biology, 2011. 7: 486.

II. **Kling, T.***, Ferrarese, R.*, Ó hAilín, D., Heiland, H. H., Dai, F., Vasilikos, I., Weyerbrock, A., Jörnsten, R., Carro**, M. S., Nelander, S**. *Integrative modeling reveals ANXA2 as a determinant of mesenchymal transformation in glioblastoma.* 2015
Submitted
*Joint first authors
**Joint last authors

III. **Kling, T.***, Johansson, P.*, Sánchez, J., Marinescu, V. D., Jörnsten, R., Nelander, S. *Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content.* Nucleic Acids Research, 2015.
*Joint first authors

PAPERS NOT INCLUDED IN THIS THESIS

I. Abenius, T., Jörnsten, R., **Kling, T.**, Schmidt, L., Sánchez, J., Nelander, S., 2012. *System-Scale Network Modeling of Cancer Using EPoC.* Goryanin, I., Goryachev, A., Eds. Advances in Systems Biology. Advances in Experimental Medicine and Biology 736, Springer, 2012.

II. Persson, M., Andrn, Y., Moskaluk, C.A., Frierson, H.F. Jr, Cooke, S.L., Futreal, P.A., **Kling, T.**, Nelander, S., Nordkvist, A., Persson, F., Stenman, G. *Clinically significant copy number alterations and complex rearrangements of MYB and NFIB in head and neck adenoid cystic carcinoma.* Genes, Chromosomes and Cancer, 2012.

III. Gerlee, P., Schmidt, L., Monsefi, N., **Kling, T.**, Jörnsten, R., Nelander, S. *Searching for synergies: matrix algebraic approaches for efficient pair screening.* PLoS One, 2013 Jul 25;8(7):e68598.

IV. **The Cancer Genome Atlas Research Network**, Weinstein, J.N., Collisson, E.A., Mills, G.B., Mills Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013 Oct;45(10):1113-20.

V. Schmidt, L., **Kling, T.**, Monsefi, N., Olsson, M., Hansson, C., Baskaran, S., Lundgren, B., Martens, U., Häggblad, M., Westermark, B., Forsberg Nilsson, K., Uhrbom, L., Karlsson-Lindahl, L., Gerlee, P., Nelander, S.*Comparative drug pair screening across multiple glioblastoma cell lines reveals novel drug-drug interactions.*
Neuro Oncol, 2013 Nov;15(11):1469-78.

Contents

Abbreviations

A	adenine
ADMM	Alternating Directions Method of Multipliers
ANXA2	Annexin A2
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
BTSC	Brain tumor stem cell
C	cytosine
C/EBP-β	CCAAT/Enhancer binding protein (C/EBP), beta
C3SE	Chalmers Centre for Computational Science and Engineering
CDKN2A	Cyclin-dependent kinase inhibitor 2A
CGP	Cancer Genome Project
CNA	Copy Number Aberration
CNS	Central Nervous System
DNA	Deoxyribonucleic Acid
EGFR	Epidermal Growth Factor Receptor
ESR1	Estrogen receptor 1
G	guanine
G-CIMP	glioma-CpG island methylator phenotype
GBM	Glioblastoma
GGM	Gaussian Graphical Models
Glasso	Graphical lasso
GSEA	Gene Set Enrichment Analysis
HGP	Human Genome Project
ICGC	International Cancer Genome Consortium
IDH1	Isocitrate dehydrogenase 1
lasso	least absolute shrinkage and selection operator
LGG	Lower Grade Glioma
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
miRNA	micro RNA
mRNA	messenger RNA

NACC1	Nucleus accumbens-associated protein 1
NCBI	The National Center for Biotechnology Information
NDN	Necdin
NF1	Neurofibromin 1
OLIG2	Oligodendrocyte transcription factor 2
PDGFRA	Platelet-derived growth factor receptor alpha
RHPN2	Rhophilin, Rho GTPase binding protein 2
RNA	Ribonucleic Acid
RNAseq	RNA sequencing
shRNA	short hairpin RNA
SICS	Sparse inverse covariance selection
STAT3	Signal transducer and activator of transcription 3
T	thymine
TAZ	Tafazzin
TCGA	The Cancer Genome Atlas
TP53	Tumor protein P53
U-CAN	Uppsala-Umeå Comprehensive Cancer Consortium
WGCNA	Weighted Correlation Network Analysis
WHO	World Health Organization

1 Introduction

Cancer is the umbrella term for more than 200 diseases[1] having in common that cells start to grow and divide uncontrollably, and with the potential to invade neighboring tissue. The term itself, *cancer*, originates in the Latin word for crab, from the veins surrounding a breast tumor visually resembling the legs of a crab. The modern science of *cancer genetics* began with Boveri in the beginning of the 20th century[2], who hypothesized that chromosomal defects in a cell underlay the process of tumor formation. During the coming decades, experiments on animals to investigate the formation of tumors suggested that multiple alterations were needed for tumors to be able to form. For instance, the sequential exposure to two different carcinogens highly increased tumor incidence rates compared to exposure to only one[3,4,5,6]. Also, the application of tar followed by cutting of the skin of mice[7] and rabbit[8] showed that tar and wound in combination increased the number of tumors. Ashley (1969[9]) and Knudson (1971[10]) compared the age patterns of incidence of inherited and non-inherited forms of colon cancer and retinoblastoma respectively, and concluded that the later onset of cancer for the non-inherited forms was due to the fact that the patients with inherited mutations already had acquired one necessary event for cancer to initiate. An early example of mathematical modeling in the field is estimations based on incidence curves of the number of independent events required for cancer to initiate, first being introduced by Fisher and Hollomon 1951[11,6].

It was clear that mutations were involved in the formation of tumors, but it was not until 1982 that the first cancer-causing mutation was localized in the RAS gene in bladder carcinoma by the Weinberg, Wigler and Barvacid research groups[12,13,14]. The rest of the 1980s established the concept of oncogenes and tumor suppressors, and also made clear that several different types of genetic rearrangements can be the source of activating an oncogene or turn off a tumor suppressor[15]. During the 1990s, technologies began to emerge for doing measurements and analyses on larger parts of the genome simultaneously, resulting in the discovery of for example activating mutations in the oncogene BRAF in a wide range of cancers[16], and the

oncogene EGFR in lung cancer[17,18,19]. Also, in 1990, The Human Genome Project (HGP) was started, partly to create a reference genome sequence for easier findings of cancer mutations, and was finished in the early 2000s.

One study in 2004[20] estimated the number of found tumor-driving genes to 291, using the criteria that at least two independent studies should have reported the gene to be genetically altered and cancer causing. A more recent estimate in 2013, based on 3284 sequenced tumors, reported 138 genes as being driver genes of tumorigenesis[21]. These genes were assigned to 12 different pathways and three core cellular processes, and the authors speculated that it is enough for a cell to accumulate 2-8 of these alterations for cancer to develop. Furthermore, Hanahan and Weinberg describe eight different traits that have to be acquired for cancer to develop, titled *Hallmarks of Cancer*. These traits are assumed to be common for all cancers but are not achieved by the same aberrations for all patients and cancer types[22,23]. Nonetheless, it remains a challenge to understand how mutations in several pathways combine to modulate the phenotype of cancer cells, resulting in the acquired phenotypes that are essential for cancer.

In spite of these advances in understanding the underlying causes and the development of tumors, cancer remains a significant health burden affecting all parts of the world. In 2012 there were 14.1 million new cancer cases reported, and 8.2 million people died because of a cancer disease, which correspond to almost 16 deaths per minute[24]. The lifetime risk, globally, of being diagnosed with cancer is around 43% for men and 38% for women[25], and the lifetime risk of dying from a cancer disease is 23% and 19% for men and women respectively. In Sweden, cancer is the second most common cause of death after cardiovascular diseases[26]. Depending on the type of cancer, the 10-year survival differs between as low as 1% for pancreatic cancer to 98% for testicular cancer[27]. The development of treatment options has improved the survival rates by as much as around 40% over the last 40 years for malignant melanoma, non-Hodgkin lymphoma, leukemia, bowel cancer and female breast cancer[27]. However, for cancers of the pancreas, esophagus, lung and adult brain, very little improvement in survival can be seen during the same period of time[27]. Risk factors also vary between cancer types, but includes inherited genetic predisposition, old age, environment such as sun or radon exposure, and lifestyle such as level of physical activity, smoking, overweight, diet and alcohol habits[24].

In order to address this huge health problem, one important component will be to leverage our molecular insight of cancer genomics into new thera-

pies. Analysis and modeling of cancer genomic datasets from many samples will provide important tools towards this goal. The subject of this thesis is to adapt statistical network modeling tools, including data preparation and normalization, for the context of large scale cancer genomic data of multiple types, and apply the developed tools on tumor data from The Cancer Genome Atlas. The created cancer network models can then be used to identify prognostic biomarkers, possible drug targets, drivers of molecular subclasses, and reveal similarities and differences between cancer types.

2 Cancer genomics

2.1 Comprehensive molecular profiling of cancers

Cancer genomics is a broad field oriented towards mapping and understanding changes in the structure or activity of genes in cancer. One important trend in the last decade has been the transition from application of a single method to characterize a set of samples (such as transcript profiling), to broader application of several methods. Applicable in both basic research and clinical settings, the resulting data from such *comprehensive profiling* gives a high-dimensional view of cancer, revealing the joint presence of acquired mutations, localized chromosomal copy number aberrations, promoter hypermethylations, transcriptional alterations affecting microRNA and mRNA levels etc. Details of some of the molecules and genetic abnormalities that are commonly being investigated are discussed next, and the properties of the technologies to detect such changes and their implication on data analysis are discussed separately in Chapter 5.

i) Observable ***genetic alterations*** in DNA can be as small as one nucleotide (A, G, C or T) or up to a whole chromosome, Figure 1A-C.

