Biomedical Data Integration in Cancer Genomics

Riku Louhimo

Research Programs Unit,
Genome-Scale Biology
Medicum,
Biochemistry and Developmental Biology
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 3, Haartmaninkatu 8, Helsinki,
on 4 September 2015, at 12 o'clock noon.

Helsinki 2015

Supervisor

Sampsa Hautaniemi, DTech, Professor
Genome-Scale Biology Research Program, Medicum
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Reviewers appointed by the Faculty

Merja Heinäniemi, PhD, Adjunct Professor
Institute of Biomedicine, University of Eastern Finland
Kuopio, Finland

Päivi Onkamo, PhD, Adjunct Professor
Department of Biosciences, University of Helsinki
Helsinki, Finland

Opponent appointed by the Faculty

Sol Efroni, PhD, Senior Lecturer
Faculty of Life Sciences, Bar-Ilan University
Ramat-Gan, Israel

ISBN 978-951-51-1433-4 (paperback)
ISBN 978-951-51-1434-1 (PDF)
`http://ethesis.helsinki.fi`
Unigrafia Oy
Helsinki 2015

To my family

In spite of what you would suppose, the facts are not reversible.
Paul Auster, In the Country of Last Things

Abstract

Cancer is one of the leading causes of death in industrialized nations and its incidence is steadily increasing due to population aging. Cancer constitutes a group of diseases characterized by unwanted cellular growth which results from random genomic alterations and environmental exposure. Diverse genomic and epigenomic alterations separately and jointly regulate gene expression and stimulate and support neoplastic growth. More effective treatment, earlier and more accurate diagnosis, and improved management of cancer are important for public health and well-being.

Technological improvements in data measurement, storing and transport capability are transforming cancer research to a data-intensive field. The large increases in the quality and quantity of data for the analysis and interpretation of experiments has made employing computational and statistical tools necessary. Data integration — the combination of different types of measurement data — is a valuable computational tool for cancer research because data integration improves the interpretability of data-driven analytics and can thereby provide novel prognostic markers and drug targets.

I have developed two computational data integration tools for large-scale genomic data and a simulator framework for testing a specific type of data integration algorithm. The first computational method, CNAmet, enhances the interpretation of genomic analysis results by integrating three data levels: gene expression, copy-number alteration, and DNA methylation. The second computational method, GOPredict, uses a knowledge discovery approach to prioritize drugs for patient cohorts thereby stratifying patients into potentitally drug-sensitive subgroups. Using the simulator framework, we are able to compare the performance of integration algorithms which integrate gene copy-number data with gene expression data to find putative cancer genes.

Our experimental results indicate in simulated, cell line, and primary tumor data that well-performing integration algorithms for gene copy-number and expression data use and process genomic data appropriately. Applying these methods to diffuse large B-cell lymphoma, integrative analysis of copy-number and expression data helps to uncover a gene with putative prognostic utility. Furthermore, analysis of glioblastoma brain cancer data with CNAmet suggests that a number of known cancer genes, including the epidermal growth factor receptor, are highly expressed due to co-occuring alterations in their promoter DNA methylation and copy-number. Finally, integration of publicly available molecular and literature data with GOPredict suggests that treating patients with FGFR inhibitors in breast cancer and CDK inhibitors in ovarian cancer could support standard drug therapies. Collectively, the methods developed here and their application to varied molecular cancer data sets illustrates the benefits of data integration in cancer genomics.

Tiivistelmä

Syöpä on yksi yleisimmistä kuolinsyistä teollisuusmaissa ja sen esiintyvyys kasvaa tasaisesti väestön vanhetessa. Syöpä käsittää joukon sairauksia, joiden yhteispiirteenä on ei-toivottu solujen uudiskasvu. Uudiskasvu on seurausta genomin sattumanvaraisista sekä ympäristövaikutteisista muutoksista. Monitahoiset genomiset ja epigenomiset muutokset yhdessä ja erikseen säätelevät ja ohjaavat geenien ilmentymistä sekä edesauttavat ja tukevat syövän kasvamista. Hoidon tehostaminen, aikaisempi ja osuvampi taudin määritys, ja parempi syövänhallinta ovat merkittäviä haasteita kansanterveydelle.

Teknologinen kehitys tiedon mittauksessa, säilömisessä ja siirrossa on muuttanut syöpätutkimuksen dataintensiiviseksi alaksi. Aineistojen määrän ja laadun suuri kasvu on tehnyt laskennallisista ja tilastollisista menetelmistä välttämättömiä työkaluja. Data-integraatio — erilaisten mitta-aineistojen yhdistäminen — on syöpätutkimukselle arvokas laskennallinen työkalu, sillä sen käyttö parantaa aineistolähteisen tutkimuksen tulkintaa ja tällä tavoin edesauttaa uusien ennustetekijöiden ja lääkekohteiden tunnistamista.

Olen kehittänyt kaksi laskennallista työkalua suurien genomiaineistojen yhdistämiseen sekä aineistosimulaattorin erityyppisten genomisten aineistojen yhdistämisohjelmien koestamiseen. Ensimmäinen laskennallinen työkalu, CNAmet, parantaa genomiaineistojen analyysin tulkintaa yhdistämällä kolmea eri tyyppistä mittaustietoa: geeni-ilmentymän, kopiolukumuutosten ja DNA-metylaation. Toinen laskennallinen työkalu, GOPredict, käyttäen automaattista tiedonmääritystä panee lääkkeet tärkeysjärjestykseen potilaskohortissa ja täten tunnistaa mahdollisesti lääkeherkät potilasalijoukot. Aineistosimulaattorilla vertailemme eri yhdistämisalgoritmien suorityskykyä menetelmillä, jotka yhdistävät geenien kopiolukumittaustietoa ja ilmentymämittaustietoa löytääkseen mahdollisia syöpägeenejä.

Kokeelliset tuloksemme simulaatio-, solulinja- ja kasvainaineistoissa osoittavat, että parhaat kopioluvun ja geeninilmentymistä yhdistävät työkalut käsittelevät kopiolukumittauksia oikealla tavalla. Kun näitä menetelmiä käytetään suurisoluiseen B-solulymfoomaan, geenien kopioluku- ja ilmentymätiedon yhdistäminen auttaa löytämään mahdollisen ennustetekijägeenin. Glioblastooma syöpäkasvaimien analysointi CNAmet-työkalulla antaa osviittaa, että osa tunnetuista syöpägeeneistä ilmenee voimakkaasti johtuen samanaikaisesti sattuvista muutoksista geenien promoottorien DNA-metylaatiossa ja geenien kopioluvussa. Lopuksi, avoimen molekulääristen ja kirjallisuusaineistojen yhdistäminen GOPredictillä antaa ymmärtää, että FGFR-estäjien käyttö rintasyövässä ja CDK-estäjien käyttö munasarjasyövässä saattaisi tukea vakiohoitoja. Kaiken kaikkiaan tässä työssä kehitetyt työkalut ja niiden käyttö monitahoisiin molekyläärisiin syöpäaineistoihin havainnollistavat data-integraation käytön hyödyllisyyden syöpägenomisten aineistojen käsittelyssä.

