UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Medicine 1426

Large scale integration and interactive exploration of cancer data – with applications to glioblastoma

PATRIK JOHANSSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2018

ISSN 1651-6206
ISBN 978-91-513-0231-7
urn:nbn:se:uu:diva-340843

Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, Dag Hammarskjölds väg 20, Uppsala, Friday, 23 March 2018 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Gary Bader (University of Toronto, Canada).

Abstract
Johansson, P. 2018. Large scale integration and interactive exploration of cancer data – with applications to glioblastoma. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 1426. 58 pp. Uppsala: Acta Universitatis Upsaliensis.
ISBN 978-91-513-0231-7.

Glioblastoma is the most common malignant brain tumor, with a median survival of approximately 15 months. The standard of care treatment consists of surgical resection followed by radiotherapy and chemotherapy, where chemotherapy only prolongs survival by approximately 3 months. There is therefore an urgent need for new approaches to better understand the molecular vulnerabilities of glioblastoma. To this end, we have conducted four interdisciplinary studies.

In study 1 we develop a method for efficiently constructing and exploring large integrative network models that include multiple cohorts and multiple types of molecular data. We apply this method to 8 cancers from The Cancer Genome Atlas (TCGA) and make the integrative network available for exploration and visualization through a custom web interface.

In study 2 we establish a biobank of 48 patient derived glioblastoma cell cultures called the Human Glioma Cell Culture (HGCC) resource. We show that the HGCC cell cultures represent all transcriptional subtypes, carry genomic aberrations typical of glioblastoma, and initiate tumors *in vivo*. The HGCC is an open resource for translational glioblastoma research, made available through *hgcc.se*.

In study 3 we extend the analysis of HGCC cell cultures both in terms of number (to over 100) and in terms of data types (adding mutation, methylation and drug response data). Large-scale drug profiling starting from over 1500 compounds identified two distinct groups of cell cultures defined by vulnerability to proteasome inhibition, p53/p21 activity, stemness and protein turnover. By applying machine learning methods to the combined drug profiling and matched genomics data we construct a first network of predictive biomarkers.

In study 4 we use the methods developed in study 1 applied to the data generated in studies 2 and 3 to construct an integrative network model of HGCC and glioblastoma data from TCGA. We present an interactive method for exploring this network based on searching for network patterns representing specific hypotheses defined by the user.

In conclusion, this thesis combines the development of integrative models with applications to novel data relevant for translational glioblastoma research. This work highlights several potentially therapeutically relevant aspects, and paves a path towards more comprehensive and informative models of glioblastoma.

Keywords: Glioblastoma, data integration, network modeling, interactive exploration, precision medicine

Patrik Johansson, Department of Immunology, Genetics and Pathology, Neuro-Oncology, Rudbecklaboratoriet, Uppsala University, SE-751 85 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-0231-7
urn:nbn:se:uu:diva-340843 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-340843)

Sub specie aeternitatis

List of papers

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

I Kling T*, **Johansson P***, Sanchez J, Marinescu VD, Jörnsten R, Nelander S.
Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content.
Nucleic Acids Research, 43(15):e98, 2015.

II Xie Y*, Bergström T*, Jiang Y*, **Johansson P***, Marinescu VD, Lindberg N, Segerman A, Wicher G, Niklasson M, Baskaran S, Sreedharan S, Everlien I, Kastemar M, Hermansson A, Elfineh L, Libard S, Holland EC, Hesselager G, Alafuzoff I, Westermark B, Nelander S, Forsberg-Nilsson K, Uhrbom L.
The Human Glioblastoma Cell Culture Resource: Validated Cell Models Representing All Molecular Subtypes.
EBioMedicine, 2(10):1351-63, 2015.

III **Johansson P**, Schmidt L, Baskaran S, Kundu S, Gallant CJ, Kling T, Awe O, Elfineh L, Holmberg Olausson K, Almstedt E, Häggblad M, Martens U, Lundgren B, Lönnstedt I, Frigault MM, Hurt E, Jörnsten R, Krona C, Nelander S.
Decoding glioblastoma drug responses using an open access library of patient derived cell models.
Manuscript.

IV **Johansson P**, Rosén E, Weishaupt H, Jörnsten R, Nelander S.
Exploring large scale integrative networks of glioblastoma using hypothesis driven pattern search.
Manuscript.

* indicates equal contribution.
Reprints were made with permission from the publishers.

Related work by the author

1 Schmidt L, Baskaran S, **Johansson P**, Padhan N, Matuszewski D, Green LC, Elfineh L, Wee S, Häggblad M, Martens U, Westermark B, Forsberg-Nilsson K, Uhrbom L, Claesson-Welsh L, Andäng M, Sintorn IM, Lundgren B, Lönnstedt I, Krona C, Nelander S.
Case-specific potentiation of glioblastoma drugs by pterostilbene.
Oncotarget, 7(45):73200-73215, 2016.

2 Kling T, Ferrarese R, O hAilin D, **Johansson P**, Heiland DH, Dai F, Vasilikos I, Weyerbrock A, Jörnsten R, Carro MS, Nelander S.
Integrative Modeling Reveals Annexin A2-mediated Epigenetic Control of Mesenchymal Glioblastoma.
EBioMedicine, 12:72-85, 2016.

3 Weishaupt H, **Johansson P**, Engström C, Nelander S, Silvestrov S, Swartling FJ.
Graph centrality based prediction of cancer genes.
Engineering Mathematics II: Algebraic, Stochastic and Analysis Structures for Networks, Data Classification and Optimization / [ed] Sergei Silvestrov; Milica Rancic, pp. 275-311, Springer, Cham, 2016.

4 Weishaupt H, **Johansson P**, Engström C, Nelander S, Silvestrov S, Swartling FJ.
Loss of conservation of graph centralities in reverse-engineered transcriptional regulatory networks.
Methodology and Computing in Applied Probability, 1-17, 2017.

Contents

Abbreviations

ADMM	alternating direction method of multipliers
APC	adenomatous polyposis coli
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
ASPM	abnormal spindle microtubule assembly
ATRX	ATRX, chromatin remodeler
BBB	blood brain barrier
CCLE	Cancer Cell Line Encyclopedia
CDK4	cyclin dependent kinase 4
CDKN2A	cyclin dependent kinase inhibitor 2A
CDKN2B	cyclin dependent kinase inhibitor 2B
CGP	Cancer Genome Project
CHI3L1	chitinase 3 like 1
ChIP	chromatin immunoprecipitation
CLR	context likelihood of relatedness
CNA	copy number alteration
COSMIC	Catalogue Of Somatic Mutations In Cancer
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CSC	cancer stem cell
CTRP	Cancer Theraputics Response Portal
DAG	direct acyclical graph
DNA	deoxyribonucleic acid
EGFR	epidermal growth factor receptor
G-CIMP	glioma-CpG island methylator phenotype
GABRA1	gamma-aminobutyric acid type A receptor alpha1 subunit
GBM	glioblastoma
GDSC	Genomics of Drug Sensitivity in Cancer
GENIE3	GEne Network Inference with Ensemble of trees
GWAS	genome-wide association study
HDAC	histone deacetylase
HDAC1	histone deacetylase 1
HGCC	Human Glioma Cell Culture
HTS	high-throughput screening
ICGC	International Cancer Genome Consortium
IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial
JIVE	Joint and Individual Variation Explained
LINCS	Library of Integrated Network-based Cellular Signatures
LOH	loss of heterozygosity
MET	MET proto-oncogene, receptor tyrosine kinase
MGMT	O-6-methylguanine-DNA methyltransferase
MI	mutual information
miRNA	micro RNA
mRNA	messenger RNA
MYC	v-myc avian myelocytomatosis viral oncogene homolog

NCI	National Cancer Institute
NEFL	neurofilament, light polypeptide
NES	nestin
NF1	neurofibromin 1
NIH	National Institutes of Health
PDGFRA	platelet derived growth factor receptor alpha
PDX	patient derived xenograft
PI	proteasome inhibitor
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PIK3R1	phosphoinositide-3-kinase regulatory subunit 1
PTEN	phosphatase and tensin homolog
RB1	RB transcriptional corepressor 1
RNA	ribonucleic acid
RTK	Receptor tyrosine kinase
SICS	sparse inverse covariance selection
SLC12A5	solute carrier family 12 member 5
SNP	Single nucleotide polymorphism
SOX2	SRY-box 2
SSC	sample size correction
STITCH	search tool for interactions of chemicals
SYT1	synaptotagmin 1
TCGA	The Cancer Genome Atlas
TERT	telomerase reverse transcriptase
TIC	tumor initiation capacity
TIGRESS	Trustful Inference of Gene REgulation with Stability Selection
TP53	tumor protein p53
TSS	transcription start site
UCSC	University of California, Santa Cruz
WGCNA	Weighted Correlation Network Analysis
WHO	World Health Organization

Introduction

Cancer

Cancer is a collective term for diseases characterized by uncontrolled cell growth invading or spreading to other parts of the body. Meaning 'crab' in Latin, from the Greek 'karkinos', cancer is said to be named after the swollen veins around certain tumors resembling the legs of a crab [1]. Cancer incidence and mortality is rising worldwide, mainly due to a growing and aging population and risk factors associated with an increasingly westernized way of living [2]. In Sweden, the incidence rate of cancer is increasing [3, 4] even when adjusting for an aging population. The mortality rate of cancer in Sweden has been slowly but steadily decreasing and is now at the lowest rate ever measured [3, 4]. Tumors located in the brain (including cranial nerves and meninges of the brain) account for around 1300 cases per year in Sweden [4]. The incidence of and mortality from these tumors has remained essentially constant when adjusting for population age (Figure 1).

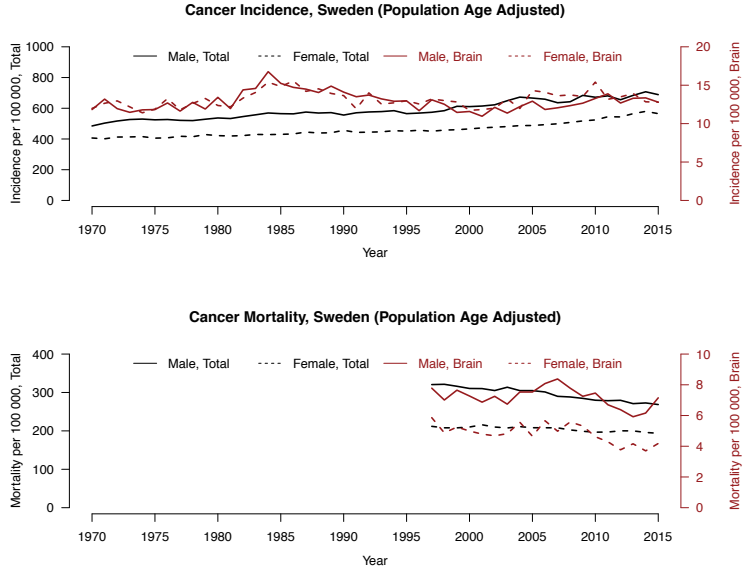

Figure 1. Cancer incidence in Sweden, data from Socialstyrelsen [4]. Population age standardized against population at year 2000.

Cancer cells are proposed to obtain their ability to multiply beyond control by acquiring a particular set of phenotypes, often referred to as the ”hallmarks of cancer” [5, 6] (Figure 2). Exactly how cells acquire these hallmarks, and how they can be counteracted is the main question at stake for developing efficacious cancer treatments. How cells acquire these hallmarks varies depending on many factors, including the tissue and normal cell being affected, but in general it consists of the disruption of the regular function of a gene, or a set of genes that contribute to maintaining normal cell function. During the last decades, attention has largely been focused on cancer-causing mutations, but interest has over time shifted to include many types of genetic, epigenetic and transcriptional alterations. Key alterations in cancer are centered around a complex network of interconnected pathways that regulate crucial cellular phenotypes. Perhaps most famous are alterations of members across intracellular pathways that control cell viability, proliferation (RTK, Ras, Myc), motility (Apc, β-catenin), cytostasis (p16, p21, pRb) and apoptotic signalling (p53) [6]. Importantly, cancer cells are also influenced by their interaction with surrounding cells (the microenvironment), creating a complex system consisting of cancer cells, cancer progenitor cells, immune cells, and surrounding normal cells.

Charting the molecular basis of cancer has been a long ongoing process. Pioneering projects, such as The Cancer Genome Atlas (TCGA, `cancergenome.nih.gov`) and the TCGA pan-cancer analysis project [7] together with other efforts (e.g. [8, 9, 10]) have unearthed the molecular basis of over 34 types of cancer. These efforts have identified mutations related to over 500 genes to be implicated in cancer, serving as a foundation for advances across the research community [9].

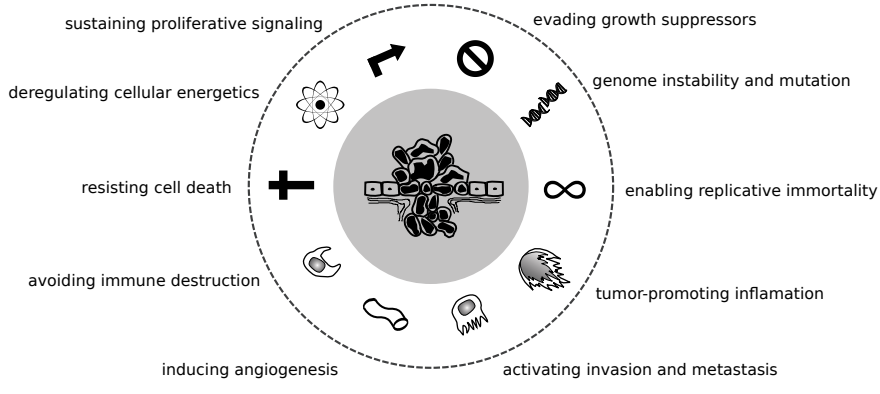

Figure 2. The *hallmarks of cancer* and enabling characteristics, proposed as an organizing principle by Hannahan and Weinberg for the main capabilities acquired during the development of human tumors. Adapted from [11].

Glioblastoma

For a long time, the WHO classification of brain tumors into grades and types was based mainly on histopathological features [12], but has recently been updated to also include molecular information [13, 14]. Astrocytes, oligodendrocytes and ependymal cells together form the supporting structure around the neuronal cells of the brain and are called *glial cells* from the Greek word meaning 'glue'. Gliomas therefore, are those tumors whose cells resemble glial cells. The annual incidence rate of gliomas in the United States is 6.3 in 100,000 individuals [15]. Based on the perceived similarity between the cells of a tumor and normal cells of the brain, gliomas are categorized as astrocytic, oligodendroglial, ependymal, and neuronal, among others [14].