- **Somatic point mutations** that occur in a cell are passed on in the next cell division and can sometimes contribute to the process of tumor formation. These mutations are acquired anytime, as opposed to germline mutations which are inherited or appear early in development and exist in all cells in the body. A point mutation is a change, a deletion, or an addition of one base in the DNA sequence, on one or both copies of the gene. A tumor suppressor can be silenced by a mutation causing the resulting protein to be non-functional. Alternatively, a mutation can alter the protein structure or affect the regulation of the gene by being located in the promoter region. Mutations in oncogenes normally occur recurrently in the same amino acid positions, whereas mutations of tumor suppressors occur anywhere in the

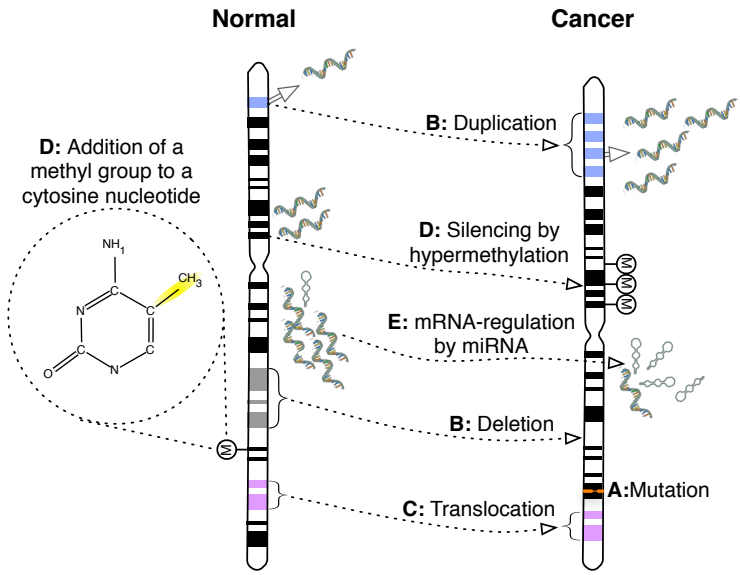

Figure 1: Some potential differences between normal and tumor cells A-C: Genetic alterations. **D:** Epigenetic modifications. M denotes methylation. **E:** Silencing of mRNA levels by miRNA.

sequence of the gene[21]. A mutation having no effect on the protein or its regulation is called *silent*.

- The event when either parts of, or a whole, chromosome exist in other numbers rather than the normal two copies is referred to as a **Copy Number Aberration** (CNA). The loss of one or two copies is referred to as a *deletion* and the gain of extra copies is referred to as *duplication* or *amplification*. Typically, tumor suppressors are deleted and oncogenes are amplified in cancer cells. There is often a positive correlation between the number of copies of a chromosomal region and the amount of mRNA for the genes located there.

ii) ***Epigenetic modifications*** affect gene expression or the phenotype of cells, without altering the DNA sequence.

- **DNA methylation** is the process of the addition of a methyl group to the adenine (A) or cytosine (C) nucleotides of the DNA, Figure 1D. Decreased levels of methylation are referred to as *hypomethylation* and increased levels are referred to as *hypermethylation*. In cancer, hypomethylation of DNA regions with repeated elements can lead to chromosomal instability. Also, methylation levels of the promoter region of a gene has been shown to sometimes correlate with the amounts of transcribed mRNA[28]. Thus, alterations in the methylome is another way for the tumor to regulate the cellular activity.

- Modifications to the **histones**, around which the DNA double strands are wrapped, affect the DNA replication and the transcription levels of closely located genes.

iii) ***Expression profiling*** measure levels of RNA or protein in the cells.

- **mRNA** (messenger RNA) are the RNA molecules that carry the template for protein construction in the cell. The DNA encoding a gene is *transcribed* into mRNA inside the nucleus, the mRNA is then transported out to the cytoplasm and is used as a template by the ribosome during protein formation, a process called *translation*. mRNA molecules are more easily measured than proteins, and mRNA is thus used as a proxy for the protein levels in a cell, although this notion is being debated[29].

- **miRNAs** (microRNA) are small non-coding RNA molecules of around 23 nucleotides. They regulate gene expression by destabilisation of the mRNA molecule or by decreasing the efficiency of the translation process, Figure 1E[30]. miRNAs bind to mRNA molecules by complementary sequences. This sequence matching does not have to be perfect, meaning that the same miRNA can have multiple mRNA targets and can be involved in several processes. Also, a single mRNA can be regulated by multiple miRNAs.

- **Proteins** are the end product of the genes encoded by the DNA. They are large molecules built from combinations of 20 different

amino acids, put together in chains. These chains fold into three-dimensional structures, can carry out very diverse functions and are involved in a majority of all cellular processes.

2.2 Cancer genome projects

One of the large scale cancer biobanking initiatives is The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov[31], see Section 2.2.1). The goal is to create a map of human cancer, by doing large-scale measurements on 500 or more patients from each of 25-30 different human cancers. Currently available data sets in TCGA include mRNA and miRNA expression, copy number alterations (CNAs), DNA methylation patterns, somatic point mutations and protein expression of selected genes. In addition, clinical information, like gender, age, treatment and survival time for patients is collected. Other similar projects include the Cancer Genome Project (CGP, http://www.sanger.ac.uk/genetics/CGP/), and the International Cancer Genome Consortium (ICGC, http://www.icgc.org/). CGP collects sequencing data and aims to present mutations together with other cancer-related information in a public database. ICGC aims to collect and put together data from all different large cancer genome projects around the world.

Apart from these projects covering many cancer types there are multiple examples of initiatives gathering larger number of samples for one specific diagnosis, e.g. the METABRIC project which has collected and analyzed around 2000 breast tumor samples from five hospitals in the UK and Canada[32]. Additionally, programs around the world have been initiated to integrate genomics with national healthcare, for example U-CAN (http://www.u-can.uu.se) which collects and profiles tumor and blood samples before, during and after treatment from patients with a wide range of cancer diagnoses in Sweden. The aim is to develop better diagnostics and characterization of cancer tumors, and to evaluate the performance of new and established treatment options.

2.2.1 TCGA

The Cancer Genome Atlas has been one of the catalyzers moving the cancer genomics field forward by making a huge amount of data publicly available for the researcher community. This has enabled the application of estab-

lished analysis methods on the collected data for individual cancers. By using clustering, new tumor subclasses based on molecular profiles have been identified for example for breast[33], ovarian[34], uterine[35] and brain[36] cancer, having implications on prognosis and treatment options. In contrast, an attempt to combine TCGA data with pathway information concluded that most of the tumors of the two different diagnoses colon and rectal cancer have similar genetic alterations[37]. The same study also, for example, identified the potential drug target ERBB2 as being frequently amplified in these cancers. By correlating mRNA and copy number measurements of ovarian tumors, the NACC1 gene has been found as a biomarker of early recurrence[38]. In kidney cancer, remodeled cellular metabolism has been proposed to be a characteristic of aggressive tumors, by integrating multiple data types with patient survival[39], and other studies have identified the expression profile of a small set of miRNAs to be associated to prognosis[40,41]. By investigating mutational patterns of squamous cell lung tumors, several new potential drug targets have been identified[42]. The results of systematic molecular profiling can further be illustrated by research that has for example found a number of copy number aberrations that predict response to therapy in metastatic colorectal cancer[43]. Despite these and more results, there is still room for development of new integrative methods that successfully model and enable interpretation of the full set of measurements and data collected for each cancer.

In 2012 the Cancer Genome Atlas Pan-Cancer analysis project was initiated[44], presenting, in a structured manner, the first 12 tumor types that had been profiled by TCGA. The Pan-Cancer project engage researchers, including ourselves, around the world to develop methods for the simultaneous analysis and interpretation of multiple cancers. The hopes are that the project will cast new light on similarities and differences between cancers of many types and tissues of origin. The joint analysis of multiple cancers has already resulted in the identification of 127 significantly mutated genes across the set of 12 pan-cancer tumor types[45]. A similar attempt identified 291 cancer driving genes across the 12 cancer types, by combining five different analysis methods[46]. In another study, 10% of the investigated tumors were shown, when studied on the molecular level[47], to belong to a different type of cancer than the histological classification indicated. Multiple Pan-Cancer studies have been performed focusing on one data type. For example, the investigation of copy number data across 11 of the Pan-Cancer diagnoses revealed that the same genomic regions often are being affected by copy number aberrations, across multiple cancer types[48]. De-

spite these advances, the full potential of the Pan-Cancer data set remains to be investigated, most likely by the invention of new advanced analysis methods adapted to the scale and high dimensionality of the data. This involve creating new infrastructure for data transfer and storage, developing normalization protocols, and producing analysis and statistical modeling techniques that reveal the full potential of the huge amount of data.

3 Cancers of the brain and the central nervous system

In paper I and II of this thesis, the central focus is data analytical problems associated with the particular type of tumor called *glioblastoma*, which belongs to a group of cancers localized in the brain or central nervous system (CNS). Malignant primary brain and CNS tumors are rare (incidence rates in USA 8.93 per 100,000 population) compared to the most common cancer types of the prostate, breast and lung (incidence rates 215.96, 173.65 and 95.40 per 100,000 population respectively). Nonetheless, these tumors are the second leading cause to die from cancer in men aged 20 to 39 years and the fifth leading cause in women aged 20 to 39 years[49]. The only found risk factors for developing brain tumors involve exposure to therapeutic radiation given to treat other conditions, and rare genetic diseases caused by mutations i.e. Li-Fraumeni Syndrome, Neurofibromatosis Type 1 and 2, and Turcot Syndrome[50]. Brain tumors are named by the type of cells they are thought to originate from, or sometimes from their growing location. Gliomas are a group of tumors that originate from glial cells, and include Ependymomas, Oligodendrogliomas and Astrocytomas. Astrocytomas are the most common, and originate from astrocytes (or astroglia), which function as support cells around the neurons in the brain[51]. They are divided into four grades, of which Astrocytoma grade I is regarded as benign, and prognosis decreases with increasing grade.

3.1 Glioblastoma

Glioblastoma (GBM), or grade IV astrocytoma, is the most common primary malignant brain tumor in adults, with a median age of diagnosis of 64[52]. It is highly aggressive and is characterized by cell proliferation, diffuse infiltration, *necrosis* i.e. unnatural cell death, and *angiogenesis* i.e. the formation of new blood vessels supplying the tumor with blood[53]. Median survival time after diagnosis is around 15 months, despite treatment including surgery, radiotherapy and chemotherapy[54]. There is therefore

great room for improvement when it comes to treatment of glioblastoma patients. A majority of the glioblastomas arises without being developed from a less malignant tumor type. However a small amount, termed *secondary glioblastomas*, are developed from lower grade astrocytomas and normally occurs in younger patients[53].