Contents

Abbreviations

aCGH	Array comparative genomic hybridization
BRCA	Invasive breast carcinoma
CDK	Cyclin dependent kinase
CNA	Copy-number alteration
CNV	germline copy-number variation
COAD	colon adenocarcinoma
COSMIC	Catalogue of Somatic Mutations in Cancer database
DLBCL	Diffuse large B-cell lymphoma
EGFR	Epidermal growth factor receptor
ERBB2	Human epidermal growth factor receptor 2
FIGO	International Federation of Gynecology and Obstetrics
FGFR	Fibroblast growth factor receptor
GBM	Glioblastoma multiforme
GO	Gene Ontology
HER2	Human epidermal growth factor receptor 2 (also known as ERBB2)
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papillomavirus
IHC	Immunohistochemistry, immunohistochemical
LOH	Loss-of-heterozygosity
LUSC	Lung squamous cell carcinoma
mRNA	messenger RNA
OVCA	Ovarian carcinoma
qPCR	Quantitative real-time polymerase chain reaction
sCNA	Somatic copy-number alteration
TCGA	The Cancer Genome Atlas project and database
TSS	Transcription start site

Publications and author's contributions

Publication I **Riku Louhimo**, Tatiana Lepikhova, Outi Monni, Sampsa Hautaniemi.
Comparative analysis of algorithms for integration of copy number and expression data.
Nature Methods, 2012, 9(4), 351-355.

Publication II **Riku Louhimo**, Sampsa Hautaniemi.
CNAmet: an R package for integrating copy number, methylation and expression data.
Bioinformatics, 2011, 27(6), 887-888.

Publication III **Riku Louhimo***, Marko Laakso*, Denis Belitskin, Juha Klefström, Rainer Lehtonen, Sampsa Hautaniemi.
Data integration to prioritize drugs using genomics and literature data.
Submitted.

* equal contribution

Author's contributions

Publication I Designed the study, developed the simulator framework for integrative copy-number gene-expression data sets, analyzed the copy-number and expression data, and wrote the paper.

Publication II Designed the study, developed the CNAmet integration algorithm, analyzed the data, and wrote the paper.

Publication III Designed the study, developed the algorithms, wrote the formal description, analyzed the data (all together with ML), and wrote the paper.

Related publication

Policies of the Faculty of Medicine at the University of Helsinki limit the use of publications in multiple doctoral dissertations. The following publication is therefore included as a "related publication".

Related Publication I Minna Taskinen, **Riku Louhimo**, Satu Koivula, Ping Chen, Ville Rantanen, Harald Holte, Jan Delabie, Marja-Liisa Karjalainen-Lindsberg, Magnus Björkholm, Øystein Fluge, Lars Møller Pedersen, Karin Fjordén, Mats Jerkeman, Mikael Eriksson, Sampsa Hautaniemi, Sirpa Leppä.
Deregulation of COMMD1 Is Associated with Poor Prognosis in Diffuse Large B-cell Lymphoma.
PLoS ONE, 2014, 9(3), e91031.

Author's contributions to related publication

Related Publication I Analyzed copy-number alteration data, integrated copy-number and expression data, and analyzed the survival impact of sCNA.

1 Introduction

We live in an increasingly data rich society. The digitalization of most daily activities creates massive amounts of structured and unstructured data. Our consumption habits offline and online, commuting habits from our cell phones and travel cards, and health records quantify who we are and what we do [1]. Increases in volume and frequency of accurate measurement data in digital form enable its beneficial use which requires data integration. Data integration — the combination of information from multiple sources — and analytics are becoming key aids in improving industrial and commercial processes as well as management decision making [2]. Data integration is a valuable tool because its usage improves the interpretability of data-driven analytics thereby increasing the value of both analytics and data. In science, the sharp increase in quantitative data drives the emergence of integrative studies in such diverse fields as musicology [3], urban geography [4] and medicine [5].

Integration of measurement data has been central for molecular biology from its beginnings in 1950's, emerging in the wake of Watson and Crick's central hypothesis of molecular biology [6]. The importance of data integration to molecular biology shows a steady increase. Searching the life science and biomedical literature database PubMed for mentions of "data integration" between 1980 and 2014 shows how between 1980 and 1989 data integration is only mentioned once. This is in sharp contrast to 209 mentions in 2014 alone. The growth has been driven by the arrival of measurement technologies that efficiently produce large quantities of molecular data and databases that store and manage the data. These and many other technological advances together enable modern integrative molecular biology.

Cancer is one of the key diseases studied by molecular biology. Cancer is a group of diseases characterized by unwanted cellular growth and many forms of cancer are still incurable. As one of the leading causes of death globally, cancer is under intensive research. The technological development in molecular biology over the last two decades has greatly increased the amount of information that can be time- and cost-efficiently quantified from tumor specimens. Although the understanding of molecular and especially genomic changes in cancer has increased, improvements have been slow to transition to patient care as better treatments or diagnostic tools.

As a specialty in cancer research, cancer genomics has taken center stage in the molecular study of cancer [5]. Genetics and genomics have in particular considerably benefited from the improvements in measurement technology. In addition to quantity, the number of different types of molecular data, that can be measured in large quantities, has grown substantially. As a result, data integration

has become an important tool for genomic analysis in improving interpretation of experimental results [7]. As a result of the increase in data volume, the use and development of computational data integration methods have become necessities. Novel insights from cancer genomics include discoveries of new cancer promoting and suppressive genes, characterization of novel drug target proteins, and discovery and improvement of molecular stratification of cancer [8].

In addition to physical structural alterations in the genome, epigenetic changes are also widespread in cancer. Evidence for the role of epigenetic modifications in cancer came to light soon after the discovery of the first cancer gene [9]. Nonetheless, little is known of the relationship between transient epigenetic changes and genomic alterations and their interplay in regulating gene expression. Data integration is necessary for clarifying this multilevel relationship.

Efforts to utilize large-scale measurement data in industrial and scientific applications have become the current fad due to the increase in available information. In genomics, large-scale data refer to the enormous quantity of molecular data that can be extracted from single tumors. With the increases in size and volume of raw data, the number and scope of entire experiments has also grown with the largest cancer studies enrolling tens of thousands of patients. In addition to raw data, storage, management and analysis of results from these experiments have needed new databases which have lately burgeoned. These databases include databases that provide interpretation-ready information — such as gene-phenotype information or gene-gene regulatory pathway data — and databases storing unprocessed measurement data from cancer specimens [10]. Harnessing the power of these databases to their full extent remains an important topic in computational biology and a challenge which data integration is fit to tackle.

Improving the diagnosis, treatment and management of cancer is important for public health and well-being. Results and methods described in this thesis provide solutions and improvements to (1) computational data integration methods via a new integration algorithm and a comprehensive framework to improve algorithm development; (2) to diagnosis via new putative prognostic markers from integrative experiments; and (3) to treatment by providing suggestions for preferential drugs to patient subgroups.