The grade of diffuse astrocytic and oligodendroglial tumors (II-IV) is decided based on criteria assessing malignancy and is highly correlated with patient outcome. In short, according to WHO

> ... "tumors with cytological atypia alone are considered grade II, those that also show anaplasia and mitotic activity are considered WHO grade III, and tumors that additionally show microvascular proliferation and/or necrosis are grade IV" [14]

Glioblastoma (*GBM*, previously glioblastoma multiforme) is a grade IV glioma with predominantly astrocytic differentiation [14]. GBM is the most common primary brain tumor with an annual incidence rate of 3.2 in 100,000 in the United States and shows extremely poor patient outcome with a 5 year survival of 5.5% [15]. Standard treatment includes surgical resection, radiotherapy and chemotherapy with the drug temozolomide. The median survival after diagnosis is a mere 14.6 months with radiation and temozolomide and 12.1 months with only radiotherapy [16]. Glioblastomas are 1.6 times more common in men and the median age at diagnosis is 64 years [15]. Environmental factors contributing to GBM are poorly understood, with the only established link being exposure to ionizing radiation [17]. A link between cell phone usage and glioma has been widely speculated, and while conclusive evidence has not been found, indications of a connection have been reported [18, 19].

Glioblastomas are categorized into *primary* and *secondary* depending on whether they arise de novo (primary) or arise from lower grade tumors (secondary). GBM is also divided into IDH-wildtype, IDH-mutant and NOS (not otherwise specified) based on the IDH1 and IDH2 mutation status. IDH-wildtype GBM accounts for about 90% of all GBM and consists mainly of primary GBM [20]. IDH mutation is considered a strong marker for secondary glioblastoma [20]. Promoter methylation of *MGMT*, present in 45% of glioblastoma patients, imparts an improved response to temozolomide but is also an independent marker for favorable prognosis [21]. Importantly, non-molecular factors, such as the extent of surgical resection, play a large part in patient survival [22, 23]. Patient survival is highly correlated with the age

at diagnosis with some patients diagnosed under the age of 20 showing prolonged survival [15].

Glioblastoma genomics

A number of important publications have contributed to the understanding of the genomic properties of glioblastoma [24, 25, 26, 27, 28, 29, 30]. The emerging picture is one implicating a number of important genes centered around the *TP53*, *RB1* and RTK/Ras/PI(3)K pathways (Figure 3B). *TP53* is a well known tumor suppressor gene recurrently mutated in many cancers [31]. Non-mutated *TP53* plays an important role in maintaining normal cell cycle regulation, proliferation control, induction of apoptosis and control of DNA repair pathways [32]. Regulation of *RB1* controls cell cycle progression and in GBM, *RB1* can either be altered directly or impacted by upstream alterations such as amplification of *CDK4* or deletion of *CDKN2A* and *CDKN2B* [26]. Receptor tyrosine kinases (RTKs) are cell surface receptors that transmit extracellular signals to activate signalling pathways within the cell, controlling crucial functions such as sustaining proliferative signalling [5, 6]. In GBM, common alterations to RTKs include *EGFR* and *PDGFRA*, together with alterations of downstream factors such as *PTEN*, *NF1* and *PI(3)K* [26, 30]. Gain of chromosome 7 and loss of chromosome 10 are characteristics of glioblastoma, containing *EGFR* and *PTEN* respectively. A number of important risk-associated SNPs for glioma have emerged from GWA studies, connected to *EGFR*, *TP53*, *CDKN2B* and regions linked to *IDH*-mutations [33, 34].

Glioblastoma heterogeneity and stratification

Glioblastoma itself is a heterogeneous disease. GBM patients are proposed to have differing prognosis depending on the genotype, epigenome and gene expression properties of the tumor. The spread across these features between tumors has given rise to subclassifications of glioblastoma based on expression profiles [24, 28, 30], methylation profiles [29, 30, 36] and mutations [13, 30, 37].

The most commonly used subclassification of GBM based on gene expression data is that of splitting into four subtypes; Proneural, Neural, Classical and Mesenchymal [28]. Tumors of the proneural subtype typically have high expression of *PDGFRA* and were originally proposed to be associated with longer patient survival [28]. However, subsequent stratification of the proneural patients based on methylation data revealed that the observed survival benefit was imparted by a subset of IDH1 mutated and hypermethylated samples termed the G-CIMP group [29, 30]. Interestingly, the G-CIMP group contains a higher proportion of young patients. These findings have since been included in the updated WHO classification of glioblastoma [13]. Tumors of the clas-

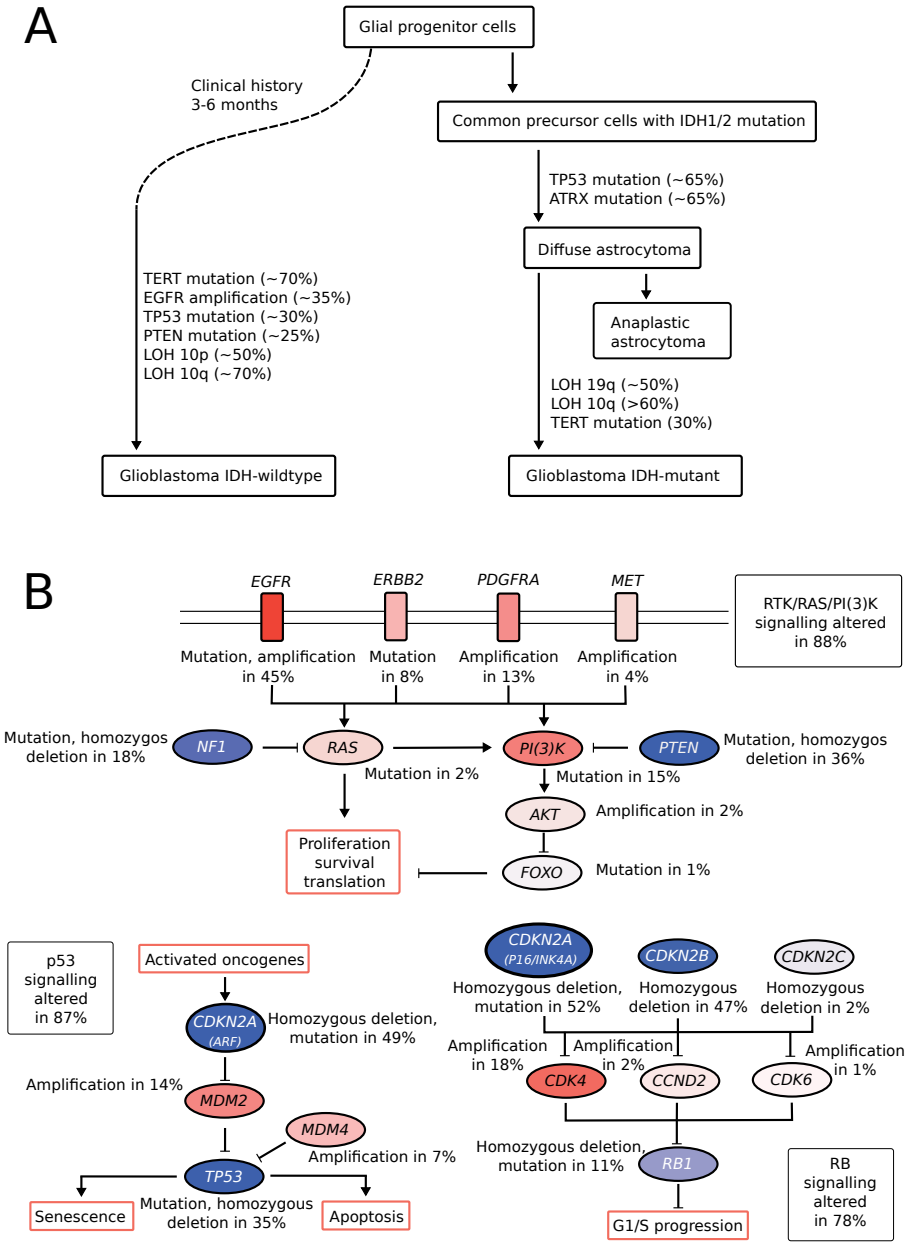

Figure 3. Main pathways and organizational structures of glioblastoma.
(A) Genetic pathways to IDH-wildtype and IDH-mutant Glioblastoma. Adapted from [35] as in [14]. (B) Main altered pathways in GBM. Reprinted with permission [26].

sical subtype typically carry amplifications of *EGFR* and deletions or loss of chromosome 10. The neural tumors are enriched for neuronal markers such as *NEFL*, *GABRA1*, *SYT1*, and *SLC12A5* while the mesenchymal tumors are enriched for mesenchymal markers such as *CHI3L1* and *MET*. Others have proposed stratifying glioblastoma and lower grade gliomas by the same criteria, based on the observation that molecularly, IDH mutated glioblastoma is more similar to lower grade glioma than to IDH-wildtype glioblastoma [37]. Over time, a suspicion has grown that the neural subtype is dominated by tumor-extrinsic factors such as stromal contamination and that glioblastoma cell-intrinsic stratification only includes the Proneural, Mesenchymal and Classical subtypes [38].

The stratification of glioblastoma patients into subtypes based on molecular properties highlights the heterogeneity of this tumor, which is argued to have important implications for therapeutic options and the necessity for personalized treatment. Heterogeneity within tumors has been shown to span all gene expression subtypes and to be present at the single-cell level [39, 40]. The evolution of a glioblastoma tumor, from the first glioblastoma cell to the tumor mass that is surgically resected, and in turn to the inevitable recurrence, is a complex process. The chronology of how recurrent aberrations appear in IDH-wildtype cases highlights a progression from a proneural-like precursor, with gain of chromosome 7 and loss of chromosome 10 being early events [41]. In secondary glioblastoma, early events include those typical of lower grade glioma, such as mutation of *IDH1*, *TP53* and *ATRX* (Figure 3A) [35]. Analysis of genetic evolution in GBM highlights mutation of *IDH1* and *TP53* as early events that commonly occur in both tumor and recurrence [42]. However, key GBM driver alterations in the recurrence frequently diverges from the primary tumor, with distant recurrences being less genetically similar than local recurrences [43, 44]. Wang et al. show that the gene expression subtype is only retained between initial tumor and recurrence in 55% of cases [38]. The intra-tumor genomic heterogeneity of GBM has been shown to be important for functional heterogeneity, also at the single cell level [44, 45, 46].

Tumor initiating cells and *in vitro* models of GBM

To explore functional and pharmacological properties of GBM an appropriate disease model is needed. Disease models can mainly be categorized as *in vitro*, *in vivo* and *ex vivo*. The choice of model system for testing a given hypothesis is a balance between the clinical relevance, practicality and cost. In this balance, *in vitro* models using patient-derived cell cultures have emerged as a popular model system in the GBM field [47, 48, 49]. To follow up on results obtained *in vitro*, *in vivo* models provide an added level of clinical relevance, utilizing for instance mice, rats or zebrafish.

Patient-derived GBM cells will typically not grow *in vitro* if not encouraged by supplements added to the cell medium. For a long time, cultured GBM cells were supplemented with serum, a non-defined mixture of growth factors. Decades ago (1966) the widely used GBM cell line U87MG was established under such conditions [50]. U87MG is still commonly discussed with over 450 mentions in PubMed for 2017 (Over 3500 mentions in total [1]). Since U87MG has been passaged many times in serum it is unclear whether this model still represents the original tumor [51]. Other frequently used serum grown cells include U251MG, LN-18, T98G and A172 [52]. It has been shown that cells grown in serum drift away from the original tumor in gene expression space, thus making them less suited for phenotypic studies [53]. To counteract this problem, cells are nowadays often grown in well defined serum-free medium [47, 48, 49]. This medium is designed to have less effect on the cells, while still providing the growth factors needed for prolonged passaging, necessary when producing material for molecular profiling and follow-up studies.

It has been proposed that only a subset of cells, so called *cancer stem cells* (CSCs), have the capacity to self-renew and to differentiate into all kinds of tumor cells, thus being the driving force behind tumor growth [54]. Cancer stem cells are proposed to be the culprit in tumor recurrence, since if radiation or chemotherapy has failed to eradicate these cells, capacity to regenerate a tumor is retained [55, 56]. In GBM, a subpopulation of cells resembling CSCs has been identified [57, 58] but molecular markers for this population have proven elusive [59, 60]. The putative glioblastoma stem cell (GSC) population has been proposed to sustain long-term tumor growth after treatment [61] and inducing GSC differentiation has been suggested as a strategy to limit tumor growth [62]. Closely related to the GSC hypothesis is the proposed *cell of origin* of GBM. Although historically referred to as Grade IV Astrocytoma, GBM is thought to arise from both astrocyte, oligodendrocyte and neuronal precursor cells [63, 64]. An important property of patient-derived GBM cells is their ability to maintain stemness characteristics, such as the ability to form neurospheres *in vitro* and to form tumors *in vivo*.

To investigate GBM in an *in vivo* setting, both rats and mice have been popular alternatives since the 1940s [65, 66]. Tumors can be induced by injecting GBM cells into immunodeficient mice, or by genetically engineering mice to develop tumors intrinsically. When injecting patient-derived cells, the cells can either be established as cell cultures first, or injected directly as patient derived xenografts (PDXs) [47, 67, 68]. Genetically engineered mouse models rely on altering signalling pathways that are known to induce glioblastoma. Models rely on different alterations, giving rise to tumors induced with different backgrounds, such as Ras and Akt [69], *PDGFB* [70, 71], *EGFR* [72], and *TP53/NF1* [73, 74].

[1] PubMed Search `http://www.ncbi.nlm.nih.gov/pubmed/?term=U87+OR+U-87` on January 11 2018

Data driven analysis of cancer

Technological development has unleashed the generation of large amounts of data from cancer samples, be it patient surgical samples or cultured cancer cells. These growing masses of data allow for a *data-driven* approach to understanding how cancer develops. In the data driven-world the inner workings of a cell is often simplified compared to the biological reality. The level of simplification is strongly dependent on the technological developments that at any given point restricts us economically to measure some, but not other molecular properties with high throughput. The data-driven world thus relies on those entities for which data can be generated and omits those for which it can not, hoping that these omissions do not skew our interpretations.

The methods available for high-throughput profiling can probe various stages in the process of converting genetic information into the molecules that ultimately determine cell function (Figure 4). Starting at the level of DNA, there are *genetic* changes, such as mutations or alterations of the copy number of sections of DNA. On top of the genetic state, there are *epigenetic* changes that impact inheritable traits such as DNA methylation and histone modifications. The genetic and epigenetic state together influence the makeup and quantity of the RNA that is transcribed from a gene together with other forces regulating gene transcription, such as the presence of transcription factors. Post-transcriptional modifications and regulation of the transcribed mRNA, such as binding of miRNAs, influences how much of the mRNA that is translated. Finally, we can measure the levels of various proteins in a sample. Since protein measurements are relatively expensive, the corresponding mRNA level is often used as a proxy for the protein level. In the following section, a number of commonly used data types are covered.