3.2 Glioblastoma subtypes

Verhaak et al.[36] defined four subtypes of glioblastoma, based on the molecular profiles of the tumors. The characteristics of the *Classical* subtype include amplification of chromosome 7 together with deletion of chromosome 10, highly increased levels of EGFR, deletions of the CDKN2A gene and lack of mutations of TP53. The *Mesenchymal* subtype is characterized by decreased expression of the NF1 gene, increased expression of genes in the tumor necrosis family and is associated with poor survival. The *Neural* subtype displays high expression of neural markers. The *Proneural* subtype harbors increased expression of PDGFRA and OLIG2, and mutations of TP53. A subgroup of the Proneural samples also has mutations in the IDH1 gene, and is further classified as belonging to the glioma-CpG island methylator phenotype (G-CIMP) and thus displays hypermethylation in very many locations[55]. The G-CIMP subgroup is associated with better survival compared to the other subtypes. Both the Classical and Mesenchymal subtypes showed response to aggressive therapy by increased survival times, which was not seen in the Proneural subtype[36]. This illustrates that molecular profiles of tumors can have an important role in determining when it is worthwhile to proceed with aggressive treatment.

The development of new statistical analysis methods that help to deepen the understanding of how the different layers of data are connected will assist in gaining knowledge of the biology underlying the formation of glioblastoma tumors. The end goal is to be able to offer new treatment and diagnostic options that improve the chances for longer survival for glioblastoma patients.

4 Network modeling of cancer

To make sense of all the collected data in the cancer genome projects, and actually make a difference for cancer patients, the data needs to be thoroughly analyzed. Through new visualization tools and modeling methods of the complex data structure, the hope is to gain better understanding of the mechanisms of cancer formation and progression and thereby also identify new possible drug targets. Other aims are to make better predictions of the likelihood of tumor recurrence and metastasis and find new biomarkers for early detection of cancer disease. Ultimately these efforts aim to offer improved prognosis estimates and the possibility to make individualized treatment decisions.

In the papers of this thesis we have chosen to use *network estimation* as a tool for exploration of large heterogeneous cancer genomic data sets. A network model, Figure 2, consists of nodes, representing data variables, connected by links (edges) representing associations between the nodes. Depending on the method, see below, and underlying data it is sometimes possible to infer causality represented by a directed network[56]. Mostly however, it is only possible to infer association (undirected network). The links can also be signed, indicating negative or positive associations between the variables.

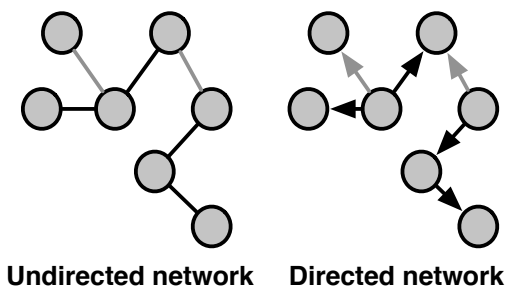

Figure 2: Black link = positive association, grey link = negative association.

Network models are suitable for several of the aims of cancer genome research by being able to capture associations between a large set of variables and present multidimensional data in a visually explorable way. The models have the potential of resulting in the proposal of new drug targets, pinpointing of groups of variables being predictive of survival, discovery of new associations between data points, and identification of common and individual properties across cancer types. This chapter presents the framework for the network modeling methods used in this thesis. Practical issues regarding data handling, preparation and normalization are discussed in Chapter 5.

4.1 Network estimation methods

There are several families of network estimation methods, of which some are presented below:

Information-theory-based methods use mutual information, which is a statistical measure that gives information about how much knowledge of one random variable reduces uncertainty about another variable and vice versa. One such method is ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks[57]). According to their proponents, information-theory methods are suited for biological applications, since they do not assume linear relationships between the variables[56].

A Bayesian network represents the probabilistic relationships between the variables and is constructed by searching for a network with a high posterior probability[58]. Some advantages of Bayesian networks are that they naturally handle missing values and, unlike the other methods, infer causal relationships. The application are mainly focused on smaller networks or structures as the construction of Bayesian networks is computationally heavy compared to for example correlation-based methods[59].

Correlation-based methods calculate the correlation coefficient (e.g. Pearson or rank correlation) between all pairs of variables and retain only the strongest associations, after different types of thresholding[59]. Advantages of correlation-based methods include that they often are fast and able to handle large data sets. As further discussed in the next section, the resulting networks will contain both indirect and direct associations between the variables, potentially resulting in dense networks that are hard to interpret. One example of a correlation-based method is WGCNA (Weighted

Correlation Network Analysis[60]).

4.2 Partial correlation estimation

Partial-correlation estimation methods are based on the theory of Gaussian Graphical Models (GGM). When the partial correlation between two variables is zero the variables are conditionally independent, given all other variables. If the partial correlation is non-zero and therefore represented by a link in the network, there is a direct interaction between the variables, when the effect of all other variables is controlled for. As opposed to correlation methods, which measure both direct and indirect associations, the partial correlation network will only include direct interactions between variables. A link is represented by a non-zero entry in the so-called *precision* matrix, which is equal to the inverse of the correlation matrix between all variables.

When the number of samples n is much smaller than the number of variables p, which is the case when working with genome-scale measurements, finding the precision matrix through direct inversion of the correlation matrix is not possible since the correlation matrix is singular. One option then is to enforce a sparse estimate of the precision matrix, i.e. it has few non-zero elements. This is also attractive for the interpretation of the resulting networks; we want the strongest interactions to emerge to be able to infer relevant biology from the model and because a fully connected network is uninformative.

Different methods have been presented for efficient estimation of the sparse precision matrix. Meinshausen and Bühlmann[61] presented an approximation that uses penalized regression on each node. Element ij of the precision matrix is set to be nonzero if either the coefficient of variable i on j, or the coefficient of variable j on i, is estimated to be nonzero. Another option, used in Paper II and III, is to estimate the sparse inverse correlation matrix by maximizing the penalized log likelihood[62]:

$$l(\Theta) = \ln(\det(\Theta)) - \mathrm{tr}(S\Theta) - P(\lambda, \Theta), \tag{4.1}$$

where $S = 1/nX^TX$ is the empirical correlation matrix, X is the $n \times p$ $N(0, \Sigma)$-data matrix, here assumed to be centered, and $\Theta = \Sigma^{-1}$. P is the penalization function which constrains Θ and is tuned by the variable λ. Different suggestions for the optimal penalization function P have been presented, of which some are outlined next.

The ***lasso***[63] penalty is in the case of graphical models (graphical lasso, ***Glasso***[62]) defined by:

$$P(\lambda, \Theta) = \lambda_1 ||\Theta||_1 = \lambda_1 \sum_{i \neq j} |\theta_{ij}| \tag{4.2}$$

This penalty controls the number of non-zero partial correlations in the model, with increasing values of λ resulting in increasing number of zeros. Glasso was applied in Paper II on a correlation matrix, S, based on mRNA, CNA, methylation, miRNA, mutation and clinical data from glioblastoma. The construction of S, in practice, with data of multiple types is discussed in the summary of Paper II, Chapter 7.

The ***Elastic Net***[64] penalty is defined by:

$$P(\lambda, \alpha, \Theta) = \lambda_1 \sum_{i \neq j} (\alpha |\theta_{ij}| + (1 - \alpha)\theta_{ij}^2) \tag{4.3}$$

This penalty is beneficial when variables are strongly correlated; by using the Elastic Net these variables tend to be zero or not simultaneously. $\alpha = 1$ is equivalent to the Glasso model and $\alpha = 0$ to the Ridge penalty.

The ***Ridge penalty***[65] does not produce a sparse model but only shrinks the variables towards zero, and is therefore not as suitable for estimation of sparse models.

Several methods for simultaneous estimation of network models for *multiple classes* of samples, e.g. cancer types, have been presented during that last couple of years. This can be done under the assumption that all, say K, classes share the same parameters. To simultaneously analyze multiple classes aims to highlight common structures at the same time as capturing the diversity between the classes.

Danaher et al.[66] presented the ***fused graphical lasso*** that encourages links to be equal across classes, by adding a term to the glasso penalty function so that the penalized log likelihood becomes:

$$l(\{\Theta\}) = \sum_{k=1}^{K} n_k [\ln(\det(\Theta^k)) - \mathrm{tr}(S^k \Theta^k)] - \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i \neq j} |\theta_{ij}^{(k)} - \theta_{ij}^{(k')}|, \tag{4.4}$$

where λ_2 also is a tunable parameter that regulates how similar the networks for the different classes should be. For large λ_2-values all links are

equal between the K estimated networks. The fused graphical lasso problem can efficiently be solved by a method called Alternating Directions Method of Multipliers (ADMM), presented in Section 6.3.

In paper III we substitute the usual lasso penalty with the Elastic net and, following Danaher, add a fused penalty, resulting in:

$$P(\Theta) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} (\alpha |\theta_{ij}^k| + (1-\alpha)(\theta_{ij}^k)^2) + \lambda_2 \sum_{k<k'} \sum_{i \neq j} |\theta_{ij}^k - \theta_{ij}^{k'}| \quad (4.5)$$

We applied this model on mRNA, CNA, methylation, miRNA and mutation data from eight cancers publicly available in TCGA.

5 Cancer as a big data problem

5.1 Data types

This chapter presents the data types used in the papers of this thesis and discuss general and data type specific complications of handling large cancer datasets. Confounding factors regarding the quality of measurements on samples from tumors include tumor heterogeneity which both can dilute signals and mean that data from different parts of the tumor may represent different subclones. Another factor is the potential mixture of non-tumor cells in the samples[67]. The biological functions of the measured entities are also discussed in Section 2.1.

5.1.1 mRNA

Established methods for measuring mRNA levels include the use of hybridization microarrays, a technique introduced in 1995[68]. Short probe sequences of DNA or RNA, designed to match specific genes, are printed on a solid surface, or attached to small beads. Complementary nucleotides (cDNA or cRNA), converted from mRNA of the sample, is hybridized to the probe surface under high-stringency conditions. A perfect probe-target hybridization match is detected by fluorophore or chemiluminescence.