2 Data integration

Why should data be integrated? Integration is the act or process of combining or fusing varied features of an object of interest. When we measure different aspects of an object, we combine as many different measurements — data levels — as possible when we want to form a complete picture of the object. To characterize, say, this book, we could measure its weight and dimensions (a few hundred grams and $17 \times 25 \times 0.5$ centimeters), its material composition (paper and ink), and count the number of pages and words it contains. Our task at hand determines which of these measurements we need. If we want to ship the book, we need its size and shape. If we wish to quantify the length of the book, we need the number of pages and words. For ascertaining that it is, in fact, a book, we need to combine all three of these measurements.

The book example illustrates an important aspect of data integration: the interdependence of what has been measured and what can be determined from the measurements. We can with some certainty predict that our object is a book even if we only measure two of the three properties. One property, however, would not be enough. Seeing the entire system at the same time instead of looking at individual parts separately contributes to better understanding of the whole. We are unlikely to be able to exhaustively measure any object but data integration is a means to see as much of the whole as possible. Working with the whole enables us to observe and infer knowledge without having to break things apart first [11].

In an actual scientific experiment, we often only measure a limited number of data levels due to costs. The parameters we measure narrow down the research questions which we can answer. On the other hand, determining the research question first enables us to choose the appropriate data levels to measure. Ideally, every type of data would be incorporated into an integrative analysis. In reality, as much as possible of the data, or the data which best answer a given research question, are utilized. Furthermore, some measurement levels are more closely related than others and provide more information when integrated. Some levels can even be completely unrelated and do not provide additional information if integrated and can even lead to false conclusions.

The current trend in science to make available bigger and more fine-grained data sets opens up the possibility of data integration [7]. The increase in size has been fueled by increases in measurement accuracy which together make integration technically and statistically challenging. The sheer size of current data sets creates technical issues and calls for computationally optimized tools. In addition, data sources are often heterogeneous and variable in quality. Biological and technical noise

can stem from impure tumor specimens, unoptimized measurement protocols, and improper computational data preprocessing. Heterogeneous data sources produce utilization challenges which need to be tackled in data processing and interpretation. Together, these technical and sampling issues create a demand for computational and statistical data integration tools which take advantage of existing subject-field knowledge, and are computationally efficient and statistically rigorous.

Molecular biology has experienced a rapid transformation into a data intensive field of science. Genome-wide measurements of DNA, mRNA and DNA methylation have increased the use of statistical and computational tools. A large application field for these measurements has been the study of tumors and tumor cell lines. Computational data integration, too, has emerged as a necessary approach to tackle the inherent noise and complexity of tumor samples [7, 8].

Two aspects of data integration are discussed in this chapter. I start with the conceptualization of two broad questions to which integration can be applied when studying cancer: to (1) combine dependent but different information sources from the same subject and (2) combine independent information from the same subjects or the same information from independent subjects. Second, I describe the four different design principles by which computational integration methods are developed. The first part focuses on data and data sets. The second part focuses on how data analysis methods are applied to integrate data. Further examples of these conceptualizations are discussed in Chapter 3.

2.1 Defining data integration

Data integration can be understood in three ways. First, data integration is a way of analyzing data by combining data originating from multiple sources. Second, data integration is a set of tools to combine data from multiple sources. Third, integration refers to the combination of evidence, supporting a given hypothesis, which accumulates through the repetition, replication and combination of observations from experiments. Here we deal with how data integration as a means and as a set of tools can be used in the molecular study of cancer.

Data can be combined and integrated in many different ways depending on the type of available data, the number of available samples and the specific research question at hand. I therefore start by introducing key concepts that define different ways of applying data integration. On a conceptual level there are three different ways to understand and use data integration (Figure 1).

In **complementary** or dependent integration, data from two or more measurements on the same object – such as a gene or sample – are combined to understand

co-dependence, concomitant changes or correlative relationships (Figure 1a). For example, we can measure both the mRNA and DNA from a tumor and find genes where changes in DNA alter abundance of mRNA. Complementary integration infers correlative and putatively causal relationships. In studies where thousands of genes are measured and appear to be altered, complementary integration helps in distinguishing truly altered genes from spurious or incidental alterations. Current large-scale data have brought computational approaches to the forefront of complementary integration but single-gene in vitro and in vivo studies are also instances of complementary integration when two or more data levels are measured.

In **parallel** or independent integration, information is combined from either measuring independent aspects of the same samples or the same measurements of independent samples (Figure 1b). The requirement, that data are independent, distinguishes parallel integration from complementary integration. For example, Yuan and colleagues combined microscopy imaging data on the cellular admixture of tumors with transcriptional and gene-gene network data to find how cellular structure can be used as a prognostic marker [12, 13]. Thus, structural data is parallel to molecular data. Many studies use parallel integration when stressing the importance of the study's results by referencing previous, independent results or replicating their results in an independent but similar sample set. In fact, parallel integration is deep-rooted in the scientific process as the growth of supporting information from independent sources over time is used to accept scientific theory as scientific knowledge.

In essence, parallel integration is the combination of all types of data from independent sources or samples whereas complementary integration is the combination of different kinds of data from the same individual samples. **Total** integration refers to applications where the two approaches are combined (Figure 1b).

The three concepts — complementary, parallel and total data integration — are independent of the size of available data sets. However, the current trends and technology in cancer studies favor the creation of data sets with hundreds of samples and millions of data points. The volume and dimensionality of the data, which large scale experiments produce, call for the development and application of computational methods capable of efficiently and rigorously integrating and analyzing complementary and parallel data.

2.2 Conceptualizing the design of data integration methods

Computational algorithms are step-by-step sets of operations starting from a predefined input and ending in a predefined output. In this thesis, I use *algorithm* and the more general word *computational method* interchangeably. Both