Common types of cancer data

Copy number alterations

Copy number alterations (CNA) are changes in the number of copies of a DNA segment. CNAs are structural changes that can occur for relatively short spans of bases, but commonly occurs for segments containing entire genes, or even many genes. In this way, genes can be deleted or duplicated, potentially leading to a decrease or increase in gene expression.

Copy number aberrations can be measured by probe based array techniques, such as Affymetrix Cytoscan HD. On such arrays, probes with an affinity for a specific DNA sequence are distributed across the genome and a fluorescent signal is used to quantify the abundance of each. A smoothing algorithm is then used to convert the probe values into copy number values for longer segments along the genome. Single nucleotide polymorphism (SNP) arrays can also be used in a similar manner.

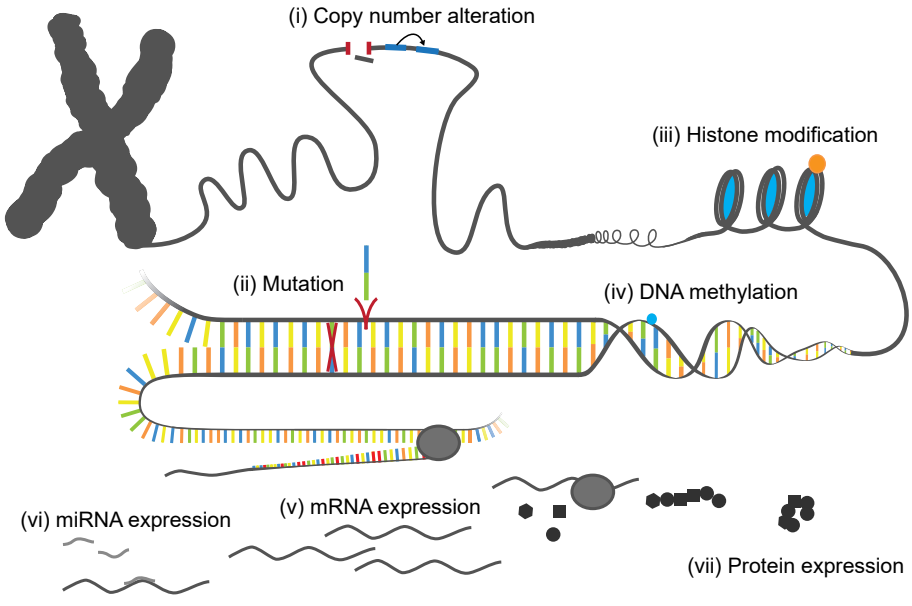

Figure 4. Conceptual overview of key data types in large scale analysis of cancer. (i) Copy number alterations, (ii) mutations, (iii) histone modifications, (iv) DNA methylation, (v) mRNA expression, (vi) miRNA expression and (vii) protein expression.

Mutations

Mutations are base-pair changes in the DNA sequence. The effect of a mutation is completely dependent on where the mutation occurs and what the change is. Mutations that have no effect on any phenotype are called *silent*. Mutations consist of the substitution, insertion or deletion of a single base (called *point-mutations*) or a set of bases. Some mutations in protein coding genes do not impact the resulting protein at all, for example if a point mutation results in a three-base-pair codon that codes for the same amino acid. Such mutations are also called *synonymous*. *Non-synonymous* mutations on the other hand are those that impact the gene product in some way. *Missense* point mutations alter a base so that it results in a different amino acid, impacting the resulting protein. *Nonsense* mutations are those that have a larger effect on the transcription of a gene, for example by introducing a new start or stop position, completely changing the resulting RNA. Insertions or deletions can cause large changes of gene transcription, *frameshift* mutations for example, are those that offset transcription so that the three-base-pair codons are grouped in a different configuration. Mutations outside genes can also have large effects, for example by impeding the binding of regulatory elements.

High-throughput methods for mutation detection are based on DNA sequencing. Sequencing of DNA can occur at varying resolution and com-

prehensiveness. Whole genome sequencing targets the entire genome but is therefore relatively expensive. Alternatively, a targeted sequencing, limited to regions of interest (e.g. specific genes or all exons) can be used. To determine the existence of mutations, a reference sequence is needed. The reference sequence can either be based on a population sample (after removing commonly occurring variation) or it can be based on a matched normal sequence.

Histone modifications

Histones are the core around which DNA is wrapped to form a *nucleosome*, the basic structure for coiling DNA into chromatin (Figure 4). How compactly coiled the nucleosomes are influences the level of gene transcription in that region. There are many modifications of the histone core units, for example acetylation, methylation and phosphorylation. These modifications serve as a dynamic system for controlling chromatin structure and can be associated with both active and repressed transcription.

ChIP-seq is an increasingly popular method for measuring histone modifications with high throughput [75]. Chromatin Immunoprecipitation (ChIP) is the process of first fixating DNA bound proteins to the DNA, followed by fragmentation of the DNA, and then using antibodies to enrich for proteins of interest. By first performing ChIP focusing on histone modifying proteins, and then sequencing the resulting DNA fragments the positions of histone modifications can be inferred.

DNA methylation

DNA methylation is the addition of a methyl group (CH_3) to a cytosine in the DNA strand, usually where the cytosine is followed by a guanine (called a CpG site). Methylations are an epigenetic mechanism by which the expression of many genes is controlled. Regions upstream of the transcription start site (TSS) of a gene often contains CpG islands (regions that are enriched for CpG sites) that are normally unmethylated. Once methylation of a previously unmethylated CpG site occurs it often leads to decreased expression of the corresponding gene. Understanding how methylation impacts gene expression, and other functions, is still an area of active research. For a recent review, see [76]. Many effects of methylation are less understood and are very context dependent, for example when a methylation interferes with the binding of transcription factors to enhancers and other regulatory elements, or the interplay between methylation and other epigenetic properties such as histone modifications.

DNA methylation is most often measured on bisulfite-treated DNA, either by subsequent sequencing or by probe-based methods. The bisulfite-treatment converts unmethylated cytosine to uracil, but leaves methylated cytosine unchanged. Thus, two probes for each interrogated site are used, one quantifying the methylated, and one quantifying the unmethylated state. The methylation state of a site is then summarized as a β-value, defined as a ratio between

the methylated and the overall intensity [77]. Each platform will cover different regions with varying number of probes, often focusing on CpG islands and other regions with known regulatory importance. For instance, the popular Illumina Infinium MethylationEPIC platform covers over 850,000 sites focusing on CpG islands, regions surrounding CpG islands, sites known to be differentially methylated in tumor samples, enhancer regions, and more.

mRNA expression

mRNA expression is one of the most abundantly generated molecular data. This is in large part due to early development of cost effective probe-based techniques to simultaneously measure large numbers of mRNA levels. Since mRNA levels can be seen as a proxy for protein levels in a cell, they are often used to interpret the state of various cellular phenotypes or molecular functions. It has long been assumed that there is a positive correlation between the mRNA level and the corresponding protein level, a notion that has been significantly nuanced over time [78, 79, 80].

Hybridization based microarrays for mRNA quantification have been extensively used since the 90s [81], but are gradually being replaced by sequencing based methods [82]. Hybridization arrays are based on specially designed probes with an affinity for regions within certain genes. These are arranged on a plate and the corresponding mRNA attaches to them in order to be quantified by measuring a fluorescence intensity. RNA sequencing (RNAseq) is based on sequencing of cDNA, used to more accurately quantify the corresponding mRNA levels.

miRNA expression

Micro RNAs (miRNA) are small non-coding RNAs that target mRNA. miRNAs can silence or regulate gene expression by post-transcriptional binding or cleaving of mRNA. The quantification of miRNA can be achieved by the same methods as for mRNA, but with specifically designed probes or by sequencing analysis.

Protein expression

Proteins, being a key determinant of cellular function, can be seen as the gold standard measurement to understand cellular phenotypes. Reducing the cost of measuring proteins is a rapidly developing field, including antibody based techniques such as proximity ligation assays [83] and reverse phase protein arrays [84], as well as various methods based on mass spectrometry [85].

Clinical and phenotypic data

Measurements of molecular properties are not inherently interesting if they can not be connected to some clinically relevant phenotype. The most common type of clinical information considered in cancer genomics studies is readily available patient specific covariates such as age, sex and postoperative survival. By linking such information to genomic data sets it is, accordingly, possible to identify molecular markers of survival, or to compare the demography between different molecular subtypes. In this thesis, we also consider a different form of phenotypic data, collected using functional assays applied to cells that were grown from the patients' tumors. One key example of such data is drug sensitivity profiles, collected by high-throughput screening (HTS) technology, which can be combined with molecular data to suggest mechanisms behind the observed drug response. A second important example of patient specific phenotypic data obtained in a preclinical setting is measurements of tumor initiating capacity of patient derived cells transplanted into immunodeficient host animals. This information can provide important molecular correlates of tumor initiating or invasive capability of the tumor cells. *In vivo* phenotypes are expensive to measure at large scale but are an important follow-up in high-throughput studies.

Integrative analysis of cancer data

Databases of multidimensional cancer data

Tremendous effort has been put into the systematic collection of multidimensional omics data across all major cancer diagnoses. The most prominent effort is *The Cancer Genome Atlas* (TCGA, `cancergenome.nih.gov`) covering over 11000 patient cases across 34 cancers. This publicly accessible database includes gene mRNA expression, copy number alterations, DNA methylation, mutation, protein expression and clinical data. Similar efforts include the International Cancer Genome Consortium (ICGC) [86], the Cancer Genome Project (CGP) [87], the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [88] and MSK-IMPACT [10].

Drug responses of traditional cancer cell lines have been studied extensively. Publicly available data repositories of perturbation experiments include NCI-60 [89], the CCLE (Cancer Cell Line Encyclopedia) [90], the GDSC (Genomics of Drug Sensitivity in Cancer) [91, 92], the CTRP (Cancer Therapeutics Response Portal) [93, 94] and the NIH LINCS program (L1000) [95]. These data sources have all been used in a large number of publications to gain understanding about different cancers (For instance, NCI-60 is mentioned by over 700 entries in Pubmed [2]).

[2]Retrieved on January 11 2018.

One of the current challenges is to develop mathematical models and visualization techniques to leverage these huge sources of available data. Integrative analysis using multiple data types and multiple patient cohorts is a rapidly developing field, aiming to increase the power of an analysis by including more data at once. Often, traditional methods can be adapted to accommodate these types of analysis, for instance by extending unsupervised clustering to a *clusters of clusters* analysis to include multiple data types, or by simply comparing summary statistics across patient cohorts.

Integrative analysis across data types and cohorts

Integrating data across many different types of cancer, so called *pan-cancer* analysis is an emerging area of research [7]. The idea is that even though different types of cancer behave differently they rely on having the same *hallmarks* and therefore exploit the same cell vulnerabilities to some extent. The structure of how different cancers are alike and different can give insight into how each cancer gains a growth advantage over normal cells. For instance, Kandot et al. analyze exome sequencing data from 12 TCGA cancers in order to identify 127 significantly mutated genes, highlighting how they are centered around 20 core cellular processes [31]. Tamborero et al. also use TCGA data from 12 cancers and utilize an ensemble of methods to identify 291 cancer driver genes [96]. Zack et al. consider the copy number profiles of 4934 cases across 11 cancers, noting that many regions with identified copy number alterations did not contain any known cancer driver genes [97]. Ciriello et al. identify two main pan-cancer tumor classes, driven predominantly by either recurrent copy number alterations or recurrent mutations [98]. Hoadley et al. [99] perform an integrative analysis of 12 cancer types and five data types. Using unsupervised clustering they identify 11 major subtypes, some of which overlap well with a particular cancer, while others contained samples from different cancers. In fact, for some cancers the samples were split between more than one of the 11 major subtypes, indicating that the cancer type itself is not the main determinant of molecular features. Iorio et al. combine drug response data and molecular data for 990 cancer cell lines to predict drug responses and drug interactions across 19 tissue types [100]. The pan-cancer notion can also be applied to *in-silico* prescription of drugs, i.e. transferring knowledge about efficacious treatments in one cancer to others with a similar molecular basis [101]. Sharing data across cancers, when implemented properly, allows us to learn more about each cancer by increasing the power of our statistical models in detecting shared processes.

Integrative network models

Statistical models for multidimensional analyses come in many flavors. One extensively used class of models are so called *network models*. A network, or graph, in this context is a set of nodes (variables, e.g. gene expression for a gene) and a set of links between pairs of nodes. The meaning of a link connecting two nodes depends on the framework used to infer the network. Usually, a link between two nodes indicates an association between them (undirected) or an influence of one on the other (directed, e.g. the expression of a gene regulating the expression of another gene) (Figure 5A-B). Different frameworks regard association differently; in correlation networks (e.g. co-expression networks), association is measured by the correlation coefficient, in graphical models, association is measured by the conditional dependence between two random variables. In a network including multiple data types, different nodes represent different data types, such as the mRNA expression or CNA of a gene, or the methylation of a certain site (Figure 5C). In a network including multiple classes (e.g. patient cohorts, cancer subtypes), a link can have different values for each class (Figure 5D). Depending on the assumptions we make about variables and the structure of our network, we must use different tools for network inference.

Correlation based networks

The simplest version of a correlation based network method is to calculate all the pairwise correlations between variables. These can then be thresholded to produce a 0-1 adjacency matrix describing a network of pairwise associations. Many methods use the correlation matrix as input to derive a network conforming to certain properties. WGCNA, for example, transforms the correlation network so that it better conforms to proposed properties of regulatory networks, such as having a scale-free topology [102]. Other methods, such as sparse inverse covariance selection (SICS, used in Papers I and IV) uses the covariance matrix to estimate the pairwise partial-correlations, i.e. correlations while controlling for confounders. Applied to GBM, co-expression networks have been used to stratify patients [103] and WGCNA has been used to identify the gene *ASPM* as a potential molecular target [104].

Information-theoretical networks

Methods based on information theory replace the linear concept of correlation with non-linear concepts like mutual information. Mutual information (MI) is a measure of how much information is shared between two variables. Similar to the correlation based approach, a network can be obtained by thresholding all pairwise MI values. Methods such as ARACNE [105] and CLR [106], use pairwise mutual information together with pruning methods that retain only the most stringent associations. Applied to GBM, ARACNE has been used to explore the activation of a mesenchymal phenotype linked to tumor aggressiveness [107].