In the last couple of years, sequencing of DNA and RNA have dropped in cost, and now RNA sequencing (RNAseq) is commonly used for measuring levels of mRNA. Briefly[69], RNA is converted to cDNA of which the exact sequence is determined using high-throughput sequencing. In paper I, only microarray data has been used, but in paper II and III both microarray and RNA sequencing data has been used.

In paper II and III, apart from the normalization done by TCGA, all RNAseq data has been $\log_2$ transformed and all mRNA arrays have been quantile normalized, across samples from the same cancer and platform. Quantile normalization ensures that the distributions of values for all arrays are the same. Furthermore, the amplitude of the mRNA levels were

standardized, each gene was centered around its mean expression level and divided by its standard deviation across the samples for the same cancer and experiment platform. We evaluated the effect of each transformation step by studying the distributions of the mRNA values and the cross-correlations between them.

5.1.2 miRNA

miRNAs can be detected in the same manners as mRNAs, by designed microarrays or by RNA sequencing. As miRNAs are involved in silencing of mRNAs a negative correlation between them can be expected. The prediction of miRNA targets can be done by sequence matching[70]. One summary database for target predictions is miRbase[71], using the method miRanda[72] that uses a scoring system to grade how well the miRNA sequence match the target and subsequently looks for target sequence conservation of at least 90% across mammal species.

5.1.3 CNA

Copy number aberrations (CNAs) are measured by microarrays, where the probe DNA sequences are designed to be more or less evenly spread across the genome. The chromosomal DNA is cut in smaller pieces and are allowed to attach to the probe sequences. The amount of emitted light is measured and the quantity of attached DNA to each probe is inferred. The data is noisy, so computational methods are used to judge where the duplicated or deleted segment starts and ends, and how many copies there are.

In paper I, CNA data from Agilents 244k CGH (Comparative Genomic Hybridization) array was used. In paper II and III, copy number data from Affymetrix Genome-Wide Human SNP Array 6.0 was used. SNP arrays are designed to detect single nucleotide polymorphisms, i.e. the probes are designed to match locations in the genome where there is a nucleotide known to vary between individuals, but are also being used to find CNAs from the intensity measures of the probes.

TCGA level 3 data supplies information about start, end and amplitude of CNA segments in each sample. As we want the copy number of each gene we have mapped the gene positions (of NCBI build 36.1) to the segments and assigned the amplitude to each gene. A correlation-based model including variables defined as the copy number of separate genes will consist

of a large number of links between genes located in the same CNA segment. Another approach would have been to use the segment as variable instead of the gene. Unfortunately it is then hard to define the variables, as most patients have differing start and end positions of the CNA segments.

5.1.4 DNA point mutations

DNA point mutations are on a large scale found by DNA sequencing. In the TCGA case, three centers, Broad Institute, Baylor College of Medicine and Washington University School of Medicine, are separately performing whole exome sequencing on both tumor sample and either blood or non-malignant tissue from the same patient. The normal sample are used as a reference to distinguish somatic mutations from germline. The discrepancies in mutation calls between the different centers are substantial[73], and are a problem yet to be resolved. Furthermore, the information provided from the centers regarding how the analysis were performed are very limited. We chose to be deliberately inclusive and used the union of mutation calls done by the three centers. Silent mutations were ignored in the analysis, and a gene was flagged whether or not it contained a mutation.

5.1.5 DNA methylation

DNA methylation levels are measured large scale for the TCGA project by Illumina Infinium Methylation assays. These arrays have earlier contained 27000 probes; the current version consists of 485000 probes spread out on gene-populated areas in the genome. The arrays use a two-color technique, where unmethylated attached DNA emits one color and methylated DNA emits the other color. The relationship between methylated and unmethylated DNA is measured as[74]:

$$\beta = \frac{\max(y_{meth}, 0)}{\max(y_{unmeth}, 0) + \max(y_{meth}, 0) + 100}, \tag{5.1}$$

where y_{meth} and y_{unmeth} are the emission intensities, $\beta = 0$ means completely unmethylated, and $\beta = 1$ means 100% methylated. As the methylation data consequently not follows a normal distribution, we have chosen in paper II to BoxCox transform the β-values, and in paper III to use rank correlation instead of Pearson correlation. As many methylation sites are not varying at all across samples, we chose to keep probes in the analysis with a standard deviation across the patients > 0.05.

5.2 Data magnitude

One challenge that arise before construction of the cancer network models is the *practical handling* of the very large datasets on a personal computer. One downloaded folder from TCGA, including all 450k methylation data for all patients from, for example, breast cancer is as large as around 17 GB. To avoid having to harbor data from multiple cancers on a personal computer hard drive, we instead downloaded the available data from TCGA and stored it in a mySQL database on a local server. It was then possible to query the database, at any given time, to get the currently needed data matrices. The mySQL database also helped to increase speed of data preparation, since it enabled extraction of a subset of the data.

5.3 Heterogeneity of data

In addition to data magnitude, a second complicating factor is the *heterogeneity* of the data. Since the application of the methods in this thesis focuses on data from The Cancer Genome Atlas, examples will be taken from that particular setup.

As the data collection for TCGA has been going on since before 2008, the techniques for large scale measuring have improved and dropped in cost massively during the project. For the cancers being investigated in the beginning of the project, e.g. glioblastoma, the used platforms are older, and the samples have not been reanalyzed with new methods at the time of the preparation of data for the papers of this thesis. The variation in coverage of different platforms complicates the comparison between cancers, as decisions have to be made on how to handle data that is not present everywhere. Additionally, TCGA provides the results given the gene *names* provided by the supplier of the platform technology. Unfortunately, since gene nomenclature is not completely unified, the result is that the same gene can be named in different ways depending on which platform the data comes from, reflecting when the annotation files were constructed. Where multiple mRNA platforms have been used, we have used the intersect of the included variables to ensure that all genes in the model are available in all platforms. The CNA gene variables were then selected to match the mRNA set.

The NCBI (National Center for Biotechnology Information) currently

is responsible for the genome assembly containing the human reference sequence. A genome assembly is the attempt to align together the short DNA sequences read by sequencing technology into the correct chromosomes and order. This assembly is then the basis for where on a chromosome a gene is situated. As many genomes include repeated sequences, the assemblies are continually updated when additional measurements are done. These assemblies are, in the human case, referred to as NCBI builds, and are labeled e.g. *Human Annotation Release 101* or *NCBI Human Build 38.1*. Results from different platforms are not always presented in TCGA using the same versions of the human genome build. This complicates comparison regarding data types that are dependent on position on the genome, like methylation and CNA. One workaround is to use a map translating the chromosomal positions between the builds. Nevertheless, this is a potential source of error.

Often, the data collected from a patient is not complete; instead there is a lack of measurements for one or several types of data. The tumor sample might have been too small or of too low quality, or some other laboratory step may have failed. In the correlation matrices we have chosen to use the maximum number of patients available for the specific combination of variables, even if some patients have missing data for other variables.

6 Estimation of network models

The previous chapter discussed practical issues regarding preparation, unification and handling of the large heterogeneous cancer data sets. This chapter address the central problems that needed to be solved for the network construction to be feasible. These problems include how to be able to produce a robust result and present balanced models with respect to datatype. Another central issue has been the time complexity of the estimation of the networks.

6.1 The bootstrap

To improve robustness of the network estimates we make use of bootstrapping (Paper I and III) and aggregate the results. Network estimation is repeated 500 or 1000 times with randomly chosen samples from the full set (details in papers). For each link, the proportion of network estimates where it is present is calculated. A link is then included in the final network if this proportion exceeds a threshold T.

To investigate how the number of bootstraps affected the stability of the networks, we, for Paper III, compared two networks summarized from 10 to 250 bootstrapped data sets. The Jaccard index, measuring the similarity in link presence between the two network estimates, reached a plateau of 0.85 at around 200 bootstraps, indicating that construction of the final network from 500 bootstrapped data sets was sufficient.

Bootstrapped networks can also be used in the validation of choice of penalty parameters (Paper II), where the stability between network solutions, based on different bootstrapped data sets, was used as a measure for how to tune the penalty.

The ability to perform many bootstraps is highly limited by the available computer power, and the dimension of a single network. Since the network estimation can be done independently for each bootstrap it is an easily parallelizable problem. Parallelization using cluster computers is further discussed in Section 6.4.

6.2 Introduction of priors in penalties

In the partial correlation models for Paper II and III we combine large data sets from multiple sources, which can result in models that are hard to survey due to dimensionality and complexity. To highlight associations that are more likely based on biological knowledge, or of more interest, we introduce a link-specific prior in the the glasso penalty (Paper II) and the fused graphical lasso penalty with elastic net (Paper III). The prior is introduced in the lasso penalty part, by replacing the scalar λ_1 by a matrix of size $p \times p$:

$$P(\lambda, \Theta) = \sum_{i \neq j} \lambda_{1,ij} |\theta_{ij}|, \tag{6.1}$$

where $\lambda_{1,ij} = \lambda_1 v_{ij}$. v_{ij} assumes different values depending on the type of link, e.g. ∞ for implausible/irrelevant associations, 1 for neutral associations, and $u < 1$ for associations that are biologically plausible. The number of different values of u in a given λ_1-matrix should be small for the sake of simplicity and for the possibility of validation of tunable parameters.

In Paper II we evaluate the stability of the network model for different prior parameter settings by calculating Kendalls W [75] (Kendalls coefficient of concordance) over network estimates based on bootstrapped data sets. For Paper III we investigated the sensitivity of the model with respect to the tuning parameter u and established that mild tuning of u had moderate effects on the resulting networks.

As previously have been mentioned in Section 5.1.3, the variables representing the copy number level of genes localized closely in the genome will generate many links, as these variables are likely to be highly correlated. To avoid that clusters of that kind dominate the models and balance the models with respect to data type representation we also use the penalty matrix to remove such links by setting $v_{ij} = \infty$.