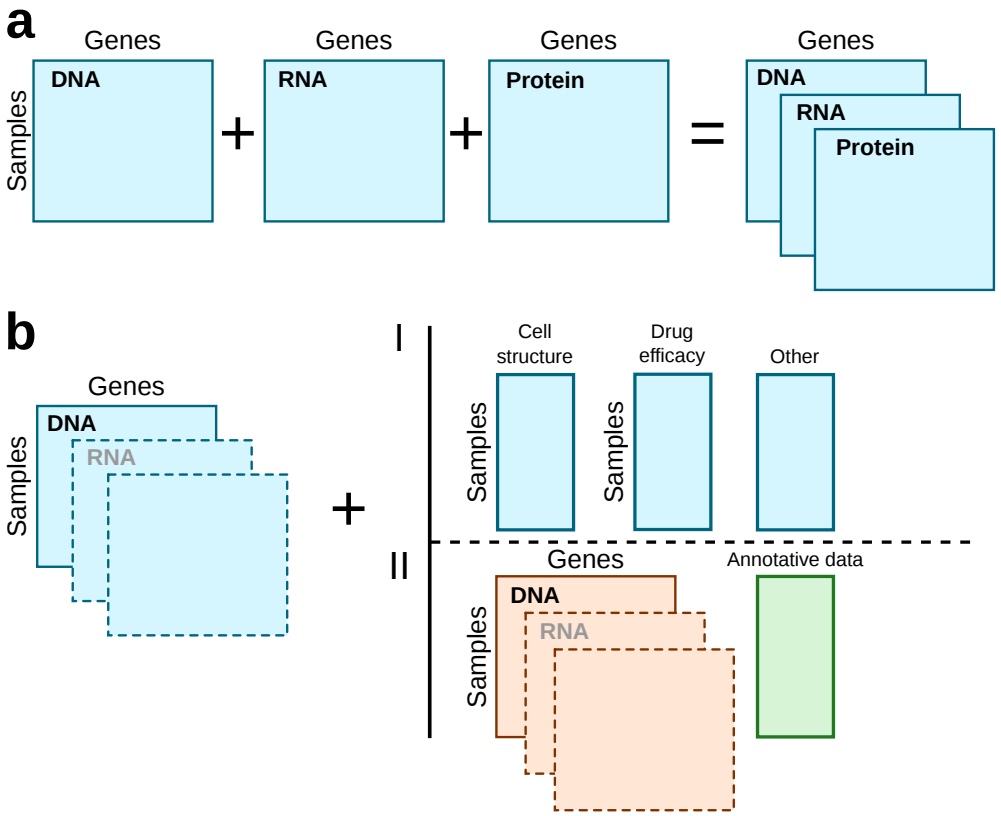

Figure 1: Different types of data integration. Rectangles are abstractions of data matrices and represent different molecular data levels. Blue, red and green colors denote sample sets: all blue boxes depict the same samples which are distinct from red samples. (a) In complementary integration, data levels overlap on both the y-axis (samples) and x-axis (genes). (b) In parallel and total integration, data are independent either on the x-axis (measurement data e.g., genes, drugs) (I) or sample-axis (II). Total integration combines complementary (rectangles with dashed borders) and parallel integration. Annotative data include, for example, gene regulatory networks, which are independent on the sample-axis.

complementary and parallel integration with modern large-scale molecular data require computational methods which I categorize into four groups. The four categories of data integration algorithms provide a backbone for the characterization of differences in methods. Furthermore, the categories conceptualize how statistical and machine learning methods are applied to integrative analyses.

Black-box algorithms treat measurements blindly as numerical data and ignore possible biological interdependence that may exist between data levels. For example, linear regression can be applied in a black-box manner: in a linear regression model, where co-variates are complementary DNA methylation and mRNA data, each gene in the model has two measurements that are known to be co-dependent. Black-box methods can be employed with cursory or no knowledge of what the data biologically represents. Furthermore, black-box methods can require only light data

preprocessing — such as standardization of covariates — and therefore are efficient to implement and use. As a downside, black-box methods do not fully or at all employ the inherent knowledge of how different types of biological data interact and are structured in reality. For example, the genome is sequential and therefore DNA abundance (copy-number) in adjacent regions is highly correlated which should be taken into account in analysis. In particular, the omission of biological knowledge hinders applying black-box methods to complementary integration. For example, DNA, mRNA and protein measurements are directly connected and with a known directionality. Lastly, interpreting results can be difficult because black-box algorithms have not been designed to answer a specific biological question.

In contrast to black-box algorithms, a **controlled** method utilizes biological knowledge on the relationship between integrated data levels. For example, changes in DNA abundance cause changes *in cis* in mRNA abundance [14]. Therefore an increase in gene copy-number should implicate an increase in mRNA level and not the other way around. Controlled methods take this causative relationship and its directionality into account. As a downside, controlled methods can be highly specific to a particular application and therefore have limited applicability to additional research questions. Furthermore, controlled methods depend on the current level of biological knowledge. Controlled methods can require more cumbersome preprocessing steps than black-box methods. Controlled methods improve interpretability of results over black-box methods but require using more time for implementation and computation, and are poorly applicable to general use if at all.

Abstraction methods are a special case of controlled methods. In abstraction methods, the data levels are categorized or labeled prior to integration and integration is carried out on the categorized, abstracted, level via logical rules. For example, gene centric data from different levels can be categorized on the gene expression (on/off, high/medium/low) and copy-number levels (amplified/deleted) and then fused. The integration model can include rules such as (gene copy-number high -> gene expression high -> 1) or (gene copy-number low -> gene expression low -> -1), that define different categories (codings) for different situations. Abstraction methods are fast and efficient computationally but they require considerable effort and expertise in preprocessing depending on the number of different data levels. Interpretability of results depends on the clarity of the logical rules. Similarly to controlled methods, abstraction methods are limited by existing biological knowledge.

In reality, many if not most approaches are **hybrids**. For example, an analysis workflow will use controlled methods for feature extraction followed by black-box or controlled methods for building predictive models or stratifying samples. The

emergence of hybrid approaches illustrates (1) the underlying biological complexity that is tackled with data integration methodology, and (2) the practical challenges in applying standard statistical and machine learning methodology to biological problems.

3 Cancer

Cancer is a collection of complex diseases with one common characteristic: unwanted proliferation of cells also known as neoplastic growth [15]. In industrial nations cancer is one of the leading causes of death driven by the overall aging of the population. Improving the diagnosis, treatment and management of cancer is therefore important for public health and well-being.

Cancer causing changes in cells can be inherited in the germline or occur somatically during a person's lifetime. Many cancers have both a hereditary and a non-hereditary, sporadic form. Hereditary forms have an earlier age of onset than somatic forms of the same cancer. The focus of this dissertation is on non-hereditary, somatic cancers.

Non-hereditary cancers arise from random genomic alterations and environmental exposure. Whether random alterations or environmental factors are more important than the other is under debate [16]. Exposure to chemicals, substances or radiation (carcinogens) and viral infections are causally linked to carcinogenesis. For example, tobacco smoke is a carcinogen which causes lung and head-and-neck cancers [17] and infection with the human papillomavirus (HPV) causes cervical as well as head-and-neck cancer [18, 19]. Globally, lung cancer has the highest incidence of all cancers and tobacco induced lung cancers constitute a clear majority of lung cancers [20]. In contrast, cancers arising from viral infections constitute less than 10 % of cancers in developed nations but as much as 20% in the developing world [21]. Cancers have varied etiology but the incidence of cancer increases with age as both random and environmental caused DNA alterations accumulate.

Most if not all tumors exhibit genomic and chromosomal instability [22, 23]. Here I use the terms genomic and chromosomal instability interchangeably. Genomic instability induces random genomic alterations which activate cancer promoting genes (oncogenes) and inactivate tumor suppressor genes. In addition to genetic alterations, epigenetic changes in DNA and histone methylation are also postulated to drive carcinogenesis by similarly activating and inactivating cancer genes [24]. The genome in itself is little more than a blueprint and it is the proteins in the cell — produced based on the genomic blueprint via an intermediate RNA molecule — which are functional. It is clear that both genetic and epigenetic alterations lead to measurable changes in transcription and protein abundance.