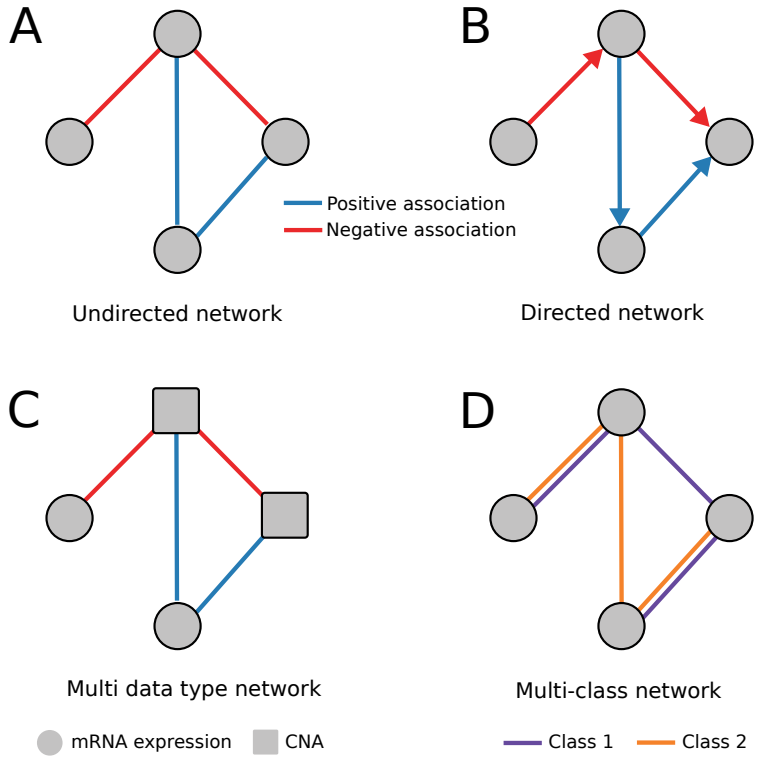

Figure 5. Overview of network model types.

Bayesian Networks / Graphical models

Bayesian networks describe probabilistic relationships between variables representing a directed acyclic graph (DAG). In such a graph, links have directions and the graph contains no cycles. I.e. it is not possible to start from one node, and to follow the directed edges to return to that node. This limits the kinds of regulatory relationships that these networks can represent, for example excluding the existence of feedback loops. However, the directed nature of the links in these graphs, and the accomodation of missing data still make them attractive for modeling regulatory networks. The inference of bayesian networks is relatively computationally demanding and is thus typically used when modelling relatively small systems [108].

Other methods

There are many other frameworks for representing networks. For instance, to model time-dependent behaviour, differential equations are often used. The state of a node is defined as a change over time, as a function of the state of other nodes. This creates a set of differential equations that are solved in order to describe the dynamics of the system. These can either be ordinary differ-

ential equations, partial differential equations, or even stochastic differential equations.

Boolean networks use a 0-1 status of nodes that are determined by a boolean function of the states of a subset of other nodes [109]. Such networks allow logical relationships between states to be modelled (e.g. if gene A is expressed *AND* gene B is mutated, gene C is not expressed), often with a time dimension.

In addition, any ad-hoc method can be used to define links between nodes. For instance, one can describe a network by connecting the variables that accurately predict another variable to each other. By applying a prediction method for each variable at a time, using a variable selection method to obtain a subset of other variables, a global network can be derived. This allows for a large degree of flexibility since there are many specialized prediction methods, suited for various interaction types. However, these methods are often computationally intensive and do not guarantee any theoretical properties for the network.

Sparse Inverse Covariance Selection

Sparse Inverse Covariance Selection (SICS) is used to derive approximations of partial correlations from covariance matrices. If we assume that our variables come from a multivariate gaussian (normal) distribution, and that the underlying network is undirected, a zero entry in the inverse of the covariance matrix of the gaussian distribution implies that the two corresponding variables are conditionally independent (no link). Thus, the problem boils down to efficiently estimating the inverse of the covariance matrix.

For typical high-throughput cancer omics data, the number of variables exceeds the number of samples. This prohibits the direct inversion of the covariance matrix, since it is a singular matrix in these cases. Alternative methods use a sparse approximation of the inverse covariance matrix, for example by solving one penalized regression problem for each variable, using the other variables as predictors [110]. Friedman et al. propose finding the partial correlations by optimizing a penalized likelihood problem [111]. In general terms, such a problem can be stated as:

$$\underset{\Theta}{\text{argmax:}}\ l(\Theta \mid S) - \text{penalty}(\Theta)$$

where Θ is the matrix describing the network and S is the sample covariance matrix. In simpler terms: given a covariance matrix from our data, we want to find an optimal network under the influence of certain penalty constraints. The first term of the equation controls that our network actually consists of partial correlations. The second term controls the sparsity of the network. The penalty function can take various forms, such as a lasso penalty [112] (then called *Glasso* [111]), or an elastic net penalty [113]. The lasso penalty applies to the sum of the absolute values of the elements of Θ, while the elastic net balances a lasso penalty with a quadratic penalty term:

$$\sum_{i \neq j} (\alpha |\theta_{ij}| + (1-\alpha)\theta_{ij}^2)$$

where θ_{ij} are elements of the network matrix. Here, $\alpha = 1$ reduces to a lasso penalty since the quadratic term is then removed.

Augmentations to the penalty terms can be made in order to accommodate additional criteria, such as modelling multiple cohorts, i.e. the *joint graphical lasso* [114]. In this setting the penalty term is used to encourage similar network structure and link values across cohorts. The penalized likelihood for the *joint graphical lasso* problem can be solved efficiently using a method called alternating direction method of multipliers (*ADMM*) [114]. The ADMM method is used to decompose the complex objective function so that it can be solved in an iterative manner [114, 115].

Assessing network importance

Network models can naturally be interpreted at the level of single links, i.e. the presence or absence of an association between two nodes. But often, to make more nuanced interpretations, methods to summarize some aspect of the network are used. Summary methods can help highlight important properties of a network and allow us to focus on only that which is relevant for a specific hypothesis. Network summary methods can be broadly categorized into those based on; *centrality*, identifying nodes that have an outstanding role in terms of their linkage to other nodes, *clustering*, identifying groups of nodes in a network that are more connected to nodes within that group than to nodes outside it, *motifs*, identifying recurring structures in the network, and *proximity*, identifying nodes 'guilty by association' to nodes of known importance.

Graph centrality

Graph centrality uses the structure of the network to identify nodes that are more important than others in some sense. A long standing contention is that in biological networks, entities that are connected to many other entities are functionally more important [116, 117]. For networks based on cancer data, centralities have been used to identify candidate cancer genes (e.g. [118, 119, 120]). Depending on which measure of centrality is used, different aspects of the network can be highlighted. *Degree centrality* measures the number (and strength, in a weighted network) of links to a node, giving highly connected nodes higher importance. *Betweenness centrality* measures how often the shortest path between any other two nodes passes through a particular node [121]. This measure for instance highlights nodes that lie between clusters of other nodes. *PageRank* is a method aimed at identifying nodes that are strongly connected to other important nodes, originally developed by Google for ranking webpages [122]. Many other methods exist, such as local clustering [123] or topology based methods [124], other shortest path based methods such as flow centrality [125] and closeness centrality [126], and other

eigenvector based methods such as Katz status [127] and eigenvector centrality [128].

Graph clustering
Graph clustering is the process of partitioning nodes into groups of nodes that belong more with each other than with nodes in other groups. This concept can be used to identify sets of nodes with a distinct function that is not strongly related to other distinct functions, for example to define an ontology of gene function [129] or to define functional modules [130]. Methods for graph clustering can be broadly categorized into two types; those based on computing similarity values between nodes and then deriving clusters from those values, and methods based on a quality function measuring how 'good' a certain partitioning of the network is, with the goal to maximize that value. Methods relying on pairwise similarities between nodes, such as the Jaccard index and topological overlap [131], often serve as the basis for traditional clustering methods such as hierarchical [132] and spectral clustering [133]. Criterion of a 'good clustering' are often built on the notion of *modularity*; determining whether subgraphs have a larger number of internal edges than expected by chance [134]. Thus, modularity depends on assumptions about a reasonable null-model, i.e. a random graph that retains important properties of the graph for which we want to identify clusters. Many methods use a global quality function and there are a vast array of methods for maximizing it, e.g. greedily making small stepwise changes [134], simulated annealing [135] or solving a linear programming problem [136]. For a more extensive overview of clustering methods, see [137].

Graph motifs
Motifs in a network are patterns that appear more often than what is expected by chance [138]. In biological networks, such patterns can be interpreted to reflect true functional relationships that are recurrently present in the underlying biological system. Commonly analyzed network motifs include feedback loops and other structures with relevance for regulatory circuits. Motifs can either be detected *de novo* or a directed search for certain patterns can be performed. To determine the statistical significance of a pattern, a null-model of a random network is needed. Often, a completely random network is not relevant, but a random network retaining key properties of the network in which we are evaluating motifs is needed. Prominent methods for finding motifs *de novo* include MFinder [139], Kavosh [140] and MODA [141]. For an longer introduction to network motifs, see [142, 143].

Graph proximity
Identifying important nodes by some measure of proximity to nodes that are known to be important is in essence a method of *guilt by association* [144]. The underlying assumption being that if we know that something is important,

things influencing that factor have a higher probability of also being important. In a biological setting, this can be useful in finding alternative strategies for targeting known mechanisms. Guilt by association has the double-edged property of relying on *a priory* information, on one hand making new findings dependendent on the completeness of the input, on the other hand making the search for new findings more directed. Notable graph proximity methods to identify candidate genes include those based on direct neighbors [145], shortest paths [146] or methods using network propagation such as random walks [147].

Web based interactive exploration of cancer data

As the size and scope of cancer omics data is expanding, every part of the process from sample gathering, statistical analysis, visualization, and biological interpretation becomes a challenge. A number of hugely popular tools are available for exploring and downloading TCGA data, e.g: TCGA Data Portal (`tcga-data.nci.nih.gov/tcga`), the cBio Portal [148], the UCSC Cancer Genomics Browser [149] and UCSC Xena (`xena.ucsc.edu`). Other popular resources are more focused towards combining statistical modeling with interactive graphics, e.g: IntOGen [150], Gitools [151] and the Cancer Regulome Explorer [152]. These resources have played an important role in making data and data analysis approachable to a wider audience.

Many applications have moved to a web based interface to allow for easier distribution across platforms and devices. Some notable examples of web based software for exploration of cancer omics data include:

Oncoscape (`oncoscape.sttrcancer.io`) [153], a platform for exploring TCGA data using interactive PCA plots, Timelines and Survival analysis.

MAGI (`magi.brown.edu`) [154], a tool for annotation and integration of cancer genomics data. MAGI allows for side by side analysis of public and private data sets.

OASIS (`www.oasis-genomics.org`) [155], developed by Pfizer, includes data from TCGA and CCLE and allows for integrative analysis of multiple data types for both tumor samples and cell lines.

GlioVis (`gliovis.bioinfo.cnio.es`) [156], contains data for over 6500 tumors samples, focusing mainly on gliomas. GlioVis provides tools for classification, gene set enrichment analysis as well as common plotting tools for visualizing data.

Most of these applications heavily utilize modern web technologies such as HTML5 and JavaScript. These technologies lie behind the growing popularity of 'web apps' such as Google Docs and Microsoft Office 365. Web apps have

a number of advantages over traditional software. Firstly, they allow for unparallelled accessibility since no installation of software is necessary and they automatically work across all platforms that support standard web browsers. Secondly, they provide an environment where data can be seamlessly transferred to the user once requested, which minimizes the total bandwidth used. Thirdly, web applications are arguably easier to develop and maintain since much of the basic functionality is provided through the web browser engine and commonly used libraries.

Clearly there are also downsides to using web applications. Chiefly among these are performance. Although applications based on JavaScript have seen impressive performance improvements during recent years, they still lag behind the performance of native desktop apps implemented in languages such as C and Java. Performance can sometimes be an issue, but this issue is alleviated when using a powerful computer or a browser with good JavaScript performance. The use of web technologies is thus in large a tradeoff between accessibility and performance, where many chose to favor accessibility.

Present investigations

Paper I
Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content

In Paper I we develop statistical network models that allow for the concurrent modeling of multiple cancers from The Cancer Genome Atlas while also including multiple types of omics data. Here, we use an augmented version of the SICS framework discussed above. The solution to this problem gives us a set of pairwise measures of conditional independence, which we can view as the adjacency matrix of a network. When including multiple cancers, we call this a pan-cancer network, where an association between two variables can exists in one cancer, all cancers, or any combination of cancers.

To solve the SICS problem for the scale and heterogeneity of data from multiple TCGA cancers (8 cancers, 3900 patients and 4 types of omics data, each with thousands of variables), a number of important extensions were made in a method we call Augmented Sparse Inverse Covariance Selection (aSICS). In particular; a *sample size correction* (SSC) to balance cancers with very different sample sizes, a *data type dependent prior* to promote links supported by external data, and a *modular constraint across cancers* to stabilize the estimated network structure across cancers. This results in an augmented version of the penalized likelihood problem:

$$\underset{\Theta}{\text{argmax:}} \quad \underbrace{l(\Theta|S)}_{\text{likelihood}} - \underbrace{\mathbf{P_S}(\Theta, \text{Prior}, \text{SSC})}_{\text{network sparsity}} - \underbrace{\mathbf{P_d}(\Theta, \text{Modularity}, \text{SSC})}_{\text{modularity constraint}}$$

where $l(\Theta|S)$ is the Gaussian log-likelihood for the networks in Θ, and S consists of one correlation matrix per cancer. To solve this optimization problem we implemented an efficient cluster solver utilizing bootstrapping procedures to improve estimate stability.

The resulting network is presented in a bespoke web based visualization platform. This platform allows for efficient exploration of the pan-cancer network and the underlying data by anyone with a web browser. The web application includes features for browsing the pan-cancer network, filtering for certain types of data, viewing links present in certain combinations of cancers, and for viewing the underlying data as scatterplots and boxplots.

Paper II

The Human Glioblastoma Cell Culture Resource: Validated Cell Models Representing All Molecular Subtypes

In Paper II we introduce a biobank of 48 glioblastoma cell cultures that we call the Human Glioma Cell Culture (HGCC) resource. Importantly, the HGCC cultures are grown under serum-free conditions, making them more representative of the original tumor than if using traditional cell culture methods. In addition, HGCC cells have a well defined origin, have case-matched omics data, and represent all molecular subtypes of glioblastoma (Classical, Mesenchymal, Proneural, and Neural), which is not the case for commonly used glioblastoma cells (e.g. U87MG). All these factors in combination make the HGCC cultures a useful resource for translational glioblastoma research.

The establishment of 48 cell cultures was based on 94 surgical samples from GBM cases presenting at the Uppsala University Hospital. Cells from surgical samples were explanted in serum-free media, grown as spheres for 5-7 days and then kept in laminin coated dishes as adherent monolayer cultures. The successful establishment of a cell culture was associated with a short patient survival. Molecular profiling of the HGCC cultures, using gene expression and copy number arrays, highlighted their similarity to the TCGA, including similar frequencies of chromosome 7 gain and chromosome 10 loss. The gene expression data was used to assign a molecular subtype to each of the HGCC cultures, and to show that the HGCC cultures represent all TCGA subtypes. However, the HGCC did not contain any IDH1-mutated cases, as indicated by exome sequencing.