6.3 ADMM algorithm

The Alternating Directions Method of Multipliers (ADMM) algorithm [76,77] builds on a concept where a complex objective function, hard to solve as a whole, is decomposed into components that are relatively easily solved separately. Danaher et al. [66] present an adaption of the ADMM algorithm to solve the fused graphical lasso problem for comparative networks, used in

paper III. This section briefly present the framework for ADMM algorithms in general, details of the adapted version for paper III can be found in the Appendix.

The ADMM algorithm was constructed to solve convex optimization problems of the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in C \end{aligned}$$

with variables $x \in R^n$ and f and C are convex. This problem can be rewritten as:

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } x = z \end{aligned} \tag{6.2}$$

where g is the indicator function of C (i.e. $g(z) = 0$ if $x \in C$, ∞ otherwise).

This problem can be rewritten in the form of a scaled augmented *Lagrangian*[77]:

$$L_\rho(x, z, u) = f(x) + g(z) + (\rho/2)||x - z + u||_F^2 \tag{6.3}$$

where u is a so-called *Lagrangian multiplier vector* or *dual variable*.

The ADMM algorithm consists of iterating through three steps. An approximate solution is achieved by first minimizing L for x holding z fixed, and then solving for z holding x fixed. In Step 3, the dual is updated, ensuring that x and z converge towards each other:

$$1: \; x_m \leftarrow \underset{x}{\operatorname{argmin}} \; L_\rho(x_{m-1}, z_{m-1}, u_{m-1}) \tag{6.4}$$

$$2: \; z_m \leftarrow \underset{z}{\operatorname{argmin}} \; L_\rho(x_m, z_{m-1}, u_{m-1}) \tag{6.5}$$

$$3: \; u_m \leftarrow u_{m-1} + x_m - z_m \tag{6.6}$$

In the adaption of the ADMM algorithm to the fused graphical lasso problem, see Appendix, the main bottleneck is Step 1 including eigendecomposition of large matrices. As we chose to implement the algorithm in Matlab, the most effective method to solve the eigendecomposition turned out to be the Matlab inbuilt function *eig*. Step 2 includes elementwise operations on all the matrix elements. Appendix includes a sketch on how vectorization can be used to rewrite Step 2 as matrix operations instead of loops going through $p \times p$ elements. As Matlab is slow at elementwise looping and fast at making matrix calculations, the vectorization led to massive speedup of Step 2.

6.4 Parallelization

As mentioned above, the bottleneck of the algorithm for paper III was the step of calculating the eigenvalue decomposition. Yang et al.[78] show that if the matrices $\theta^{(k)}$ can be reordered in such a way that they become block-diagonal (with the same size of the blocks for all classes k) with

$$\Theta^{(k)} = \begin{pmatrix} \theta_1^{(k)} & 0 \\ 0 & \theta_2^{(k)} \end{pmatrix}, \tag{6.7}$$

then one can solve the optimization problem for each of the blocks. This greatly reduces the computational time as the calculation time for the eigenvalue function eig grows exponentially with number of variables. Sufficient conditions for dividing the problem into subblocks θ_1 and θ_2 are[78]:

$$\begin{cases} |\sum_{k=1}^{t} S_{ij}^{(k)}| \leq t\lambda_1 + \lambda_2, \\ |\sum_{k=0}^{t-1} S_{ij}^{(r+k)}| \leq t\lambda_1 + 2\lambda_2, \ 2 \leq r \leq K - t, \\ |\sum_{k=1}^{t} S_{ij}^{(K-t+k)}| \leq t\lambda_1 + \lambda_2, \\ |\sum_{k=1}^{K} S_{ij}^{(k)}| \leq K\lambda_1 \end{cases} \tag{6.8}$$

for all $i \in \theta_1, j \in \theta_2, t = 1, \ldots, K-1$.

This not only enables faster calculations, it also means that each problem can be run in parallel on different processors. For this, either a personal computer with multiple processors with e.g. Matlab's inbuilt parallelization program, or a computer cluster can be used. Benchmarking tests is called for, however, in the case of using a personal computer, to check that the overhead time of sending data do not exceed the time gained in the calculations. On a computer cluster it is suitable to send each bootstrap network to a separate cluster node, and in that way remove the need of sending data between the nodes. The execution time per bootstrap for a sequence of 12 values for λ_1 and 8 values for λ_2 varied between 24 and 30 hours, using a computer cluster located at Chalmers University of Technology consisting of 268 nodes, with 8 cores each.

7 Summary of papers

7.1 I: Network modeling of the transcriptional effects of copy number aberrations in glioblastoma

When the work for paper I started, the collection of data for The Cancer Genome Atlas had just begun. The first available data were mRNA and CNA measurements and clinical data for a set of 186 glioblastoma patients. Given the opportunity to analyze these data, we set out to create a network model that connects the DNA copy number aberrations to the transcriptome. The goal was not only to investigate the well-documented direct local effects of increased or decreased copy number on the expression of the genes in the same locus in form of increased or decreased expression[79,80,81]. Our models instead also aimed to catch regulation of expression of genes localized outside the CNA by indirect mechanisms, for example by the deletion of a transcriptional repressor that increases the expression of its targets or the amplification of a kinase that drives a signaling cascade. Global network models based on the combined data of mRNA and CNA from the same samples could potentially be used to identify genes whose copy number aberration have an impact on expression levels of other genes, propose possible drug targets by matching model-identified regulators or their targets to pharmacological databases, and pinpoint CNA and mRNA features that are predictive of patient prognosis.

The framework for creating the network models required preprocessed and normalized data matrices in a format that the developed algorithm could use. Furthermore, in the case of mRNA, we were presented with data from two different platforms, Agilent 44k and Affymetrix U133. As it has been shown to stabilize the signals[36], the mRNA levels were averaged across the two platforms. In the case of CNA, we used discretized estimates of the copy number of each gene. The datasets were synced in matter of patients and genes, so that patients and genes not present in all datasets were discarded from the analysis.

We introduced a set of differential equations modeling the change rate of mRNA levels as the difference between the *synthesis rate*, proportional to the genes own copy number and regulatory effects of other mRNAs, and the *decay rate*, proportional to regulatory effects of other mRNAs. The steady state solution to the set of differential equations could be seen as two linear systems, both of which could be represented by a network with different interpretation:

1. $A\Delta Y + \Delta U + R = 0$
2. $\Delta Y = G\Delta U + \Gamma,$

where ΔY and ΔU denote matrices containing the log-transformed and centered mRNA and CNA profiles, respectively. R and Γ are treated as noise in the estimations. A represents the transcriptional network of interactions a_{ij} between transcript i and transcript j, after correcting for the impact of the CNA of each gene. G represents the CNA-driven transcriptional activation or inhibition g_{ij} of CNA i on transcript j. A relates to G by $G = -A^{-1}$.

In short, the G network was estimated by gene-level lasso regression, repeated 1000 times on pseudo-bootstrap data sets to improve robustness. First, the direct effect of each genes CNA on its transcript was estimated by

$$d = max(0, \Delta U_i^T \Delta Y_i), \tag{7.1}$$

where ΔU_i, ΔY_i represent the transposed row i of ΔU, ΔY.

Second, the following lasso problem was solved for each gene i:

$$\min_{G_i} ||(\Delta Y_i - d\Delta U_i) - \Delta U_{H\setminus i}^T G_i||_F^2 + \lambda \sum_{j \in H\setminus i} |G_i[j]|, \tag{7.2}$$

where H is a set of candidate hub CNA genes, where CNAs that show no selection towards either deletion or amplification have been filtered out.

After estimation of the network models, experimental follow-up was done in four glioblastoma-derived cell-lines on a small subnetwork, including the recurrently copy number deleted gene NDN (Necdin) which was connected to five mRNA transcripts in the G network. NDN belongs to the melanoma-associated antigen (MAGE) family and is shown to interact with the p53 protein [82]. To investigate the importance of NDN in glioblastoma, NDN was overexpressed and the growth of treated and untreated

cells was estimated. To test for difference in growth we used a t-statistic:

$$T = \frac{k_{CTRL} - k_{NDN}}{\sqrt{s^2_{CTRL} + s^2_{NDN}}} \sim t_{2n-4}, \tag{7.3}$$

where k is the slope of the log transformed growth curve:

$$log_2(h(t)) = log_2(h_0) + kt, \tag{7.4}$$

t = time in days, h is proportional to number of cells, h_0 is number of cells at $t = 0$, and s is the standard deviation of k.

Overexpression of NDN was shown to decrease cell cycling time in three of four cell lines, and it was confirmed as a hub in the network. Thus, the CNA-driven network model could be used to find unknown regulators of cell growth in glioblastoma.

7.2 II: Integrative modeling reveals ANXA2 as a determinant of mesenchymal transformation in glioblastoma

The collection and publication of data in TCGA moved forward and the full set of measurements for around 600 glioblastoma patients was presented. The data included not only mRNA and CNA estimates, but also measurements of methylation, miRNA, and mutations. We developed a more general network model for integration of all presented data types, including clinical data such as survival and affiliation to subtype[36]. The aim of the model was to identify possible genomic, epigenetic and transcriptional regulators of the four glioblastoma subtypes; classical, neural, proneural, mesenchymal. As the subtypes have different properties in term of aggressiveness and response to therapy, it is of importance to learn about underlying mechanisms by, for example, finding candidate drivers. Previous research on the activation of the mesenchymal gene signature has identified the transcription factors C/EBP-β and STAT3[83], and overexpression of TAZ[84] and RHPN2[85] as being associated with the mesenchymal subtype. Additionally, the miRNAs miR-128a and miR-504[86] have been identified to negatively correlate with mesenchymal marker genes.

For network construction to be feasible in Matlab, the data was stored in data type specific matrices, of size *number of variables* $\times$ *number of*

patients. The full correlation matrix was then built up of blocks:

$$S = \begin{bmatrix} S_{11} & \dots & S_{1d} \\ \vdots & \ddots & \vdots \\ S_{d1} & \dots & S_{dd} \end{bmatrix}, \tag{7.5}$$

where each block S_{ab} is the cross-correlation matrix between data type a and b, and which uses the maximum number of patients available for that combination of data types.