Genomic instability, defective DNA repair, and faulty cell cycle control over recurrent cell divisions all cause DNA alterations that enable neoplastic growth. Cells in the body are constantly being replaced through the process of cell division. The frequency of cell division depends on the type of cell. Parts of DNA are

damaged in each cell division. Although cells actively repair damaged DNA, alterations accumulate over repeated cell divisions. Most alterations are silent meaning that they do not confer a detectable phenotypic change. Alterations which activate oncogenes or inactivate tumor suppressor genes lead to increased genomic and chromosomal instability and neoplastic growth. In addition, alterations, which disable genes responsible for maintaining genome integrity such as DNA repair and cell cycle control genes, accelerate the alteration rate.

This chapter describes the general forms of genomic and epigenomic alterations in cancer. Exceptions to these general characteristics exist but are beyond the scope of this dissertation. I describe how genomic and epigenomic alterations separately and together regulate gene expression and promote and maintain neoplastic growth. In addition, I describe relevant computational methods to integratively analyze these data. Finally, I briefly discuss these alterations in relation to four specific cancers — glioblastoma, diffuse large B-cell lymphoma, and breast and head-and-neck carcinomas — and shortly describe methods to measure gene copy number, DNA methylation and gene expression.

3.1 Back to basics: DNA, mRNA and proteins

A gene is expressed when its genetic sequence is transcribed into mRNA. In turn, the mRNA transcript of the gene is translated into a protein. Genes, by default, are present in two copies in the human genome. The number of copies shows genetic variability and genes can have additional copies. The number of copies of a gene (also known as gene dosage) directly influences the abundance of mRNA [25] which, in turn, correlates with the amount of protein in the cell. Since DNA, mRNA and protein are tightly linked, finding alterations in the genome implies alterations in mRNA and proteins. Thus, increases in gene dosage increases the abundance of corresponding mRNA and protein. When gene dosage increase affects the mRNA levels of the mRNA transcribed at the same locus, the effect is referred to as happening *in cis* [26]. When the alteration affects a locus somewhere else, it is *in trans*. This thesis focuses on alterations whose effects occur *in cis*.

This straightforward assumption of *in cis* effects has proved to be a powerful tool in discovering and characterizing cancer related alterations [8] eventhough the abundance of DNA and mRNA show only intermediate correlation [27] as do the levels of mRNA and protein [28]. Furthermore, directly comparing DNA-mRNA and mRNA-protein relationships simplifies how cells work since the direct comparison ignores post-transcriptional control (such as small non-coding RNAs) and post-translational modifications affecting protein expression [25]. Analysis of these *in trans* effects is beyond the scope of this dissertation.

3.2 Types of chromosomal alterations in cancer

Many cancers have an altered chromosomal structure where parts of or entire chromosomes have been lost or gained. Genomic alterations in cancer range in size from single nucleotide to entire chromosomes. Small, single nuclotide variants (SNVs or point mutations), small 1-20 basepairs (bp) long insertion deletions, and alterations (gains and deletions) which affect entire genes or chromosomes (somatic copy-number alterations or sCNA) all play a part in carcinogenesis [22]. Genomic alterations constitute the essential mechanism by which tumors activate oncogenes and deactive tumor suppressor genes [23].

At baseline, most tumors are diploid but the distribution is bimodal with additional peaks at 3.31 copies and more than five copies [29]. Large, chromosome arm or whole chromosome sCNA gains rarely exceed three copies [30]. In contrast, focal sCNA with a high number of copies encompass smaller regions and have a median size of 1.8 megabasepairs (Mbp) [30]. Arm-level sCNA are enriched for deletions whereas focal sCNA are slightly more often amplifications [29]. The length distribution of sCNA is strongly bi-modal where focal and approximately arm-level sCNA greatly outnumber intermediate length sCNA [30].

In loss of heterozygosity (LOH), one of two alleles present in the normal genome has been deleted. This can lead to reduced gene expression from one allele (allelic imbalance) and is a mechanism to partly deactivate tumor suppressor genes. LOH can retain one copy of an allele in which case it is a hemizygous deletion. Alternatively, the deleted copy can be replaced by the DNA repair machinery with one or more copies of the remaining allele. Two copy LOH events comprise more than 50% of all LOH loci in such disparate tumors as breast [31] and glioblastoma [32] incidating that they are more frequent and span larger genomic areas than copy-altered LOH loci.

A breakpoint is a locus where two regions of a chromosome, that would normally be contiguous, have been separated in a way that we can observe. Entire chromosomes can be shattered and then pieced back together without substantial loss of genomic material (chromothripsis) [33]. Contiguous regions of a chromosome can become detached and then reattach in an inverted orientation (inversions). The reattachment of breakpoints between loci that were not juxtaposed before the breakage is called a translocation. Sometimes inversions, translocations and deletions form chains of alterations called chromoplexy [34].

Cancers can be roughly divided into two groups: mutation dominated and sCNA dominated [35]. Mutation dominated cancers show preferential activation and inactivation of cancer genes via somatic point mutations rather than sCNA. The

reverse is true for sCNA dominated cancers. The difference, however, shows a gradient over cancers and some mutation dominated cancers nonetheless activate key oncogenes via sCNA.

3.3 DNA methylation at gene promoter regions regulates gene expression epigenetically

DNA methylation is an epigenetic form of gene expression regulation. Approximately 60% the genes in the genome have promoter regions with stretches of DNA in which cytosine to guanine alternations are enriched [36]. These regions are called **CpG islands**. When DNA is methylated, a methyl group is added to the end of a cytosine in the DNA sequence. DNA methylation functions in gene transcription regulation in CpG islands located at some kilobases up- and downstream of transcription start sites (TSSs) [36]. If a gene's CpG islands are saturated with methylation, transcription is silenced. If a gene's CpG islands are completely unmethylated, transcription is possible. Indeed, in normal somatic mammalian cells most TSS CpG islands are unmethylated [36]. When a gene exhibits more methylation than a reference, the gene is hypermethylated. Conversely, a gene exhibiting less methylation than a reference is hypomethylated.

Cancer cells show large, hypomethylated regions containing approximately one-third of TSSs and consistent hypermethylation of gene bodies as well as intergenic regions [24]. Hypomethylation leads to uncontrolled expression and is a potent way to activate oncogenes. In addition to oncogene activation, targeted hypermethylation of tumor suppressor gene promoter regions occurs [37] although whether this happens before or after these genes have been silenced is debated [38]. Nonetheless, DNA methylation is an important mechanism for tumors to develop drug resistance [39] and activate oncogenes [24].

3.4 Cancer genomics from an integrative perspective

A gene in part or completely responsible for the cancer phenotype is called a driver or cancer gene [40]. Cancer genes are activated and inactivated via genetic and epigenetic alterations. A gene that has genomic alterations in a cancer cell but does not participate in carcinogenesis or maintenance of cancer is called a passenger. These concepts are however fluid as a gene can be a driver at some stage of carcinogenesis and a passenger at another stage or vice versa [40].