Experimental profiling of the HGCC cultures consisted of measuring proliferation capacity, tumor initiation capacity and staining for cell lineage markers. Proneural cell lines tended to have a higher proliferative capacity and cell lines with high proliferative capacity were associated with shorter patient survival. All tested cell lines expressed the stem cell markers Nestin and Sox2, consistent with an enrichment of stem-like cells in these cultures. All tested cell cultures of the Proneural, Classical and Neural subtypes initiated tumors in mice, while 7 out of 17 tested Mesenchymal cell cultures gave rise to tumors.

HGCC cells are distributed to collaborators and academic partners via the `hgcc.se` website, in accordance with applicable biobanking laws. Through the website, researchers can explore molecular and clinical data to select the cell cultures of interest to their particular research question. To date, HGCC has received over 100 inquiries and so far distributed cell cultures to over 30 academic partners. Thus, the knowledge surrounding the HGCC will continue to grow, both by our efforts (as in Paper III and IV), and by the efforts of others in the scientific community.

Paper III

Decoding glioblastoma drug responses using an open access library of patient derived cell models

In Paper III we extend the analysis of the HGCC both in terms of number of cell cultures (to over 100) and in terms of the profiling data (adding mutation, methylation and drug response data). A substantial part of this effort was the systematic collection of omics and drug response data for such a large number of cell cultures. We show that the HGCC cultures mimic the TCGA in important respects, such as frequencies of canonical genomic alterations and main axes of transcriptional variation. We perform extensive drug profiling, starting from over 1500 compounds, screening for efficacy to select a panel of 262 compounds that were subsequently profiled in over 100 cell cultures.

To better understand the relationship between HGCC and larger cohorts such as the TCGA we quantify the main sources of shared variation between the datasets using a method called JIVE (Joint and Individual Variance Explained) [157]. This allowed us to determine that the main component of shared transcriptional variation between the datasets corresponds to transcriptional subtype, and to quantify how much of the total variation that is shared between the cohorts. This is different from classifying samples from one dataset based on a signature derived from the other, since it includes no preference of what cohort is considered the baseline.

Unsupervised consensus clustering of cell cultures using the drug response data revealed two main clusters of cell cultures defined by their sensitivity to a set of proteasome inhibitors (we call these groups PI+ and PI-). This feature of the data was both striking and unexpected, warranting extensive experimental follow-up. Briefly, these groups are defined by proteasome inhibitor sensitivity, p53/p21 activity, stemness and protein turnover, and the differential response was also shown in a mouse model.

Based on the omics data we use machine learning to predict the drug response of a cell line to a given drug. For this purpose, we use both linear (elastic net) and non-linear (Random Forest) methods. In general, these methods show similar performance across drugs, instead highlighting the importance of collecting multiple data types, since different data was needed for predicting different drugs with accuracy. Prediction performance is important when prognosticating on future data, and gives a reasonable approximation of the suitability of a personalization strategy. Furthermore, the variables used to make a good prediction can be viewed as putative biomarkers, or used to propose mechanistic hypotheses for the drug effect. To this end, we assemble a network of drug-predictor variable associations for those variables that were important in predicting a drug response with reasonable accuracy. Finally, we propose a simple method for suggesting efficacious drug combinations based on the drug profiling data.

Paper IV

Exploring large scale integrative networks of glioblastoma using hypothesis driven pattern search

In this work we use the methods developed in Paper I applied to the data generated in Papers II and III. Conceptually, there is not much difference between including multiple cancers in an aSICS model and including multiple cohorts for the same cancer. In the first case, we use multiple cancers to improve our estimate of features that they have in common by 'borrowing' information between them. In the second case we 'borrow' information between cohorts, essentially using the larger set of data to improve the model estimate for the smaller set of data. We can use this fact to our advantage when considering a large set of data with limited potential for experimental follow up (TCGA) and a smaller set of data that is more relevant in this regard (HGCC).

Here, we derive an integrated network for HGCC and glioblastoma data from TCGA using two network inference methods; WGCNA and aSICS. We then contextualize the integrated network by annotating it with other relevant data, such as drug target information from STITCH [158] and known cancer genes from COSMIC [9]. To interpret the resulting model, we propose a search method based on pattern-matching. A pattern in the integrated network is a set of links and nodes with certain properties imposed. For example, a gene that is a drug target and a known cancer gene, for which a mutation of that gene is associated with patient survival in both HGCC and TCGA.

To allow the network to be efficiently searched for any such patterns, we develop a web app solution. Here, a user can define a pattern by simply drawing one that corresponds to a particular hypothesis, and submit it as a query to the database. After searching the network using an efficient graph database implementation, the server returns a ranked list of high-scoring subgraphs (based on node importance and link strengths). The user can then click a subgraph and view the underlying data to better understand how it was constructed. Crucially, the user only needs to access a subset of the complete model and data, relaxing an important limitation when we want to expand the scope and size of our models.

Discussion and future perspectives

In conclusion, the work in this thesis follows a trajectory of including more and more relevant data, while developing the methods to accommodate this inclusion in order to understand the therapeutic vulnerabilities of glioblastoma. In Paper I we develop statistical network models that allow for the concurrent modeling of multiple cancers and multiple data types from the TCGA. This method is also applicable to modeling multiple cohorts and data types from the same cancer, as in Paper IV. In Paper II we introduce the HGCC, a new resource of primary patient-derived glioblastoma cell cultures useful for translational research. In paper III we extend the HGCC resource with more extensive molecular profiling and drug response data, identifying a functional dichotomy defined by proteasome inhibitor sensitivity. In Paper IV we use the methods developed in Paper I applied to the data generated in Papers II and III to construct an integrated model of HGCC and TCGA. An overview of the methods and data included in the corresponding papers is shown in Figure 6.

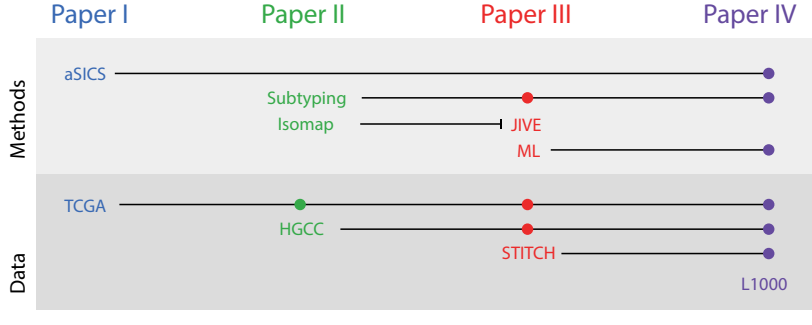

Figure 6. Overview of key methods and data used across the papers in this thesis.

Clinical relevance of HGCC

In Figure 6, the importance of the HGCC resource is made apparent. A key point for the relevance of HGCC cell cultures is whether they are representative of the relevant molecular diversity observed in glioblastoma tumors. In Papers II and III we employ a number of methods to show that HGCC is representative in terms of demographic, clinical and molecular features. One clear limitation is that the HGCC does not seem to contain any IDH-mutated cases, probably due to the tendency of such cells to not grow *in vitro* [159]. Patient

surgical samples that were successfully established as cell cultures were associated with a shorter patient survival, but it is unclear how much of this trend is driven by an exclusion of IDH-mutated cases. Given the estimate of IDH-mutated case frequency at 10% of all glioblastoma [20], and our establishment rate of 56%, there were likely a significant number of IDH-wildtype cases that failed to establish as cell cultures. However, the selection bias seems to favor cell cultures from patients with poor prognosis, which is arguably better than the converse.

In Paper II we use the gene signature proposed by Verhaak et al. [28] and a k-nearest neighbor classifier within a bootstrap step to simultaneously assign subtypes to HGCC cell cultures, and to estimate the stability of that assignment. The bootstrap approach highlighted that most assignments were stable, but a significant instability was observed for some samples. These cases can be seen as intermediaries between subtypes, possibly due to within-sample heterogeneity. As a point of comparison, Patel et al. show that out of 5 samples, 4 bulk sample subtype assignments corresponded to the most frequent assignment for the corresponding single cells [40]. In a similar analysis, Wang et al. report a 4 out of 5 concurrence between bulk and most frequent single cell subtype assignment [38]. Interestingly, in the one exceptional case, 60% of the single cells were classified as mesenchymal, while the bulk tumor was classified as the third most frequently occuring single cell subtype. Such a discrepancy either highlights a methodological instability, or heterogeneity between parts of a single bulk sample, even at a population level, c.f. Sottoriva et al. [39]. In Paper III we use an ensemble of classification methods combined by a majority vote to improve the methodological stability in assigning subtypes to HGCC cell cultures.

The significant heterogeneity between initial tumor and recurrence reported by others [38] has serious implications for the suitability of personalized treatment based on profiling of the initial tumor. If the driver alterations detected in the initial tumor are not the ones driving the recurrence, no amount of prospective predictive power is useful, unless we can also predict the transformation that occurs between initial tumor and recurrence. In fact, the one thing we know for certain is that the cells we study from the surgical sample of the initial tumor are *not* the ones that give rise to the recurrence. There is significant transcriptional diversity within the HGCC cultures [46] and there are significant transcriptional and genomic changes induced by passaging (Baskaran et al. 2018, Accepted). How these factors influence *in vitro* drug response is unclear, and should be studied further. The growing popularity of single cell RNA-seq methods will make this possible at larger scale. Methylation profiles have been shown to be more stable than transcriptional profiles ([38], Baskaran et al. 2018, Accepted), making them an attractive data type for therapeutic markers.

Recent publications exploring the intra-tumor heterogeneity and the similarity between initial tumor and both local and distant recurrences [38, 42, 43,

44] pave the way for a more complete understanding of the hierarchical organization and evolution of glioblastoma. Better understanding of how *in vitro* and *in vivo* models relate to this hierarchy is of key importance. Although a single sample from one tumor does not represent the entire tumor of that patient, a large enough library of samples from different patients can still represent all relevant diversity, but a direct matching between patient and cell model may not be optimal. A personalized treatment approach may require multiple samples to be taken from the initial tumor, in order to understand its therapeutically relevant genetic diversity. Optimally, patient-specific models would consist of multiple samples established from different parts of the tumor, in combination with analysis of clonal populations or single-cell profiling.

Continued work on Paper III

In Paper III, key questions remain unanswered about the mechanism underlying the PI+ and PI- groups. In ongoing work, we are continuing the exploration of the p53 dependency, and the relationship between *TP53* mutation and *TP53* activation. We are also extending the scope of *in vivo* testing to include additional cell cultures and mice. Importantly, the clinical relevance of bortezomib is limited due to lacking blood brain barrier (BBB) penetration. However, second generation proteasome inhibitors such as marizomib have shown promising results in preclinical studies, including BBB penetration [160]. Evaluating the PI+/PI- dichotomy for marizomib in our material is therefore of highest interest. Much effort has been directed towards understanding the PI+/PI- dichotomy, but there are several other drug classes worth exploring based on their differential response across cell cultures, such as HDAC inhibitors and *EGFR* and *PI(3)K* targeting compounds.

In Paper III we show that for at least 25% of drugs the responses of individual cell cultures can be predicted with reasonable accuracy. The prediction performance may not be enough to directly motivate a personalized treatment approach, but can be used to understand underlying mechanisms and to derive putative biomarkers. However, we believe that the prediction performance can be improved and we are pursuing a number of extensions to the current method. These extensions include filtering for variables with known relevance, either by driver events (c.f. [100]) or by membership in frequently altered pathways. Another approach is to include prior information, such as associations between drug effect and drug targets using data from L1000, STITCH or CCLE. We are also exploring the inclusion of additional prediction methods in order to create an ensemble method, combining them based on their individual prediction performance.

Network models and Paper IV

Networks representing a biological system exist somewhere between fully reflecting the underlying regulatory system and being a method for conceptually grouping biological entities. Often, it is not exactly known where on that spectrum a certain network lies. In Paper I we adhere to the SICS framework in order to interpret our links as partial correlations. In Paper IV we include multiple inference methods to create a combined network. Here, we rely on the user defining patterns that are inherently interesting. The inclusion of all glioblastoma subtypes as one class in a multi-class network is bound to introduce spurious correlations based on differences in mean values between subgroups. This problem could be approached by constructing a network where the classes represent different subtypes, ideally in a hierarchical model also including multiple cohorts. This alternative is limited by classes becoming prohibitively small, in particular in the HGCC material, where a subtype may contain as few as 20 samples. However, treating all subtypes as one class may not only be detrimental, since there is significant plasticity between them [38, 41], and such plasticity may represent therapeutic opportunities, c.f. [161, 162]. Another possible extension to the aSICS framework is to hierarchically group nodes into local entities that allow for modeling of non-linear relationships at the local scale, but linear relationships to connect groups of local entities.

An important effort throughout this work has been to generate modular and reproducible results in terms of code and data. The experience gained throughout this work has culminated in a framework for sharing code and data that is being used in two ongoing multi-group collaborative projects. This framework is already being used in Paper IV and makes the inclusion of new data, network inference methods, or additional visualizations straightforward. Comparing the quality of web applications developed in Paper I and Paper IV, significant improvements have been made. The web application in Paper IV is modular, based around data types, cohorts, and data views, while the application in Paper I was built around the network models themselves. However, many improvements can still be made to the modularity of the web application in Paper IV, with a focus on improving maintainability, portability and reproducibility [163, 164].

In Paper IV we combine two network inference methods for HGCC and TCGA data with a number of external sources of data. The network inference methods are used to construct links between variables, and external data sources are used as importance measures for both links and nodes in the network. Once a user searches for a pattern, they are presented with a toplist of matching subgraphs. A clear limitation of this approach is that it encourages a very local interpretation of a limited set of nodes and links. This can be counteracted by visualizing a neighborhood around a given subgraph, for example indicating functional cluster memberships. The pattern search approach is closely related to the concept of network motifs, but does not impose any

criteria that a pattern should be overrepresented in the network, instead using other measures of importance. Graph centralities and graph proximities could be precomputed and included as node importance measures, reserved for future work. In ongoing work we are adding several other network inference methods to the integrated network, including GENIE3 [165] and TIGRESS [166]. The flexibility of the pattern search implementation allows for other future extensions, including addition of other inference methods, datasets and data types.

Throughout this work we combine the development of integrative models with applications to novel data relevant for translational glioblastoma research. This work highlights several potentially therapeutically relevant aspects, and paves a path towards more comprehensive and informative models of glioblastoma. Moving forward there are a number of exciting opportunities to extend both the scope of the included data and the scope of the analysis methods. An important challenge ahead is to balance the inclusion of new data with the interpretability of the resulting models. Developing the results in Paper IV is a promising approach to this challenge.