The network models are created using *glasso*[62], with the addition of a datatype-specific prior implemented in a link-specific λ_1-penalty variable. This model can be seen as a modified single-cancer version of the aSICS model presented in Paper III. The prior is necessary for balancing the model when including multiple data types. The specific prior used in this paper was:

$$\Lambda_{ij} = \lambda_1 \lambda_{ab}^{block} \lambda_{ij}^{link}, \tag{7.6}$$

where λ_1 is a general sparsity parameter, tuning the global sparsity of the network. λ_{ab} is used to tune the sparsity of variables from *different* data types a and b; when $a = b$, then $\lambda_{ab} = 1$. λ_{ij} is a link specific sparsity parameter, allowing to alter the penalty for specific pairs of variables i and j, e.g. when i is the variable for promoter methylation of a gene with mRNA j.

The resulting network model included 19 potential regulators of glioblastoma subtypes, defined as a methylation, miRNA or mutation node connected to an mRNA, subsequently linked to a subtype. We chose to further investigate, and experimentally validate, one of these regulators, Annexin A2 (ANXA2) which was linked in the network to the mesenchymal subtype and a methylation site located in the ANXA2 promoter.

We investigated the expression and methylation status of ANXA2 in TCGA lower grade glioma (LGG) and GBM datasets, and showed that expression increases and methylation decreases with grade of glioma. RNAseq values were used for the expression comparison, since other platforms were not available for LGG; although it meant there were fewer patients available for GBM, illustrating the trade off that sometimes has to be done to be able to compare datasets. We also compared ANXA2 expression of glioblastoma cell lines and tumor samples from University of Freiburg and there was a significant difference between mesenchymal and not mesenchymal samples, where mesenchymal had higher expression.

Gene Set Enrichment Analysis (GSEA[87,88]) is a computational method used to investigate whether a defined set of genes differs significantly between two groups of samples, e.g. before and after treatment. We used GSEA to investigate the effect on mesenchymal subtype signature mRNAs before and after knockdown of ANXA2 with shRNA in two glioblastoma-derived cell lines, and showed that there was a significant downregulation of the group of mesenchymal signature transcripts after knockdown. The mesenchymal signature genes were taken from the subtype definition by Verhaak[36]. Knockdown of ANXA2 by shRNA also reduced cell proliferation and invasiveness.

In summary, network modeling of multiple data types and the inclusion of subtypes as nodes made new interesting predictions about potential subtype regulators in glioblastoma. As the knockdown experiments displayed both an effect on the mesenchymal signature genes, and on physical properties like proliferation and invasiveness, ANXA2 might have the potential of being a new therapeutic target. Since the mesenchymal phenotype is associated with poor survival, the hope is that a reduction of the mesenchymal profile genes possibly could decrease the tumor invasiveness and thereby improve prognosis for the patients.

7.3 III: Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content

Following the publishing of the full data set for glioblastoma, TCGA collected and presented a series of other cancer diagnoses, which eventually became part of the Pan-Cancer initiative[44] (see Section 2.2.1). By the time of eight published cancer types with at least 200 patients each, we developed a network model harboring not only multiple data types, but also multiple cancer types. Our application of partial correlation-based modeling presents a variable focused alternative to multi-platform analyses that use patient-based clustering methods[47].

The model is based on an extension of sparse inverse covariance selection (SICS), which was adapted and optimized for modeling of genetic, epigenetic, and transcriptional data across cancers, see Section 4.2, Chapter 6, and Appendix. To evaluate the potential of the method, we applied it to the eight TCGA cancers and published the model online at cancerland-

scapes.org, using a web interface that was designed to enable navigation and exploration of the networks. The resulting network models showed good performance in terms of enrichment of known interactions in PathwayCommons (http://www.pathwaycommons.org), also when compared to the correlation based network estimation method WGCNA [60].

Simultaneous analysis of multiple cancer types aims to highlight similarities and differences between them. To be able to conduct the network construction it is absolutely necessary to align all the different datasets so that the same variables are present across the cancers. As described above, the discrepancies between data coming from different platforms enforce a sequence of choices. These choices include what to do with missing data, both regarding variables and samples, which normalization to choose and how to filter variables to reduce data dimensionality. A select choice of stability analyses of parameters included in the network construction were done, including sensitivity analysis with respect to the elastic net parameter α and the choice of prior strength u (Section 6.2), and stability of networks with respect to the number of bootstraps (Section 6.1). Nonetheless, since the dependencies are complex in this kind of multi-step procedure, this issue warrant further study in future work.

The derived multi-cancer network detected known interactions in pathway databases and contained interesting predictions, including functionally coupled network structures shared between cancers. Examples include a network module enriched for mitosis related genes mainly shared by glioblastoma, head and neck, kidney and uterine cancers, and a network module enriched for immune response dominated by head and neck, uterine and lung cancer. The network model also contained modules specific to singular cancers, e.g. TP53 point mutation linked to a number of TP53 targets in uterine cancer. Another cancer specific network module represents mutation of IDH1 linked to over 600 methylation probes in glioblastoma, probably reflecting the IDH1 driven hypermethylated G-CIMP phenotype discussed in Section 3.2. The same module also proposes a link between loss at the end of the short arm of chromosome 11 and mutation of IDH1. Furthermore, overlaying the network with survival association and drug target information highlighted for example the estrogen receptor ESR1 in breast and ovarian cancer, linked to the known modulators of estrogen receptor signaling (GATA3, EYA2) [89,90] and the gene GPR77 previously being unreported in the context of the estrogen receptor.

Pan-Cancer studies focusing on one datatype have resulted in estima-

tions of the number of significantly mutated cancer driving genes[45,46], common patterns of copy number alterations[48], and similarities in miRNA regulation across cancers[91]. Our multi-cancer network model has the potential of giving new insight to cancer biology and aid in the hunt for similarities and differences across cancers also by connecting the different data types to each other.

8 Conclusion and future perspectives

The papers of this thesis show that network modeling of cancer genomics data from multiple sources can aid in the search for unknown regulators, and that simultaneous analysis of multiple cancers enable identification of common and unique properties across diagnoses. The developed methods should be regarded as one addition to the range of analysis tools that assist in the creation of a map covering the whole variation of diseases that is cancer. Unless large parts of the map are explored, it is not possible to pinpoint the location of the single individual and reach the ultimate goal of being able to offer personalized medicine where the treatment is adapted to the aberrations and conditions of each patient.

The modeling framework of paper II and III make it possible to add more types of data, the most obvious being protein level measurements of tumors. As the number of samples and cancer diagnoses increase in TCGA, it will be interesting to further develop the models to encompass not only more types of cancer but allow for division into molecular subtypes, as subsets of tumors have been shown to molecularly resemble tumors with another histological classification[47]. The aim is also that the models in the future should be able to incorporate data from other databases and sources than TCGA.

Towards the goal of creating network models encompassing more variables and diagnoses, and that are surveyable and informative, there is potentially a need both to redesign the models, for example by redefining what a variable represents, and to finetune the presentation and communication of the results. It would be interesting to investigate the use of alternative types of association summaries such as *low order partial correlations*, that condition on a subset of the other variables rather than the full set[92]. Another option is the estimation of directed acyclic graphs (DAG), using for example the PC algorithm[93], that instead of measuring the effect on variable i of variable j by keeping all the covariates fixed, includes the indirect effect of all variables. Further alternatives include to introduce more data types in a regression based model with genetic and epigenetic alterations as input and mRNA or proteins as output. Another idea would be to also

account for diverse cellular composition of different tumor types in models that compare multiple cancer classes.

Regarding improvement of data quality lies the possibility of full genome sequencing, which will reveal mutations not only in gene coding regions but also in other possibly regulatory segments of the DNA, such as in long non-coding RNAs. Also, improvements in mass spectrometry analysis or array technologies will enable more global measurements of protein levels, which will provide a new level of insight to cancer mechanisms [29]. However, although computer power increases fast, and solutions for using the capacity of cloud computers emerge, a challenge lies in the development of data-handling algorithms to match the speed of data collection and hardware development.

Further ahead is the possibility of cancer genome studies that also harbor recurred tumors, metastasis and treatment information, which would open for modeling of treatment efficiency in patients with different molecular profiles. The development of measuring techniques that requires less tumor material could possibly enable separate analysis of several parts of a tumor, or even multi-platform analysis of tumor biopsies. Also, paired samples of for example blood will possibly aid in the finding of new biomarkers, to be used in screening programs for early detection of different cancer diagnoses.

Acknowledgements

First, I would like to thank my supervisor Sven Nelander for entrusting me to be your first PhD student, and for all the conversations during this somewhat prolonged process, about mathematics, biology, medicine and subjects not at all related to science. Thank you for always trusting me to do my very best, regardless of us sharing the same office or working 600 km apart. (Big thanks to Skype and Google Drive for creating a framework that enables long-distance collaborations.) The interdisciplinary environment you created has been very educational.

My co-supervisor Rebecka Jörnsten, thank you for interesting discussions and inspiring brainstorming sessions and for adding extra theoretical depth to the group and the projects.

Maria Stella Carro and Roberto Ferrarese at University of Freiburg, thank your for the rewarding and stimulating collaboration.

Thanks to Olle Nerman, for originally introducing me to Sven. Co-supervisor Frida Abel, thanks for the valuable comments in the end.

Debora Marks and Chris Sander, thanks for the support during my first stumbling year as a young PhD student.

Linnéa Schmidt, as you know, without you this thesis would not be. To work with you have teached me a lot about communication across the disciplines. Your support, pep-talks and friendship are invaluable!

Maja Ohlson, thanks for being a very reliable colleague and good friend. Thanks for the patience with my complaining days, and for saving me from my home-office.

Patrik Johansson and José Sánchez, thanks for the good collaboration and all the productive and nice conversations. The end result was not so bad after all.

Former group members Caroline Hansson, Philip Gerlee, Mariana Buongermino Pereira, Bodil Nordlander and Linda Karlsson-Lindahl, thank you for being good colleagues and friends, you made work-life more enjoyable.

Tajana Tešan Tomić thanks for the talks about life in general and kids in particular.

To all of my colleagues who, depending on your expertise, scrutinized different parts of the text, thank you for the support and the feedback!

Thanks to the people at Sahlgrenska Cancer Center for making me feel welcome and for all the nice lunches.