The genomic analysis of a single sample yields lists of genes and genomic regions with gains, deletions, mutations, and other genomic alterations. A simple, non-integrative analysis approach is to repeat the genomic analysis in multiple samples

and count the frequency of each alteration. The logic here is that the more often we see an alteration the more likely it is related to cancer due positive selection of clones carrying the alteration [40]. For example, targeted sequencing of the *BRAF* kinase gene uncovered a high frequency of oncogenic V600E (change of valine to glutamic acid in codon 600) mutations [41]. Similarly, *EGFR* was discovered to be amplified and concomitantly overexpressed with a high frequency in glioblastoma brain tumors [42] and the amplification frequency exceeds 50% [43]. As was true for both *BRAF* and *EGFR* [44, 45], such initial discoveries in genomic studies need to be followed up by appropriate protein level and eventually in vivo studies to validate an association between the genomic alteration and phenotype.

Although successful, concentrating on high frequency alterations on one measurement level poses problems. First, a large proportion — on average one-third of the genome [30] — can be genomically altered in a sample which makes pinpointing of relevant cancer genes challenging because a region encompassing several genes is frequently altered. Data integration can tackle this problem using complementary data. Second, important alterations which affect a small subset of samples can be overlooked due to their low frequency. Again, data integration helps to overcome this challenge. For example, *MET* is amplified in 1% of head-and-neck squamous cell carcinoma (HNSCC) tumors [46] but its copy-number amplification and concomitant overexpression is nonetheless related to HNSCC growth in vitro and angiogenesis in vivo [47]. Third, only a subset of genes are expressed in any type of tissue, including different types of tumors, and this pattern varies from tissue to tissue.

Complementary integration can aid in tackling these challenges. For instance, a straightforward improvement on calculating the frequency of genomic alterations is to combine genomic and transcriptomic measurements. The logic here is to identify genes whose expression is altered *in cis* by the underlying sCNA. This approach was successfully used to characterize the oncogenic roles of *EGFR*, *MITF* and *SOX2*, to name a few [42, 48, 49].

Concomitant amplification and overexpression of genes is recognized as one of the central ways to activate oncogenes [50]. A census of amplified and overexpressed genes in cancer qualitatively identified 77 genes for which there was sufficient evidence to consider those genes as cancer genes due to complementary integration. For tumor suppressor genes, the TSGene database lists 716 human tumor suppressors, which have been collected and curated from literature [51]. The genes in TSGene have not, however, been originally characterized based on complementary genomic and expression changes.

Less than half of genes in sCNA regions show any correlation between mRNA

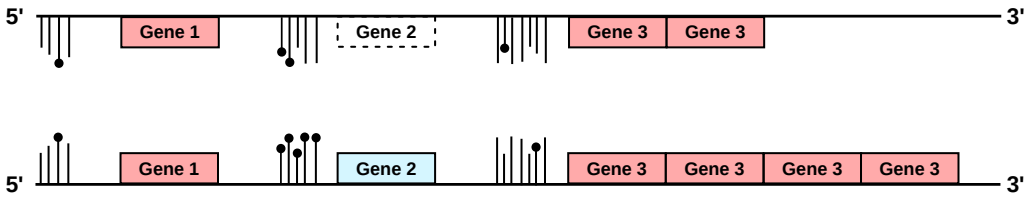

Figure 2: Two copies of a chromosomal region with promoter CpG islands (rods), methylated residues (black circles) and genes (rectangles). Red background denotes upregulated expression and blue background downregulated expression. Dashed border indicates a deleted gene. The depicted chromosomal region contains genes with varying copy-number normal, gain and deleted states. In addition, the promoter regions of the genes are varyingly methylated. Gene 1 is overexpressed due to promoter hypomethylation. For gene 2, one copy of the gene is deleted whereas the second, genomically intact copy is hypermethylated: both result in gene downexpression. For gene 3, both copies are hypomethylated and have gained additional copies resulting in overexpression.

and sCNA levels but the correlation increases with the number of copies so that highly amplified genes show better correlation [27]. A more functionally relevant alternative to integration of genomic and transcriptomic data would be to combine genomic copy-number or DNA methylation and proteomic information but this approach is severely limited by the number of proteins whose expression can be cost-efficiently quantified. For example, the reverse-phase protein array only probes approximately 170 proteins. Although the availability of large-scale protein data is limited, complementary integration can instead be enhanced by including more than two data levels. Figure 2 depicts how DNA methylation and structural genomic alterations jointly influence gene expression *in cis*.

Most solid tumors are a mixture of several cell types including cancer cells, fibroblasts, immune cells, and normal cells. Both integrative and non-integrative analyses of genomic data are affacted by this heterogeneity in patient tumor samples. The admixture creates challenges for accurate and robust identification of molecular alterations. In addition, the cellular microenvironment of the tumor is emerging as an important component for carcinogenesis [23]. Together these different cells form a complex ecosystem which both aids and fights the tumor. For example, different types of immune cells can be recruited by cancer cells to fend of other cells eliciting the host-immune response [52].

Since tumors are sampled as bulks of cells containing an admixture of cells, both computational analysis and interpretation of results become challenging. The type of cancer influences the amount of admixture and purity of sampling [53]. For example, in pseudomyxoma peritonei cancer cells make up only 5-30% of a tumor sample whereas many solid tumors can be sampled with near 100% cancer cell

yield [29]. In many cases, the signal of genomic alterations from the cancer cells is diluted and therefore harder to detect and, in addition, cells of the tumor microenvironment are also susceptible to genomic alterations [54] creating a source of false positives.

3.5 Existing data integration approaches in cancer genomics

The emergence of large-scale public data sets and databases has motivated and facilitated the development of computational tools [7]. The Cancer Genome Atlas (TCGA) is a large public database containing molecular and clinical data of 33 cancers and approximately 11,000 tumors [55]. The TCGA consortium was the first large-scale effort to systematically collect multilevel molecular and clinical data from large patient cohorts in several cancers. The systematized and harmonized sample collection as well as measurement protocols and technologies from multiple participating institutes provide high quality data which is important for integrative studies.

Whereas TCGA gathers data from multiple cancers, METABRIC is a unique compendium of molecular, microscopy cell imaging and clinical data of around 2,000 breast tumors specifically [26]. METABRIC data provide sufficient large sample numbers to enable breast cancer subtype specific analyses. In addition to TCGA and METABRIC, the NCBI Gene Expression Omnibus (GEO) contains a collection of microarray based genomic and transcriptomic data from more than 1 million samples [56]. The utility of GEO is hampered by less complete or missing clinical information and completely unharmonized sample collection practices in comparison to TCGA and METABRIC.

Two frequent applications for data integration in cancer genomics are gene prioritization for cancer gene discovery and sample classification for cancer subtyping. Numerous complementary, parallel and total integration methods have been developed to address these questions. The following paragraphs describe examples of each type of integration tool and their applications.