Future integrative models of glioblastoma have the potential to include multiple cohorts, multiple types of data (at multiple levels of the hierarchy representing *inter-tumor*, *intra-tumor*, *single-cell* and *longitudinal* heterogeneity), to better understand functional diversity of cell models both *in vitro* and *in vivo*, as well as how these factors influence patient survival and therapeutic response. Such methods should not omit other important data such as results from large GWA studies and the significant clinical diversity and non-molecular factors that influence glioblastoma patient survival.

Populärvetenskaplig sammanfattning

Glioblastom är en aggressiv men relativt ovanlig hjärncancer som främst drabbar personer över 40 år. Standardbehandling för patienter med glioblastom innefattar operation där så mycket som möjligt av tumören tas bort och följs sedan av behandling med strålning och cytostatika. Trots aggressiv behandling överlever de flesta patienter inte längre än cirka 15 månader och endast 5% överlever längre än 5 år. Så gott som alla patienter med glioblastom avlider av sin sjukdom, i många fall efter att tumören har vuxit tillbaka trots både operation och efterföljande behandling.

För många typer av cancer har diagnostik och behandlingsmetoder förbättrats väsentligt de senaste årtionden i en glädjande utveckling där patienter överlever längre, får lindrigare sidoeffekter av behandling, och i många fall kan botas helt och hållet. Denna utveckling har dessvärre inte gagnat patienter med glioblastom, men lovande framsteg ger hopp för framtiden. Nya molekylära tekniker har gjort det möjligt att i stor skala studera de underliggande faktorer som ger upphov till glioblastom och att förstå vad som gör vissa tumörer mer aggressiva än andra.

Denna avhandling består av fyra delarbeten med fokus på två nya sätt att förstå glioblastom. Dels genererar vi stora mängder data från en unik samling patientprover och dels utvecklar vi statistiska metoder som vi använder för att tolka denna data. En genomgående ambition i detta arbete har varit att förstå vår egen data bättre genom att jämföra den med data från stora internationella studier.

I delarbete ett utvecklar vi en metod för att generera så kallade statistiska nätverksmodeller som jämför flera uppsättningar patienter med varandra samtidigt som flera molekylära datatyper vägs samman. En nätverksmodell är en matematisk abstraktion där olika mätpunkter i vår data representeras av noder, och kopplingar mellan noder representeras av länkar. Dessa modeller kan användas för att förstå hur olika biologiska mekanismer påverkar varandra, och hur dessa eventuellt kan exploateras för att avväpna tumörceller. I delarbete två introducerar vi en biobank av cellkulturer odlade från glioblastomtumörer som vi kallar HGCC (Human Glioma Cell Culture, `hgcc.se`) och visar att dessa cellkulturer är ett representativt modellsystem för att studera glioblastom. I delarbete tre utför vi omfattande analys av fler än 100 sådana cellkulturer där vi kopplar deras molekylära tillstånd till deras respons då de utsätts för en panel av läkemedel. Vi identifierar två tydliga grupper av cellkulturer som svarar olika på behandling med så kallade proteasomhämmare och kopplar detta beteende till tumörsuppressorgenen *TP53* och cellernas förmåga

att omsätta protein. I delarbete fyra applicerar vi den metod vi utvecklade i delarbete ett till den data vi genererade i delarbete två och tre vilket resulterar i nätverksmodeller som inkluderar såväl data från HGCC som data från en rad stora internationella databaser. För att göra det möjligt att tolka dessa nätverksmodeller utvecklar vi ett websystem där användare kan söka efter små delnätverk som representerar experimentella hypoteser. På detta sätt använder vi de internationella databaserna för att föreslå experimentella försök i cellkulturer från HGCC.

En genomgående utmaning i detta arbete har varit att utveckla de statistiska modeller och de verktyg som behövs för att tolka dessa, samtidigt som vi har inkluderat mer och mer data för att göra dessa modeller så kliniskt relevanta som möjligt. Detta arbete bidrar med att identifierat flera kliniskt relevanta aspekter av glioblastom och banar en väg för framtida modeller med ännu större klinisk relevans.

Acknowledgements

This work was conducted with support from the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, the Swedish Research Council, the Swedish Foundation for Strategic Research, AstraZeneca and Science for Life Laboratory.

This work would not have been possible without a large number of important colleagues and friends. Thank you all!

Most importantly, this work would not have been relevant without the selfless contributions made by patients to the TCGA and HGCC. Your contribution is far above my poor power to thank you.

I would like to thank my supervisor **Sven Nelander** for constant support, inspiration and invaluable guidance. While you are extremely good at taking care of the boring parts of research (applications, deadlines, office politics) you have not once let them stand in the way of good ideas, which is remarkably impressive.
Thank you, my co-supervisor, **Rebecka Jörnsten** for introducing me to the field of statistics applied to biological problems. Without you I would probably be working at Volvo. Your sharp and creative mind (if anything, too creative) has always pushed us forward in new directions.
Thank you, my co-supervisor, **Lene Uhrbom** for providing expertise in the glioblastoma field and for always being grounded and honest.
The three of you have given me a sense of security, a source of inspiration, and faith in the scientific community.

I would like to thank all current and former members of the Nelander Group. **Teresia** for being there in the very beginning, **Linnéa** for your enthusiasm, **Cecilia** for taking the group to new heights, **Torbjörn** for teaching me the importance of constant honesty, **Ludmilla** for making the lab-lab run smoothly, **Karl** for explaining biology in a way I understand, **Evgenia**, **Ingrid**, **Soumi**, **Vikki**, **Caroline** and finally **Elin** and **Emil** - now it's on you.

I would like to thank my office-mates **Holger** and **Anders**. No matter how stressed I have been, somehow I always feel better after a good lunch.

I would like to thank all collaborators and co-authors. In particular **Yuan**, **Tobias** and **Yiwen** for hard work on the HGCC project and **José** for your work on the pan-cancer models. Tobias, for someone who likes to talk a lot you are a great listener.

I would like to thank the whole **Neuro-oncology** constellation at IGP for creating such a positive environment. In particular **Bengt**, **Karin**, **Anna** and **Fredrik**, you set the standards that we all follow. We are clearly stronger in numbers, building on a long tradition so nicely established in Uppsala.

I would like to thank **Uppsala University**, in particular the **Department of Immunology Genetics and Pathology**, for providing a safe and stable research environment. Thank you **Christina** and **Helene** for always keeping track of me, making sure my papers were in order and that I got paid.
Thank you **Lena Claesson-Welsh** and **Karin Forsberg Nilsson** for keeping track of Sven.

I would like to thank a few friends in particular. **Sathish**, you have inspired me with your sense of duty towards your work, but more importantly, towards your family. **Vivek**, **Mohan** and **Priya**, thank you for always helping out and for making me feel welcome in your homes.

Finally, I would like to thank my current and future family.
Work is just work - after all.

Bibliography

[1] *Oxford English Dictionary "cancer, n. and adj."*.

[2] Stewart, B. W. and Wild, C. P. *World Cancer Report 2014*. International Agency for Research on Cancer. World Health Organization, 2014.

[3] Cancerfonden. *Cancerfondsrapporten 2017*. 2017.

[4] Socialstyrelsen. *Antal nya cancerfall, Dödsorsaksstatistik, Antal döda*. [www.socialstyrelsen.se/statistik/statistikdatabas, 2017-08-29] Stockholm, 2017.

[5] Hanahan, D and Weinberg, R. A. "The hallmarks of cancer." *Cell* 100.1 (2000), pp. 57–70. DOI: 10.1016/S0092-8674(00)81683-9.

[6] Hanahan, D. and Weinberg, R. A. "Hallmarks of cancer: the next generation." *Cell* 144.5 (2011), pp. 646–74. DOI: 10.1016/j.cell.2011.02.013.

[7] Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., et al. "The Cancer Genome Atlas Pan-Cancer analysis project." *Nature genetics* 45.10 (2013), pp. 1113–20. DOI: 10.1038/ng.2764.

[8] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., et al. "Cancer genome landscapes." *Science (New York, N.Y.)* 339.6127 (2013), pp. 1546–58. DOI: 10.1126/science.1235122.

[9] Forbes, S. A., Beare, D., Boutselakis, H., et al. "COSMIC: Somatic cancer genetics at high-resolution". *Nucleic Acids Research* 45.D1 (2017), pp. D777–D783. DOI: 10.1093/nar/gkw1121.

[10] Zehir, A., Benayed, R., Shah, R. H., et al. "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients." *Nature medicine* 23.6 (2017), pp. 703–713. DOI: 10.1038/nm.4333.

[11] Hupe, P. *Own work, Licensed under CC BY-SA 3.0 via Wikimedia Commons*. 2017.

[12] Louis, D. N., Ohgaki, H., Wiestler, O. D., et al. "The 2007 WHO classification of tumours of the central nervous system". *Acta Neuropathologica* 114.2 (2007), pp. 97–109. DOI: 10.1007/s00401-007-0243-4.

[13] Louis, D. N., Perry, A., Reifenberger, G., et al. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary." *Acta neuropathologica* 131.6 (2016), pp. 803–20. DOI: 10.1007/s00401-016-1545-1.

[14] Louis, D., Ohgaki, H, Wiestler, O., et al. *WHO Classification of Tumours of the Central Nervous System*. IARC WHO Classification of Tumours Series v. 1. France: International Agency for Research on Cancer, 2016.

[15] Ostrom, Q. T., Gittleman, H., Xu, J., et al. "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2009-2013." *Neuro-oncology* 18.suppl_5 (2016), pp. v1–v75. DOI: 10.1093/neuonc/now207.

[16] Stupp, R., Mason, W. P., Bent, M. J. van den, et al. "Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma". *New England Journal of Medicine* 352.10 (2005), pp. 987–996. DOI: 10.1056/NEJMoa043330.

[17] Schwartzbaum, J. A., Fisher, J. L., Aldape, K. D., et al. "Epidemiology and molecular pathology of glioma." *Nature clinical practice. Neurology* 2.9 (2006), 494–503; quiz 1 p following 516. DOI: 10.1038/ncpneuro0289.

[18] "Brain tumour risk in relation to mobile telephone use: results of the INTERPHONE international case-control study". *International Journal of Epidemiology* 39.3 (2010), pp. 675–694. DOI: 10.1093/ije/dyq079.

[19] Wyde, M., Cesta, M., Blystone, C., et al. "Report of Partial findings from the National Toxicology Program Carcinogenesis Studies of Cell Phone Radiofrequency Radiation in Hsd: Sprague Dawley® SD rats (Whole Body Exposure)". *bioRxiv* (2016).

[20] Nobusawa, S., Watanabe, T., Kleihues, P., et al. "IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas." *Clinical cancer research : an official journal of the American Association for Cancer Research* 15.19 (2009), pp. 6002–7. DOI: 10.1158/1078-0432.CCR-09-0715.

[21] Hegi, M. E., Diserens, A.-C., Gorlia, T., et al. "MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma". *New England Journal of Medicine* 352.10 (2005), pp. 997–1003. DOI: 10.1056/NEJMoa043331.

[22] Li, Y. M., Suki, D., Hess, K., et al. "The influence of maximum safe resection of glioblastoma on survival in 1229 patients: Can we do better than gross-total resection?" *Journal of neurosurgery* 124.4 (2016), pp. 977–88. DOI: 10.3171/2015.5.JNS142087.

[23] Brown, T. J., Brennan, M. C., Li, M., et al. "Association of the Extent of Resection With Survival in Glioblastoma: A Systematic Review and Meta-analysis." *JAMA oncology* 2.11 (2016), pp. 1460–1469. DOI: 10.1001/jamaoncol.2016.1373.

[24] Phillips, H. S., Kharbanda, S., Chen, R., et al. “Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis”. *Cancer Cell* 9.3 (2006), pp. 157–173. DOI: 10.1016/j.ccr.2006.02.019.

[25] Beroukhim, R., Getz, G., Nghiemphu, L., et al. “Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma”. *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 20007–20012. DOI: 10.1073/pnas.0710052104.

[26] Cancer Genome Atlas Research Network. “Comprehensive genomic characterization defines human glioblastoma genes and core pathways.” *Nature* 455.7216 (2008), pp. 1061–8. DOI: 10.1038/nature07385.

[27] Parsons, D. W., Jones, S., Zhang, X., et al. “An Integrated Genomic Analysis of Human Glioblastoma Multiforme”. *Science* 321.5897 (2008), pp. 1807–1812. DOI: 10.1126/science.1164382.

[28] Verhaak, R. G. W., Hoadley, K. a., Purdom, E., et al. “Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1”. *Cancer Cell* 17.1 (2010), pp. 98–110. DOI: 10.1016/j.ccr.2009.12.020.

[29] Noushmehr, H., Weisenberger, D. J., Diefes, K., et al. “Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma”. *Cancer Cell* 17.5 (2010), pp. 510–522. DOI: 10.1016/j.ccr.2010.03.017.

[30] Brennan, C. W., Verhaak, R. G., McKenna, A., et al. “The Somatic Genomic Landscape of Glioblastoma”. *Cell* 155.2 (2013), pp. 462–477. DOI: 10.1016/j.cell.2013.09.034.

[31] Kandoth, C., McLellan, M. D., Vandin, F., et al. “Mutational landscape and significance across 12 major cancer types”. *Nature* 502.7471 (2013), pp. 333–339. DOI: 10.1038/nature12634.

[32] Surget, S., Khoury, M. P., and Bourdon, J.-C. “Uncovering the role of p53 splice variants in human malignancy: a clinical perspective.” *OncoTargets and therapy* 7 (2013), pp. 57–68. DOI: 10.2147/OTT.S53876.

[33] Yung, W. K. A. “From GWAS risk foci to glioma molecular subclass”. *Neuro-Oncology* 15.5 (2013), pp. 513–514. DOI: 10.1093/neuonc/not061.

[34] Melin, B. S., Barnholtz-Sloan, J. S., Wrensch, M. R., et al. “Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors”. *Nature Genetics* 49.5 (2017), pp. 789–794. DOI: 10.1038/ng.3823.

[35] Ohgaki, H. and Kleihues, P. “The definition of primary and secondary glioblastoma.” *Clinical cancer research : an official journal of the American Association for Cancer Research* 19.4 (2013), pp. 764–72. DOI: 10.1158/1078-0432.CCR-12-3002.

[36] Sturm, D., Witt, H., Hovestadt, V., et al. “Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma”. *Cancer Cell* 22.4 (2012), pp. 425–437. DOI: 10.1016/j.ccr.2012.08.024.

[37] Ceccarelli, M., Barthel, F. P., Malta, T. M., et al. “Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.” *Cell* 164.3 (2016), pp. 550–63. DOI: 10.1016/j.cell.2015.12.028.

[38] Wang, Q., Hu, B., Hu, X., et al. “Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment.” *Cancer cell* 32.1 (2017), 42–56.e6. DOI: 10.1016/j.ccell.2017.06.003.