To friends outside of academia for being just friends, not caring particularly about my research or when I will be finished.

My parents Berit and Ingolf, thank you for always having very strong faith in my abilities and capacity. Thanks also to my brothers Olof and Alexander with families, and Einar, whom I know I can always rely on.

Finally, the biggest thanks to the three loves of my life. Emanuel, my life companion, for being my very steady rock during my storms. Leonora, my brilliant, beautiful girl, who keeps me alert by asking all the right questions. Valdemar, my stubborn, gorgeous boy, who constantly reminds me what is truly important in life.

References

[1] Cancer Research UK. All cancers combined key stats. *Cancer Research UK 2014, http://publications.cancerresearchuk.org, 2015-06-24*, 2014.

[2] Theodor Boveri. Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris. *Journal of cell science*, 121(Supplement 1):1–84, 2008.

[3] CC Twort and JM Twort. Observations on the reaction of the skin to oils and tars. *Journal of Hygiene*, 28(03):219–227, 1928.

[4] JM Twort and CC Twort. Comparative activity of some carcinogenic hydrocarbons. *The American Journal of Cancer*, 35(1):80–85, 1939.

[5] Ii Berenblum. The cocarcinogenic action of croton resin. *Cancer Research*, 1(1):44–48, 1941.

[6] Steven A Frank. *Dynamics of cancer*. Princeton University Press, 2007.

[7] HT Deelman. The part played by injury and repair in the development of cancer; with some remarks on the growth of experimental cancers. *Proceedings of the Royal Society of Medicine*, 20(7):1157, 1927.

[8] Ian MacKenzie and Peyton Rous. The experimental disclosure of latent neoplastic changes in tarred skin. *The Journal of experimental medicine*, 73(3):391–416, 1941.

[9] DJ Ashley. Colonic cancer arising in polyposis coli. *Journal of medical genetics*, 6(4):376, 1969.

[10] Alfred G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

[11] JC Fisher and JH Hollomon. A hypothesis for the origin of cancer foci. *Cancer*, 4(5):916–918, 1951.

[12] Clifford J Tabin, Scott M Bradley, Cornelia I Bargmann, Robert A Weinberg, Alex G Papageorge, Edward M Scolnick, Ravi Dhar, Douglas R Lowy, and Esther H Chang. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149, 1982.

[13] Elizabeth Taparowsky, Yolande Suard, Ottavio Fasano, Kenji Shimizu, Mitchell Goldfarb, and Michael Wigler. Activation of the t24 bladder carcinoma transforming gene is linked to a single amino acid change. *Nature*, 300(5894):762–765, 1982.

[14] E Premkumar Reddy, Roberta K Reynolds, Eugenio Santos, and Mariano Barbacid. A point mutation is responsible for the acquisition of transforming properties by the t24 human bladder carcinoma oncogene. *Nature*, 300(5888):149–152, 1982.

[15] Laura E MacConaill and Levi A Garraway. Clinical implications of the cancer genome. *Journal of Clinical Oncology*, 28(35):5219–5228, 2010.

[16] Helen Davies, Graham R Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J Garnett, William Bottomley, et al. Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–954, 2002.

[17] Thomas J Lynch, Daphne W Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A Okimoto, Brian W Brannigan, Patricia L Harris, Sara M Haserlat, Jeffrey G Supko, Frank G Haluska, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004.

[18] J Guillermo Paez, Pasi A Jänne, Jeffrey C Lee, Sean Tracy, Heidi Greulich, Stacey Gabriel, Paula Herman, Frederic J Kaye, Neal Lindeman, Titus J Boggon, et al. Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500, 2004.

[19] William Pao, Vincent Miller, Maureen Zakowski, Jennifer Doherty, Katerina Politi, Inderpal Sarkaria, Bhuvanesh Singh, Robert Heelan, Valerie Rusch, Lucinda Fulton, et al. Egf receptor gene mutations are common in lung cancers from never smokers and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of*

the National Academy of Sciences of the United States of America, 101(36):13306–13311, 2004.

[20] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

[21] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.

[22] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[23] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[24] Olle Bergman, Lotta Fredholm, Micke Jaresand, Elizabeth Johansson, and Sara Nilsson. Cancerfondsrapporten 2015. *http://www.cancerfonden.se/publikationer/cancerfondsrapporten, 2015-08-10*, 2015.

[25] N Howlader, AM Noone, M Krapcho, J Garshell, D Miller, SF Altekruse, CL Kosary, M Yu, J Ruhl, Z Tatalovich, et al. Seer cancer statistics review, 1975–2011. bethesda, md: National cancer institute, 2014.

[26] Socialstyrelsen Sverige. Dödsorsaker 2013 = causes of death 2013. *http://www.socialstyrelsen.se, 2015-04-29*, 2015.

[27] Cancer Research UK. Cancer statistics report: Survival. *Cancer Research UK 2014, http://publications.cancerresearchuk.org, 2015-06-24*, 2014.

[28] Theresa Phillips. The role of methylation in gene expression. *Nature Education*, 1(1):116, 2008.

[29] Nancy Kendrick. A gene's mrna level does not usually predict its protein level. *Kendrick Laboratories*, 2014.

[30] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[31] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68, 2015.

[32] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[33] The Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[34] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

[35] The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

[36] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

[37] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[38] Ie-Ming Shih, Kentaro Nakayama, Gang Wu, Naomi Nakayama, Jinghui Zhang, and Tian-Li Wang. Amplification of the ch19p13. 2 nacc1 locus in ovarian high-grade serous carcinoma. *Modern Pathology*, 24(5):638–645, 2011.

[39] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013.

[40] Yann Christinat and Wilhelm Krek. Integrated genomic analysis identifies subclasses and prognosis signatures of kidney cancer. *Oncotarget*, 2015.

[41] Yu-Zheng Ge, Hui Xin, Tian-Ze Lu, Zheng Xu, Peng Yu, You-Cai Zhao, Ming-Hao Li, Yan Zhao, Bing Zhong, Xiao Xu, et al. Microrna expression profiles predict clinical phenotypes and prognosis in chromophobe renal cell carcinoma. *Scientific reports*, 5, 2015.

[42] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

[43] Josien C Haan, Mariette Labots, Christian Rausch, Miriam Koopman, Jolien Tol, Leonie JM Mekenkamp, Mark A van de Wiel, Danielle Israeli, Hendrik F van Essen, Nicole CT van Grieken, et al. Genomic landscape of metastatic colorectal cancer. *Nature communications*, 5, 2014.

[44] The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[45] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013.

[46] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3, 2013.

[47] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[48] Travis I Zack, Steven E Schumacher, Scott L Carter, Andrew D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H Mermel, et al. Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, 45(10):1134–1140, 2013.

[49] Haley R Gittleman, Quinn T Ostrom, Chaturia D Rouse, Jacqueline A Dowling, Peter M de Blank, Carol A Kruchko, J Bradley Elder, Steven S Rosenfeld, Warren R Selman, Andrew E Sloan, et al. Trends in central nervous system tumor incidence relative to other common cancers in adults, adolescents, and children in the united states, 2000 to 2010. *Cancer*, 121(1):102–112, 2015.

[50] Hiroko Ohgaki. Epidemiology of brain tumors. In *Cancer Epidemiology*, pages 323–342. Springer, 2009.

[51] Damien Ricard, Ahmed Idbaih, François Ducray, Marion Lahutte, Khê Hoang-Xuan, and Jean-Yves Delattre. Primary brain tumours in adults. *The Lancet*, 379(9830):1984–1996, 2012.

[52] Quinn T Ostrom, Haley Gittleman, Peter Liao, Chaturia Rouse, Yanwen Chen, Jacqueline Dowling, Yingli Wolinsky, Carol Kruchko, and Jill Barnholtz-Sloan. Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2007–2011. *Neuro-oncology*, 16(suppl 4):iv1–iv63, 2014.

[53] Santosh Kesari. Understanding glioblastoma tumor biology: the potential to improve current diagnosis and treatments. In *Seminars in oncology*, volume 38, pages S2–S10. Elsevier, 2011.

[54] Derek R Johnson and Brian Patrick ONeill. Glioblastoma survival in the united states before and during the temozolomide era. *Journal of neuro-oncology*, 107(2):359–364, 2012.

[55] Houtan Noushmehr, Daniel J Weisenberger, Kristin Diefes, Heidi S Phillips, Kanan Pujara, Benjamin P Berman, Fei Pan, Christopher E Pelloski, Erik P Sulman, Krishna P Bhat, et al. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, 17(5):510–522, 2010.

[56] YX Rachel Wang and Haiyan Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of theoretical biology*, 362:53–61, 2014.

[57] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[58] David Heckerman. *A tutorial on learning with Bayesian networks*. Springer, 1998.

[59] Jeffrey D Allen, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. Comparing statistical methods for constructing large scale gene networks. *PloS one*, 7(1):e29348, 2012.

[60] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[61] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[62] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[63] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[64] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[65] Arthur E Hoerl and Robert W Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[66] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[67] Levi A Garraway and Eric S Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, 2013.

[68] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[69] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[70] Hao Zheng, Rongguo Fu, Jin-Tao Wang, Qinyou Liu, Haibin Chen, and Shi-Wen Jiang. Advances in the techniques for the prediction of microrna targets. *International journal of molecular sciences*, 14(4):8179–8187, 2013.

[71] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. mirbase: tools for microrna genomics. *Nucleic acids research*, 36(suppl 1):D154–D158, 2008.

[72] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, Debora S Marks, et al. Human microrna targets. *PLoS Biol*, 2(11):e363, 2004.

[73] Su Y Kim and Terence P Speed. Comparing somatic mutation-callers: beyond venn diagrams. *BMC bioinformatics*, 14(1):189, 2013.

[74] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

[75] Maurice G Kendall and B Babington Smith. The problem of m rankings. *The annals of mathematical statistics*, 10(3):275–287, 1939.

[76] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[77] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[78] Sen Yang, Zhaosong Lu, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2):916–943, 2015.

[79] Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David

Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968, 2002.

[80] Hyunju Lee, Sek Won Kong, and Peter J Park. Integrative analysis reveals the direct and indirect interactions between dna copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–896, 2008.