One of the most successful applications of complementary integration has been the characterization of new cancer subtypes. The METABRIC breast cancer is a prime example of complementary data integration in which copy-number and expression data were integrated to define ten novel breast cancer subtypes [26]. Similar integrated subtypes were also discovered in glioblastoma [57], colorectal [58] and prostate cancer [59].

Multiple complementary integration approaches have been used to discover putative cancer genes. In general, analyses of multilevel complementary data — even

when working on more than two measurement types — tend to look at pair-wise correlations between mRNA and other data levels. For example, integIRTy is a controlled algorithm that in one pass simultaneously determines whether gene expression is deregulated due to either DNA methylation or sCNA [60]. MIGHT is a visual inspection tool which in contrast to integIRTy can also find genes whose expression is synergistically deregulated by DNA methylation and sCNA [61]. Yang and colleagues focused on a subtype of ovarian cancer and were able to extract a core subtype-specific regulatory microRNA network by correlating mRNA, CNA, miRNA, methylation with a measure of gene differential expression [62]. Jörnsten and colleagues integrated sCNA to mRNA data to construct de novo signaling pathways which they used to identify genes related to glioblastoma survival [63]. Instead of a single gene approach, Tyekucheva and colleagues constructed integrated gene sets using multivariate logistic regression over two types of copy-number and two types of expression data and discovered gene sets related to glioblastoma [64]. Both of these glioblastoma studies were carried out with TCGA data.

Some abstraction based methods exist for driver gene discovery in complementary data [65] and an interesting approach from Ciriello and colleagues uses an abstraction based fusion integration to combine genomic alterations with networks [66]. In this approach, gene modules comprising several genes are prioritized so that alterations in different genes are mutually exclusive.

To date, the largest TCGA analysis used parallel integration to uncover pan-cancer cancer genes in approximate 5,000 tumors over 12 cancers [67]. Many of the integrative analyses from TCGA have been from combining expression data with networks (e.g., gene co-expression in four TCGA cancers [68], tools for handling the multilevel data with networks [69]). Parallel integration of METABRIC data from Yuan combined spatial data on the structure of tumor cells with changes in mRNA abundance [13]. In addition, the METABRIC breast cancer subtypes were later validated in a large parallel cohort [70].

HELIOS, a total integration approach, first combines complementary mutational, sCNA and expression data with a controlled Bayesian approach and then prioritizes genes based on parallel, RNA-interference screening data [71]. Osmanbeyogly and colleagues combine mRNA and RPPA protein data from TCGA breast cancer primary tumors to predict protein activity in breast cancer cell lines, which is used to group drug sensitivity profiles in these cell lines [72]. Furthermore, their model successfully predicted survival in parallel METABRIC primary tumor data.

Building survival models with single level genome-wide molecular data has not yielded substantially improved prognostic models compared to models incorporating clinical variables only [73, 74]. In addition, randomly chosen gene sets have a

90% chance of showing an association between survival and expression in breast cancer [75]. Interestingly, depending on the integration model, integrative analysis can yield better survival predictability power than single molecular data types alone [76]. The importance of survival analysis for result interpretation is shown by (1) several integrative analyses where prognostics models were built guided by the integration results; and (2) analyses where integration was used for subtype discovery [77, 26, 70].

3.6 Glioblastoma, diffuse large B-cell lymphoma, head-and-neck and breast carcinomas

I will discuss the alterations of four cancers in more detail: glioblastoma (GBM), head and neck squamous cell carcinoma (HNSCC), diffuse large B-cell lymphoma (DLBCL), and breast cancer (BRCA).

Most aggressive or stage IV glioma brain cancers are called **glioblastomas**. GBM is the most common brain cancer and its prognosis is abysmal due to the advanced stage at which it is most often diagnosed , the difficulty of surgical removal due to the anatomical location and the diffuse characteristic of the tumor cell spread in brain tissue [78]. In terms of genomic alterations, glioblastomas are heavily mutation dominated but exhibit highly frequent focal sCNA of key oncogenes (*EGFR*, over 50% of samples) and tumor suppressors (*CDKN2A*, over 60%). Molecular classification with genomic and gene expression data has suggested that glioblastoma has 2-3 subtypes [79, 80]. Clinically, glioblastoma has two subtypes: primary and de novo [78]. The primary subtype is characterized by a slow progression from lower grade gliomas to glioblastoma, whereas de novo glioblastomas present at stage IV [78].

Head and neck squamous cell carcinoma is the second most common cancer caused by tobacco smoke after lung cancer [81]. HNSCC is an epithelial cancer that shares both etiology and many genomic features of lung squamous cell carcinoma [82]. Resectable, local tumors have a good prognosis whereas non-resectable tumors show modest response to chemotherapy and targeted drugs [83]. In addition to sporadic and carcinogen induced tumors, infection of the human papilloma virus (HPV) can also induce HNSCC. The distinction between human papilloma virus positive and negative HNSCC is notable, and the two forms of HNSCC are considered separate entities [84]. Key sCNA in HNSCC with concomitant sCNA and expression changes include gains of chromomsome 11q13 (*CCND1, CTTN* and *FADD*) and — similarly to glioblastoma — *EGFR* and deletion of *CDKN2A* [85, 50]. HNSCC tumors show features of both mutation and sCNA dominance [35]. Expression profiling of HNSCC has yielded four molecular subtypes [86, 87, 46]

but currently only HPV status is used clinically to molecularly subtype HNSCC tumors [88].

Invasive **breast carcinomas** are sCNA dominated epithelial tumors [35]. Molecularly breast tumors are divided into five subtypes according to the PAM50 classification [77]: normal, basal, luminal A, luminal B and HER2 enriched. In the clinical setting in Finland, subtyping is carried out by IHC staining for the aestrogen receptor, progesteron receptor, Ki67 (proliferation index), and the human epidermal growth factor receptor 2 (HER2) coded by the *ERBB2* gene [89]. Breast cancer has 26 commonly amplified and overexpressed genes according to a recent review [50]. The most well characterized of these are *ERBB2* and *CCND1* which are commonly amplified in breast tumors [50]. Similarly to HNSCC, the *CCND1*/11q13 amplification region in breast tumors spans several putative cancer genes [90].

In contrast to glioblastoma, breast cancer and HNSCC, **diffuse large B-cell lymphoma** is a hematological malignancy originating from lymph nodes. It is the most common type of lymphoma [91]. DLBCL has three classical molecular subtypes defined by gene expression [92]: activated B-cell (ABC), germinal center B-cell (GCB), and the unclassified subtype [93]. These three subtypes show preferential patterns of somatic alterations but most alterations occur in all molecular subtypes making genomically distinguishing one subtype from another challenging [94]. The most frequently activated DLBCL oncogene, *BCL2*, is activated via a translocation with an immunoglobulin locus or a sCNA [92]. *REL* is another frequently copy-number amplified DLBCL oncogene which is amplified in up to 16% of tumors [95, 96]. *CDKN2A* is frequently deleted in DLBCL and shows preference to occur in the ABC subtype [92]. DLBCL tumors are characterized by a high number of somatic point mutations [95]. In addition to genome-wide mutation patterns, a DLBCL specific somatic hypermutation pattern occurs in immunoglobulin loci e.g., in chromosome 14q [92]. These suggest that DLBCL is a mutation dominated cancer but the occurance of site specific somatic hypermutation not seen in mutation dominated solid tumors can bias the classification.