[39] Sottoriva, A., Spiteri, I., Piccirillo, S. G. M., et al. “Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics.” *Proceedings of the National Academy of Sciences of the United States of America* 110.10 (2013), pp. 4009–14. DOI: 10.1073/pnas.1219747110.

[40] Patel, A. P., Tirosh, I., Trombetta, J. J., et al. “Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.” *Science (New York, N.Y.)* 344.6190 (2014), pp. 1396–401. DOI: 10.1126/science.1254257.

[41] Ozawa, T., Riester, M., Cheng, Y.-K., et al. “Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma.” *Cancer cell* 26.2 (2014), pp. 288–300. DOI: 10.1016/j.ccr.2014.06.005.

[42] Wang, J., Cazzato, E., Ladewig, E., et al. “Clonal evolution of glioblastoma under therapy”. *Nature Genetics* 48.7 (2016), pp. 768–776. DOI: 10.1038/ng.3590.

[43] Kim, J., Lee, I.-H., Cho, H. J., et al. “Spatiotemporal Evolution of the Primary Glioblastoma Genome.” *Cancer cell* 28.3 (2015), pp. 318–28. DOI: 10.1016/j.ccell.2015.07.013.

[44] Lee, J.-K., Wang, J., Sa, J. K., et al. “Spatiotemporal genomic architecture informs precision oncology in glioblastoma”. *Nature Genetics* 49.4 (2017), pp. 594–599. DOI: 10.1038/ng.3806.

[45] Meyer, M., Reimand, J., Lan, X., et al. “Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity”. *Proceedings of the National Academy of Sciences* 112.3 (2015), pp. 851–856. DOI: 10.1073/pnas.1320611111.

[46] Segerman, A., Niklasson, M., Haglund, C., et al. “Clonal Variation in Drug and Radiation Response among Glioma-Initiating Cells Is Linked to Proneural-Mesenchymal Transition.” *Cell reports* 17.11 (2016), pp. 2994–3009. DOI: 10.1016/j.celrep.2016.11.056.

[47] Galli, R., Binda, E., Orfanelli, U., et al. “Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma”. *Cancer Research* 64.19 (2004), pp. 7011–7021. DOI: 10.1158/0008-5472.CAN-04-1364.

[48] Pollard, S. M., Yoshikawa, K., Clarke, I. D., et al. “Glioma Stem Cell Lines Expanded in Adherent Culture Have Tumor-Specific Phenotypes and Are Suitable for Chemical and Genetic Screens”. *Cell Stem Cell* 4.6 (2009), pp. 568–580. DOI: 10.1016/j.stem.2009.03.014.

[49] Ledur, P. F., Onzi, G. R., Zong, H., et al. “Culture conditions defining glioblastoma cells behavior: what is the impact for novel discoveries?” *Oncotarget* 8.40 (2017), pp. 69185–69197. DOI: 10.18632/oncotarget.20193.

[50] Pontén, J and Macintyre, E. H. “Long term culture of normal and neoplastic human glia.” *Acta pathologica et microbiologica Scandinavica* 74.4 (1968), pp. 465–86.

[51] Allen, M., Bjerke, M., Edlund, H., et al. “Origin of the U87MG glioma cell line: Good news and bad news.” *Science translational medicine* 8.354 (2016), 354re3. DOI: 10.1126/scitranslmed.aaf6853.

[52] Ishii, N, Maier, D, Merlo, A, et al. “Frequent co-alterations of TP53, p16/CDKN2A, p14ARF, PTEN tumor suppressor genes in human glioma cell lines.” *Brain pathology (Zurich, Switzerland)* 9.3 (1999), pp. 469–79.

[53] Lee, J., Kotliarova, S., Kotliarov, Y., et al. “Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines.” *Cancer cell* 9.5 (2006), pp. 391–403. DOI: 10.1016/j.ccr.2006.03.030.

[54] Kumar, R., Sharma, A., Pattnaik, A., et al. “Stem cells: An overview with respect to cardiovascular and renal disease”. *Journal of Natural Science, Biology and Medicine* 1.1 (2010), p. 43. DOI: 10.4103/0976-9668.71674.

[55] Clarke, M. F., Dick, J. E., Dirks, P. B., et al. “Cancer stem cells–perspectives on current status and future directions: AACR Workshop on cancer stem cells.” *Cancer research* 66.19 (2006), pp. 9339–9344. DOI: 10.1158/0008-5472.CAN-06-3126.

[56] Bomken, S, Fišer, K, Heidenreich, O, et al. “Understanding the cancer stem cell”. *British Journal of Cancer* 103.4 (2010), pp. 439–445. DOI: 10.1038/sj.bjc.6605821.

[57] Singh, S. K., Hawkins, C., Clarke, I. D., et al. “Identification of human brain tumour initiating cells.” *Nature* 432.7015 (2004), pp. 396–401. DOI: 10.1038/nature03128.

[58] Vescovi, A. L., Galli, R., and Reynolds, B. A. “Brain tumour stem cells.” *Nature reviews. Cancer* 6.6 (2006), pp. 425–36. DOI: 10.1038/nrc1889.

[59] Beier, D., Hau, P., Proescholdt, M., et al. “CD133(+) and CD133(-) glioblastoma-derived cancer stem cells show differential growth characteristics and molecular profiles.” *Cancer research* 67.9 (2007), pp. 4010–5. DOI: 10.1158/0008-5472.CAN-06-4180.

[60] Wang, J., Sakariassen, P. Ø., Tsinkalovsky, O., et al. “CD133 negative glioma cells form tumors in nude rats and give rise to CD133 positive cells.” *International journal of cancer* 122.4 (2008), pp. 761–8. DOI: 10.1002/ijc.23130.

[61] Chen, J., Li, Y., Yu, T.-S., et al. “A restricted cell population propagates glioblastoma growth after chemotherapy.” *Nature* 488.7412 (2012), pp. 522–6. DOI: 10.1038/nature11287.

[62] Park, N. I., Guilhamon, P., Desai, K., et al. “ASCL1 Reorganizes Chromatin to Direct Neuronal Fate and Suppress Tumorigenicity of Glioblastoma Stem Cells.” *Cell stem cell* 21.2 (2017), 209–224.e7. DOI: 10.1016/j.stem.2017.06.004.

[63] Zong, H., Verhaak, R. G., and Canoll, P. “The cellular origin for malignant glioma and prospects for clinical advancements”. *Expert Review of Molecular Diagnostics* 12.4 (2012), pp. 383–394. DOI: 10.1586/erm.12.30.

[64] Zong, H., Parada, L. F., and Baker, S. J. “Cell of Origin for Malignant Gliomas and Its Implication in Therapeutic Development”. *Cold Spring Harbor Perspectives in Biology* 7.5 (2015), a020610. DOI: 10.1101/cshperspect.a020610.

[65] Harry S. N. Greene and Hildegarde Arnold. “The Homologous and Heterologous Transplantation of Brain and Brain Tumors”. *Journal of Neurosurgery* 2.4 (1945), pp. 315–331. DOI: 10.3171/jns.1945.2.4.0315.

[66] Greene, H. S. N. “The significance of the heterologous transplantability of human cancer.” *Cancer* 5.1 (1952), pp. 24–44.

[67] Romaguera-Ros, M., Peris-Celda, M., Oliver-De La Cruz, J., et al. “Cancer-initiating enriched cell lines from human glioblastoma: preparing for drug discovery assays.” *Stem cell reviews* 8.1 (2012), pp. 288–98. DOI: 10.1007/s12015-011-9283-1.

[68] Joo, K. M., Kim, J., Jin, J., et al. “Patient-specific orthotopic glioblastoma xenograft models recapitulate the histopathology and biology of human glioblastomas in situ.” *Cell reports* 3.1 (2013), pp. 260–73. DOI: 10.1016/j.celrep.2012.12.013.

[69] Holland, E. C., Celestino, J, Dai, C, et al. “Combined activation of Ras and Akt in neural progenitors induces glioblastoma formation in mice.” *Nature genetics* 25.1 (2000), pp. 55–7. DOI: 10.1038/75596.

[70] Uhrbom, L, Hesselager, G, Nistér, M, et al. “Induction of brain tumors in mice using a recombinant platelet-derived growth factor B-chain retrovirus.” *Cancer research* 58.23 (1998), pp. 5275–9.

[71] Uhrbom, L, Hesselager, G, Ostman, A, et al. “Dependence of autocrine growth factor stimulation in platelet-derived growth factor-B-induced mouse brain tumor cells.” *International journal of cancer* 85.3 (2000), pp. 398–406.

[72] Zhu, H., Acquaviva, J., Ramachandran, P., et al. “Oncogenic EGFR signaling cooperates with loss of tumor suppressor gene functions in gliomagenesis.” *Proceedings of the National Academy of Sciences of the United States of America* 106.8 (2009), pp. 2712–6. DOI: 10.1073/pnas.0813314106.

[73] Zhu, Y., Guignard, F., Zhao, D., et al. “Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma.” *Cancer cell* 8.2 (2005), pp. 119–30. DOI: 10.1016/j.ccr.2005.07.004.

[74] Wang, Y., Yang, J., Zheng, H., et al. “Expression of mutant p53 proteins implicates a lineage relationship between neural stem cells and malignant astrocytic glioma in a murine model.” *Cancer cell* 15.6 (2009), pp. 514–26. DOI: 10.1016/j.ccr.2009.04.001.

[75] Furey, T. S. “ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.” *Nature reviews. Genetics* 13.12 (2012), pp. 840–52. DOI: 10.1038/nrg3306.

[76] Allis, C. D. and Jenuwein, T. “The molecular hallmarks of epigenetic control.” *Nature reviews. Genetics* 17.8 (2016), pp. 487–500. DOI: 10.1038/nrg.2016.59.

[77] Du, P., Zhang, X., Huang, C.-C., et al. “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.” *BMC bioinformatics* 11 (2010), p. 587. DOI: 10.1186/1471-2105-11-587.

[78] Vogel, C. and Marcotte, E. M. “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.” *Nature reviews. Genetics* 13.4 (2012), pp. 227–32. DOI: 10.1038/nrg3185.

[79] Payne, S. H. “The utility of protein and mRNA correlation.” *Trends in biochemical sciences* 40.1 (2015), pp. 1–3. DOI: 10.1016/j.tibs.2014.10.010.

[80] Edfors, F., Danielsson, F., Hallström, B. M., et al. “Gene-specific correlation of RNA and protein levels in human cells and tissues.” *Molecular systems biology* 12.10 (2016), p. 883.

[81] Schena, M, Shalon, D, Davis, R. W., et al. “Quantitative monitoring of gene expression patterns with a complementary DNA microarray.” *Science (New York, N.Y.)* 270.5235 (1995), pp. 467–70.

[82] Wang, Z., Gerstein, M., and Snyder, M. “RNA-Seq: a revolutionary tool for transcriptomics.” *Nature reviews. Genetics* 10.1 (2009), pp. 57–63. DOI: 10.1038/nrg2484.

[83] Fredriksson, S., Gullberg, M., Jarvius, J., et al. “Protein detection using proximity-dependent DNA ligation assays.” *Nature biotechnology* 20.5 (2002), pp. 473–7. DOI: 10.1038/nbt0502-473.

[84] Spurrier, B., Ramalingam, S., and Nishizuka, S. “Reverse-phase protein lysate microarrays for cell signaling analysis.” *Nature protocols* 3.11 (2008), pp. 1796–808. DOI: 10.1038/nprot.2008.179.

[85] Aebersold, R. and Mann, M. “Mass-spectrometric exploration of proteome structure and function.” *Nature* 537.7620 (2016), pp. 347–55. DOI: 10.1038/nature19949.

[86] International Cancer Genome Consortium, Hudson, T. J., Anderson, W., et al. “International network of cancer genome projects.” *Nature* 464.7291 (2010), pp. 993–8. DOI: 10.1038/nature08987.

[87] Pleasance, E. D., Cheetham, R. K., Stephens, P. J., et al. “A comprehensive catalogue of somatic mutations from a human cancer genome.” *Nature* 463.7278 (2010), pp. 191–6. DOI: 10.1038/nature08658.

[88] Ellis, M. J., Gillette, M., Carr, S. A., et al. “Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium.” *Cancer discovery* 3.10 (2013), pp. 1108–12. DOI: 10.1158/2159-8290.CD-13-0219.

[89] Kutalik, Z., Beckmann, J. S., and Bergmann, S. “A modular approach for integrative analysis of large-scale gene-expression and drug-response data.” *Nature biotechnology* 26.5 (2008), pp. 531–9. DOI: 10.1038/nbt1397.

[90] Barretina, J., Caponigro, G., Stransky, N., et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. *Nature* 483.7391 (2012), pp. 603–307. DOI: 10.1038/nature11003.

[91] Garnett, M. J., Edelman, E. J., Heidorn, S. J., et al. “Systematic identification of genomic markers of drug sensitivity in cancer cells”. *Nature* 483.7391 (2012), pp. 570–575. DOI: 10.1038/nature11005.

[92] Yang, W., Soares, J., Greninger, P., et al. “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.” *Nucleic acids research* 41.Database issue (2013), pp. D955–61. DOI: 10.1093/nar/gks1111.

[93] Basu, A., Bodycombe, N. E., Cheah, J. H., et al. “An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules.” *Cell* 154.5 (2013), pp. 1151–61. DOI: 10.1016/j.cell.2013.08.003.

[94] Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., et al. “Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset.” *Cancer discovery* 5.11 (2015), pp. 1210–23. DOI: 10.1158/2159-8290.CD-15-0235.

[95] Subramanian, A., Narayan, R., Corsello, S. M., et al. “A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.” *Cell* 171.6 (2017), 1437–1452.e17. DOI: 10.1016/j.cell.2017.10.049.

[96] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., et al. “Comprehensive identification of mutational cancer driver genes across 12 tumor types”. *Scientific Reports* 3 (2013). DOI: 10.1038/srep02650.

[97] Zack, T. I., Schumacher, S. E., Carter, S. L., et al. “Pan-cancer patterns of somatic copy number alteration”. *Nature Genetics* 45.10 (2013), pp. 1134–1140. DOI: 10.1038/ng.2760.

[98] Ciriello, G., Miller, M. L., Aksoy, B. A., et al. “Emerging landscape of oncogenic signatures across human cancers”. *Nature Genetics* 45.10 (2013), pp. 1127–1133. DOI: 10.1038/ng.2762.

[99] Hoadley, K. A., Yau, C., Wolf, D. M., et al. “Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.” *Cell* 158.4 (2014), pp. 929–944. DOI: 10.1016/j.cell.2014.06.049.

[100] Iorio, F., Knijnenburg, T. A., Vis, D. J., et al. “A Landscape of Pharmacogenomic Interactions in Cancer.” *Cell* 166.3 (2016), pp. 740–754. DOI: 10.1016/j.cell.2016.06.017.