[81] Björn Nilsson, Mikael Johansson, Anders Heyden, Sven Nelander, and Thoas Fioretos. An improved method for detecting and delineating genomic regions with altered gene expression in cancer. *Genome Biol*, 9(1):R13, 2008.

[82] Hideo Taniura, Kuniharu Matsumoto, and Kazuaki Yoshikawa. Physical and functional interactions of neuronal growth suppressor necdin with p53. *Journal of Biological Chemistry*, 274(23):16242–16248, 1999.

[83] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P Sulman, Sandrine L Anne, Fiona Doetsch, Howard Colman, et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, 2010.

[84] Krishna PL Bhat, Katrina L Salazar, Veerakumar Balasubramaniyan, Khalida Wani, Lindsey Heathcock, Faith Hollingsworth, Johanna D James, Joy Gumin, Kristin L Diefes, Se Hoon Kim, et al. The transcriptional coactivator taz regulates mesenchymal differentiation in malignant glioma. *Genes & Development*, 25(24):2594–2609, 2011.

[85] Carla Danussi, Uri David Akavia, Francesco Niola, Andreja Jovic, Anna Lasorella, Dana Pe'er, and Antonio Iavarone. Rhpn2 drives mesenchymal transformation in malignant glioma by triggering rhoa activation. *Cancer research*, 73(16):5140–5150, 2013.

[86] Xinlong Ma, Koji Yoshimoto, Yaulei Guan, Nobuhiro Hata, Masahiro Mizoguchi, Noriaki Sagata, Hideki Murata, Daisuke Kuga, Toshiyuki Amano, Akira Nakamizo, et al. Associations between microrna expression and mesenchymal marker gene expression in glioblastoma. *Neuro-oncology*, 14(9):1153–1162, 2012.

[87] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[88] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273, 2003.

[89] Bin Yuan, Long Cheng, Huai-Chin Chiang, Xiaojie Xu, Yongjian Han, Hang Su, Lingxue Wang, Bo Zhang, Jing Lin, Xiaobing Li, et al. A phosphotyrosine switch determines the antitumor activity of erβ. *The Journal of clinical investigation*, 124(8):3378, 2014.

[90] Vasiliki Theodorou, Rory Stark, Suraj Menon, and Jason S Carroll. Gata3 acts upstream of foxa1 in mediating esr1 binding by shaping enhancer accessibility. *Genome research*, 23(1):12–22, 2013.

[91] Mark P Hamilton, Kimal Rajapakshe, Sean M Hartig, Boris Reva, Michael D McLellan, Cyriac Kandoth, Li Ding, Travis I Zack, Preethi H Gunaratne, David A Wheeler, et al. Identification of a pan-cancer oncogenic microrna superfamily anchored by a central core seed motif. *Nature communications*, 4, 2013.

[92] Yiming Zuo, Guoqiang Yu, Mahlet G Tadesse, and Habtom W Ressom. Biological network inference using low order partial correlation. *Methods*, 69(3):266–273, 2014.

[93] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

[94] Daniela M Witten and Robert Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.

Appendix

ADMM algorithm

In the case of the fused graphical lasso problem with elastic net the optimization problem 6.2 can be written as:

$$\begin{aligned} &\underset{\{\Theta\},\{Z\}}{\text{minimize}}\ f\{\Theta\}) + g(\lambda, \{Z\}) \\ &\text{subject to } \Theta^k = Z^k,\ k = 1, \ldots, K \end{aligned}$$

where S^k is the empirical correlation matrix for cancer type k, and

$$f(\{\Theta\}) = \sum_{k=1}^{K} n_k[\ln(\det(\Theta^k)) - \text{tr}(S^k\Theta^k)]$$

$$g(\lambda, \{Z\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} (\alpha |Z_{ij}^k| + (1-\alpha)(Z_{ij}^k)^2) + \lambda_2 \sum_{k<k'} \sum_{i \neq j} |Z_{ij}^k - Z_{ij}^{k'}|$$

The augmented Lagrangian, corresponding to Equation 6.3, becomes:

$$L(\{\Theta\}, \{Z\}, \{U\}) = f(\{\Theta\}) + g(\lambda, \{Z\}) + \frac{\rho}{2} \sum_{k=1}^{K} ||\Theta^k - Z^k + U^k||_F^2$$

The three steps of the ADMM algorithm becomes:

STEP 1: $\Theta_{(m)}^k \leftarrow \underset{\{\Theta\}}{\text{argmin}}\ \{L(\{\Theta_{(m-1)}\}, \{Z_{(m-1)}\}, \{U_{(m-1)}\}))$

It has been shown that this step can be solved with the help of an eigendecomposition[94], see line 10-11 in algorithm below.

STEP 2: $Z_{(m)}^k \leftarrow \underset{\{Z\}}{\text{argmin}}\ \{L(\{\Theta_{(m)}\}, \{Z_{(m-1)}\}, \{U_{(m-1)}\}))$

This step is separable for each element (i, j). A faster implementation using vectorization is sketched in the next section.

STEP 3: $U_{(m)}^k \leftarrow U_{(m-1)}^k + \Theta_{(m)}^k - Z_{(m)}^k$

The algorithm is outlined in detail next.

Variable initialization:

(Z^k, U^k becomes matrices of the same size as S^k, I is the identity matrix)

$l \leftarrow 0$

for $k = 1, \ldots, K$, where $K = \#$ cancer classes **do**

  $Z^k_{(0)} \leftarrow (S^k + \epsilon * I)^{-1}$

  $U^k_{(0)} \leftarrow 0$

end for

Select a scalar $\rho > 0$ (e.g. $\rho = \bar{n}$ or $2\bar{n}$)

Algorithm:

while not converged **do**

  $m \leftarrow m + 1$

  for $k = 1, \ldots, K$ **do**

  .....

  STEP 1:

  Let $VDV^T =$ eigendecomposition of

  $$[S^k - \rho/n_k(Z^k_{(m-1)} + U^k_{(m-1)})]$$

  $\Theta^k_{(m)} \leftarrow V\tilde{D}V^T$, where $\tilde{D}$ is a diagonal matrix with element $jj =$

  $$\frac{n_k}{2\rho}(-D_{jj} + \sqrt{D^2_{jj} + 4\rho/n_k})$$

  .....

  STEP 2:

  $A^k \leftarrow \Theta^k_{(m)} + U^k_{(m-1)}$ ▷ Introduce helper matrix A.

  for all i,j **do**

    Assume $A^1_{ij} \leq A^2_{ij} \leq ... \leq A^k_{ij}$

    $\tilde{Z}^k_{ij} \leftarrow A^k_{ij} - \lambda_2(2k - (K+1))$

    if $\tilde{Z}^{k_0}_{ij} > \tilde{Z}^{k_0+1}_{ij}$ **then**

      $\tilde{Z}^{k_0}_{ij} = \tilde{Z}^{k_0+1}_{ij} \leftarrow (\tilde{Z}^{k_0}_{ij} + \tilde{Z}^{k_0+1}_{ij})/2$

    end if

    $Z^k_{(m)ij} \leftarrow \alpha \ \text{sign}(\tilde{Z}^k_{ij}) \max(|\tilde{Z}^k_{ij}| - \lambda_{1,ij})/(1 + (1-\alpha)\lambda_{1,ij})$

  end for

  .....

  STEP 3:

  $U^k_{(m)} \leftarrow U^k_{(m-1)} + \Theta^k_{(m)} - Z^k_{(m)}$

  .....

  end for

end while

Vectorization

Lines 13-20 in the ADMM algorithm above include elementwise operations. Instead of looping through all $p \times p$ elements for the K matrices we used vectorization for massive speed up in Matlab, algorithm is sketched below.

1: **Let $\boldsymbol{B}$** (size $p(p-1)/2 \times K$) **be the rearranged version of $\boldsymbol{A}$** (size $p \times p$), where column k of B corresponds to the lower triangular part of $A^{(k)}$ (A is symmetrical, i.e. only half is needed for the calculations):

$$B \leftarrow \begin{vmatrix} A^1_{11} & A^2_{11} & \dots & A^K_{11} \\ A^1_{21} & A^2_{21} & \dots & A^K_{21} \\ \vdots & \vdots & \ddots & \vdots \\ A^1_{p(p-1)} & A^2_{p(p-1)} & \dots & A^K_{p(p-1)} \end{vmatrix}$$

2: $B \leftarrow$ sort each row of B. Save the order of the sorting.

3: $$H \leftarrow \begin{vmatrix} 1 & 2 & \dots & K \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \dots & K \end{vmatrix}$$ ▷ H is size $p(p-1)/2 \times K$

4: $B \leftarrow B - \lambda_2(2H - (K+1))$

Initialize:

5: $b = \textbf{FALSE}$ ▷ boolean matrix size $p(p-1)/2 \times K$

6: $m = \mathbf{0}$ ▷ vector size $p(p-1)/2 \times 1$

7: $l = \mathbf{1}$ ▷ vector size $p(p-1)/2 \times 1$

8: **for** $iter = 1 \rightarrow K$ **do**
9: **for** $k = 1 \rightarrow K - 1$ **do**
10: $f \leftarrow B_{.k} > B_{.k+1}$ ▷ boolean vector
11: $b_{.k} = b_{.k+1} \leftarrow f$
12: $l_{f==1} \leftarrow l_{f==1} + 1$ ▷ Counter of number of fused classes
13: $l_{f==0} \leftarrow 1$

14: **Calculate the new values for the** $fused$ **elements:**
$m_{f==1} \leftarrow [b_{f==1,k}(l_{f==1} - 1) + b_{f==1,k+1})]/l_{f==1,k}$

15: $m \leftarrow \begin{vmatrix} m_1 & \dots & m_1 \\ \vdots & \ddots & \vdots \\ m_{p(p-1)/2} & \dots & m_{p(p-1)/2} \end{vmatrix}$ ▷ matrix size $p(p-1)/2 \times K$
16: $B_{b==1} \leftarrow m_{b==1}$
17: **end for**
18: **end for**

19: $B \leftarrow$ reorder B with the help of the sorting stored in step 2.
20: **for all** i, j, k **do**
21: $Z_{ij}^{(k)} \leftarrow B_{corresponding\ row\ element,k}$
22: **end for**