[101] Rubio-Perez, C., Tamborero, D., Schroeder, M. P., et al. “In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities”. *Cancer Cell* 27.3 (2015), pp. 382–396. DOI: 10.1016/j.ccell.2015.02.007.

[102] Langfelder, P. and Horvath, S. “WGCNA: an R package for weighted correlation network analysis”. *BMC Bioinformatics* 9.1 (2008), p. 559. DOI: 10.1186/1471-2105-9-559.

[103] Sun, Y., Zhang, W., Chen, D., et al. “A glioma classification scheme based on coexpression modules of EGFR and PDGFRA.” *Proceedings of the National Academy of Sciences of the United States of America* 111.9 (2014), pp. 3538–43. DOI: 10.1073/pnas.1313814111.

[104] Horvath, S., Zhang, B., Carlson, M., et al. “Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target”. *Proceedings of the National Academy of Sciences* 103.46 (2006), pp. 17402–17407. DOI: 10.1073/pnas.0608396103.

[105] Margolin, A. A., Nemenman, I., Basso, K., et al. “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context”. *BMC Bioinformatics* 7.Suppl 1 (2006), S7. DOI: 10.1186/1471-2105-7-S1-S7.

[106] Faith, J. J., Hayete, B., Thaden, J. T., et al. “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.” *PLoS biology* 5.1 (2007), e8. DOI: 10.1371/journal.pbio.0050008.

[107] Carro, M. S., Lim, W. K., Alvarez, M. J., et al. “The transcriptional network for mesenchymal transformation of brain tumours”. *Nature* 463.7279 (2010), pp. 318–325. DOI: 10.1038/nature08712.

[108] Allen, J. D., Xie, Y., Chen, M., et al. “Comparing statistical methods for constructing large scale gene networks.” *PloS one* 7.1 (2012), e29348. DOI: 10.1371/journal.pone.0029348.

[109] Shmulevich, I., Dougherty, E. R., Kim, S., et al. “Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks.” *Bioinformatics (Oxford, England)* 18.2 (2002), pp. 261–74.

[110] Meinshausen, N. and Bühlmann, P. “High-dimensional graphs and variable selection with the Lasso”. *Annals of Statistics* 34.3 (2006), pp. 1436–1462. DOI: 10.1214/009053606000000281.

[111] Friedman, J., Hastie, T., and Tibshirani, R. “Sparse inverse covariance estimation with the graphical lasso”. *Biostatistics* 9.3 (2008), pp. 432–441. DOI: 10.1093/biostatistics/kxm045.

[112] Tibshirani, R. “Regression Selection and Shrinkage via the Lasso”. *Journal of the Royal Statistical Society B* 58.1 (1996), pp. 267–288. DOI: 10.2307/2346178.

[113] Zou, H. and Hastie, T. “Regularization and variable selection via the elastic net”. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67.2 (2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

[114] Danaher, P., Wang, P., and Witten, D. M. “The joint graphical lasso for inverse covariance estimation across multiple classes”. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 76.2 (2014), pp. 373–397. DOI: 10.1111/rssb.12033.

[115] Gabay, D. and Mercier, B. “A dual algorithm for the solution of nonlinear variational problems via finite element approximation”. *Computers & Mathematics with Applications* 2.1 (1976), pp. 17–40. DOI: 10.1016/0898-1221(76)90003-1.

[116] Barabási, A.-L. and Bonabeau, E. “Scale-free networks.” *Scientific American* 288.5 (2003), pp. 60–9.

[117] Goh, K.-I., Cusick, M. E., Valle, D., et al. “The human disease network.” *Proceedings of the National Academy of Sciences of the United States of America* 104.21 (2007), pp. 8685–90. DOI: 10.1073/pnas.0701361104.

[118] Wachi, S., Yoneda, K., and Wu, R. “Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues.” *Bioinformatics (Oxford, England)* 21.23 (2005), pp. 4205–8. DOI: 10.1093/bioinformatics/bti688.

[119] Xu, J. and Li, Y. “Discovering disease-genes by topological features in human protein-protein interaction network.” *Bioinformatics (Oxford, England)* 22.22 (2006), pp. 2800–5. DOI: 10.1093/bioinformatics/btl467.

[120] Jonsson, P. F. and Bates, P. A. “Global topological features of cancer proteins in the human interactome.” *Bioinformatics (Oxford, England)* 22.18 (2006), pp. 2291–7. DOI: 10.1093/bioinformatics/btl390.

[121] Freeman, L. C. “A Set of Measures of Centrality Based on Betweenness”. *Sociometry* 40.1 (1977), p. 35. DOI: 10.2307/3033543.

[122] Brin, S and Page, L. “The anatomy of a large scale hypertextual Web search engine”. *Computer Networks and ISDN Systems* 30.1/7 (1998), pp. 107–17. DOI: 10.1.1.109.4049.

[123] Watts, D. J. and Strogatz, S. H. “Collective dynamics of ’small-world’ networks.” *Nature* 393.6684 (1998), pp. 440–2. DOI: 10.1038/30918.

[124] Stelzl, U., Worm, U., Lalowski, M., et al. “A human protein-protein interaction network: a resource for annotating the proteome.” *Cell* 122.6 (2005), pp. 957–68. DOI: 10.1016/j.cell.2005.08.029.

[125] Freeman, L. C., Borgatti, S. P., and White, D. R. “Centrality in valued graphs: A measure of betweenness based on network flow”. *Social Networks* 13.2 (1991), pp. 141–154. DOI: 10.1016/0378-8733(91)90017-N.

[126] Freeman, L. C. “Centrality in social networks conceptual clarification”. *Social Networks* 1.3 (1978), pp. 215–239. DOI: 10.1016/0378-8733(78)90021-7.

[127] Katz, L. “A new status index derived from sociometric analysis”. *Psychometrika* 18.1 (1953), pp. 39–43. DOI: 10.1007/BF02289026.

[128] Bonacich, P. “Technique for Analyzing Overlapping Memberships”. *Sociological Methodology* 4 (1972), pp. 176–185.

[129] Dutkowski, J., Kramer, M., Surma, M. A., et al. “A gene ontology inferred from molecular networks.” *Nature biotechnology* 31.1 (2013), pp. 38–45. DOI: 10.1038/nbt.2463.

[130] Mitra, K., Carvunis, A.-R., Ramesh, S. K., et al. “Integrative approaches for finding modular structure in biological networks.” *Nature reviews. Genetics* 14.10 (2013), pp. 719–32. DOI: 10.1038/nrg3552.

[131] Ravasz, E, Somera, A. L., Mongru, D. A., et al. “Hierarchical organization of modularity in metabolic networks.” *Science (New York, N.Y.)* 297.5586 (2002), pp. 1551–5. DOI: 10.1126/science.1073374.

[132] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Vol. 2. New York: Springer, 2009, p. 758.

[133] Donath, W. E. and Hoffman, A. J. “Lower Bounds for the Partitioning of Graphs”. *IBM Journal of Research and Development* 17.5 (1973), pp. 420–425. DOI: 10.1147/rd.175.0420.

[134] Newman, M. E. J. and Girvan, M. “Finding and evaluating community structure in networks.” *Physical review. E, Statistical, nonlinear, and soft matter physics* 69.2 Pt 2 (2004), p. 026113. DOI: 10.1103/PhysRevE.69.026113.

[135] Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. “Modularity from fluctuations in random graphs and complex networks.” *Physical review. E, Statistical, nonlinear, and soft matter physics* 70.2 Pt 2 (2004), p. 025101. DOI: 10.1103/PhysRevE.70.025101.

[136] Agarwal, G. and Kempe, D. “Modularity-maximizing graph communities via mathematical programming”. *European Physical Journal B* 66.3 (2008), pp. 409–418. DOI: 10.1140/epjb/e2008-00425-1.

[137] Fortunato, S. “Community detection in graphs”. *Physics Reports* 486.3-5 (2010), pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

[138] Milo, R, Shen-Orr, S, Itzkovitz, S, et al. “Network motifs: simple building blocks of complex networks.” *Science (New York, N.Y.)* 298.5594 (2002), pp. 824–7. DOI: 10.1126/science.298.5594.824.

[139] Kashtan, N, Itzkovitz, S, Milo, R, et al. “Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs.” *Bioinformatics (Oxford, England)* 20.11 (2004), pp. 1746–58. DOI: 10.1093/bioinformatics/bth163.

[140] Kashani, Z. R. M., Ahrabian, H., Elahi, E., et al. “Kavosh: a new algorithm for finding network motifs.” *BMC bioinformatics* 10 (2009), p. 318. DOI: 10.1186/1471-2105-10-318.

[141] Omidi, S., Schreiber, F., and Masoudi-Nejad, A. “MODA: an efficient algorithm for network motif discovery in biological networks.” *Genes & genetic systems* 84 (2009), pp. 385–395. DOI: 10.1266/ggs.84.385.

[142] Alon, U. “Network motifs: theory and experimental approaches.” *Nature reviews. Genetics* 8.6 (2007), pp. 450–61. DOI: 10.1038/nrg2102.

[143] Wong, E., Baur, B., Quader, S., et al. “Biological network motif detection: principles and practice.” *Briefings in bioinformatics* 13.2 (2012), pp. 202–15. DOI: 10.1093/bib/bbr033.

[144] Oliver, S. “Guilt-by-association goes global.” *Nature* 403.6770 (2000), pp. 601–3. DOI: 10.1038/35001165.

[145] Oti, M, Snel, B, Huynen, M. A., et al. “Predicting disease genes using protein-protein interactions.” *Journal of medical genetics* 43.8 (2006), pp. 691–8. DOI: 10.1136/jmg.2006.041376.

[146] Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., et al. “Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease.” *Proceedings of the National Academy of Sciences of the United States of America* 101.42 (2004), pp. 15148–53. DOI: 10.1073/pnas.0404315101.

[147] Köhler, S., Bauer, S., Horn, D., et al. “Walking the interactome for prioritization of candidate disease genes.” *American journal of human genetics* 82.4 (2008), pp. 949–58. DOI: 10.1016/j.ajhg.2008.02.013.

[148] Gao, J., Aksoy, B. A., Dogrusoz, U., et al. “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal”. *Science Signaling* 6.269 (2013), pl1–pl1. DOI: 10.1126/scisignal.2004088.

[149] Goldman, M., Craft, B., Swatloski, T., et al. “The UCSC Cancer Genomics Browser: update 2013”. *Nucleic Acids Research* 41.D1 (2013), pp. D949–D954. DOI: 10.1093/nar/gks1008.

[150] Gundem, G., Perez-Llamas, C., Jene-Sanz, A., et al. “IntOGen: integration and data mining of multidimensional oncogenomic data”. *Nature Methods* 7.2 (2010), pp. 92–93. DOI: 10.1038/nmeth0210-92.

[151] Perez-Llamas, C. and Lopez-Bigas, N. “Gitools: Analysis and Visualisation of Genomic Data Using Interactive Heat-Maps”. *PLoS ONE* 6.5 (2011). Ed. by Aerts, S., e19541. DOI: 10.1371/journal.pone.0019541.

[152] Cancer Genome Atlas Network. “Comprehensive molecular characterization of human colon and rectal cancer.” *Nature* 487.7407 (2012), pp. 330–7. DOI: 10.1038/nature11252.

[153] Bolouri, H., Zhao, L. P., and Holland, E. C. “Big data visualization identifies the multidimensional molecular landscape of human gliomas”. *Proceedings of the National Academy of Sciences* 113.19 (2016), pp. 5394–5399. DOI: 10.1073/pnas.1601591113.

[154] Leiserson, M. D. M., Gramazio, C. C., Hu, J., et al. “MAGI: Visualization and collaborative annotation of genomic aberrations”. *Nature Methods* 12.6 (2015), pp. 483–484. DOI: 10.1038/nmeth.3412.

[155] Fernandez-Banet, J., Esposito, A., Coffin, S., et al. “OASIS: Web-based platform for exploring cancer multi-omics data”. *Nature Methods* 13.1 (2015), pp. 9–10. DOI: 10.1038/nmeth.3692.

[156] Bowman, R. L., Wang, Q., Carro, A., et al. “GlioVis data portal for visualization and analysis of brain tumor expression datasets.” *Neuro-oncology* 19.1 (2017), pp. 139–141. DOI: 10.1093/neuonc/now247.

[157] Lock, E. F., Hoadley, K. A., Marron, J. S., et al. “Joint and individual variation explained (JIVE) for integrated analysis of multiple data types”. *Annals of Applied Statistics* 7.1 (2013), pp. 523–542. DOI: 10.1214/12-AOAS597.

[158] Kuhn, M., Szklarczyk, D., Franceschini, A., et al. “STITCH 3: zooming in on protein-chemical interactions.” *Nucleic acids research* 40.Database issue (2012), pp. D876–80. DOI: 10.1093/nar/gkr1011.

[159] Balvers, R. K., Kleijn, A., Kloezeman, J. J., et al. “Serum-free culture success of glial tumors is related to specific molecular profiles and expression of extracellular matrix-associated gene modules.” *Neuro-oncology* 15.12 (2013), pp. 1684–95. DOI: 10.1093/neuonc/not116.

[160] Di, K., Lloyd, G. K., Abraham, V., et al. “Marizomib activity as a single agent in malignant gliomas: ability to cross the blood-brain barrier.” *Neuro-oncology* 18.6 (2016), pp. 840–8. DOI: 10.1093/neuonc/nov299.

[161] Bhat, K. P. L., Balasubramaniyan, V., Vaillant, B., et al. “Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma.” *Cancer cell* 24.3 (2013), pp. 331–46. DOI: 10.1016/j.ccr.2013.08.001.

[162] Halliday, J., Helmy, K., Pattwell, S. S., et al. “In vivo radiation response of proneural glioma characterized by protective p53 transcriptional program and proneural-mesenchymal shift.” *Proceedings of the National Academy of Sciences of the United States of America* 111.14 (2014), pp. 5248–53. DOI: 10.1073/pnas.1321014111.

[163] Schultheiss, S. J. “Ten simple rules for providing a scientific Web resource.” *PLoS computational biology* 7.5 (2011), e1001126. DOI: 10.1371/journal.pcbi.1001126.

[164] Helmy, M., Crits-Christoph, A., and Bader, G. D. “Ten Simple Rules for Developing Public Biological Databases.” *PLoS computational biology* 12.11 (2016), e1005128. DOI: 10.1371/journal.pcbi.1005128.

[165] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., et al. “Inferring regulatory networks from expression data using tree-based methods.” *PloS one* 5.9 (2010). DOI: 10.1371/journal.pone.0012776.

[166] Haury, A.-C., Mordelet, F., Vera-Licona, P., et al. “TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.” *BMC systems biology* 6 (2012), p. 145. DOI: 10.1186/1752-0509-6-145.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1426

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-340843

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2018