Interestingly, several alterations such as amplifications in the epidermal growth factor receptor family members one (*EGFR*) and two (*ERBB2*) occur in multiple cancer types here - *EGFR* in breast, HNSCC and GBM and *ERBB2* in breast and HNSCC. In addition, *ERBB2* is amplified in ovarian carcinoma, which is molecularly closely related to breast cancer [97]. Furthermore, *CDKN2A* is frequently deleted in GBM, HNSCC and DLBCL but not in breast cancer [29, 92]. *ERBB2*, *CCND1* and *EGFR* are classic examples of cancer genes that show consistent and common concomitant expression and dosage alterations [50].

Although DLBCL shows consistent genome-wide chromosomal instability, it's dis-

tinctive mutation pattern due to somatic hypermutation distinguishes DLBCL from the other cancers described here. Akin to somatic hypermutation, breast cancers show localized hypermutation termed kateagis [98]. Though the hypermutation pattern in hematological malignancies appears to be attributable to the deregulation of the *AID* enzyme, these two mutation patterns seem related [99].

3.7 Measurement technologies

Microarrays as a measuring technology for DNA and RNA emerged in the 1980's and proliferated in the late 1990's [100]. The main applications of microarrays include gene expression profiling and genomic profiling. Microarrays have probes with a known DNA sequence which bind specific target molecules. The probes are specific to a genomic locus. Probes can be evenly spaced on the genome or only reside in certain regions such as genes. The location of probes depends on the type of array and its purpose. In this section, I introduce microarray types relevant for this work.

Array comparative genomic hybridization Array comparative genomic hybridization (aCGH or CGH microarray) is used to measure the genomic abundance of a sample. In aCGH, DNA from a test sample (e.g., tumor) and a control sample are labeled with fluorescent cyanine dyes Cy3 and Cy5, and co-hybridized onto an array using oligonucleotide probes whose length varies from array type to array type. The array is scanned to quantify signals from the two separate channels for computational analysis. Details of the computational analysis of aCGH data are discussed in Chapter 5.2.

CGH microarrays enable detecting sCNA amplifications and deletions. These chromosomal alterations are detected by first finding approximate breakpoints in the genome. Changes in the abundance of DNA on different sides of the breakpoint are then used to estimate genomic copy-number at this locus, and extrapolating this procedure over all chromosomes yields a copy-number profile for a sample. The accuracy of breakpoint detection depends on the density of probes on the array. Since probes are placed nearly uniformly along the genome (spacing is slightly denser near protein coding regions), the number of probes on the array determines the accuracy of breakpoint detection. For example, the Agilent 244A Oligo CGH microarrays have 236,381 probes with a median spacing of 8,900 basepairs.

CGH microarrays have a number of limitations when it comes to detecting genomic alterations. LOH, inversions, translocations or point mutations cannot be detected with aCGH. In addition to gains and deletions, SNP based arrays enable detecting LOH. Furthermore, whole-genome sequencing enables detecting all types of

structural alterations but their use is beyond the scope of this work.

Gene expression microarrays Expression microarrays quantify the amount of mRNA transcripts in a sample. The process is highly similar to aCGH except probes are designed to specifically capture mRNA as opposed to DNA. Furthermore, probes only target protein coding regions of the genome. Also, most expression arrays contain a single channel or only utilize one channel of a two-channel array.

Preprocessing of gene expression arrays is specific to the array type. Between sample normalization is necessary and is carried out, for example, with robust multi-array average (RMA) for AffyMetrix and LOWESS for Agilent arrays [101, 102]. Exon arrays require specialized normalization methods – such as MEAP [103] – to accurately quantify transcript level expression. MEAP can also summarize data on the gene, transcript or exon level.

Different manufacturers have differences in their probe design which necessitates specific preprocessing to be used for each platform. Using standardized pipelines to preprocessing of different array platforms, such as those found in workflow systems like Anduril [43], enable rapid and reproducable preprocessing of large numbers of varied arrays. Measurement data from different arrays have substantial differences in how they quantify gene expression (e.g., scale and distribution of measurement intensities) and mixing measurements produced with different array types in a single analysis should be avoided.

DNA methylation microarrays Illumina Methylation BeadArray's design has two probes per each locus. One probe measures the methylated signal and a second probe the unmethylated signal [104]. There are two sizes of the chip, the smaller covering 27,000 CpG loci in promoters of approximately 14,500 genes and the larger being "whole-genome" in the sense that there are over 485,000 CpG sites covered. The larger chip can, in addition to DNA methylation, be used to accurately analyze gene copy-number alterations similarly to SNP arrays [105]. The abundance of methylation at a CpG site is quantified as the ratio of methylated probe signal divided by the sum of signals from both unmethylated and methylated probes.

4 Aims of the study

My research was concentrated on optimizing, developing and applying data integration algorithms for computational analysis of cancer data. In particular, my research dealt with analyzing and integrating genomic, transcriptomic and DNA methylation data in conjunction with clinical and drug-target data to aid in interpretation of genomic data, find putative prognostic markers, and suggest priority drugs for potentially drug-sensitive patients.

The specific goals of my research were to

1. Compare algorithms for integrating genomic copy-number and gene expression data *(complementary integration, black-box, controlled and abstraction methods)*.
2. Integrate copy-number and expression data with clinical data to identify survival associated sCNA for development of a prognostic marker *(complementary integration, controlled method)*.
3. Develop a tool which improves complementary integration by simultaneously analyzing copy-number, DNA methylation and gene expression *(complementary integration, controlled method)*.
4. Develop a tool to prioritize drugs for potentially drug-sensitive patients through integration of existing local and open-access molecular, literature and signaling pathway data. *(parallel and total integration, controlled and abstraction methods)*.

5 Materials and methods

In this chapter, I will summarize the cancer data sets used in these studies and the computational and statistical methods with which these data were analyzed. Detailed descriptions of the cancer material and methods can be found in each publication. The following chapter first introduces the sample material and data sets. Then, I introduce the computational analysis of cancer copy-number data followed by the computational framework and database infrastructure which were used to build analyses. Finally, I will briefly summarize central statistical tools.

5.1 Biological sample material

Table 1 lists the cancer samples sets analyzed in each publication, the type of samples, and their origin. For mutational data, we used fully processed mutational calls as provided by TCGA. For other data types, we preprocessed the data ourselves (in data type specific ways) during the course of the analyses.

Publication	Material	Cancer	Data type
Publication I	Primary tumors	LUSC [82]	aCGH, gene expression
	Cell lines	HNSCC	aCGH, exon array
Related Publication I	Primary tumors	DLBCL	aCGH, exon array
Publication II	Primary tumors	GBM [106], OVCA [107]	aCGH, exon array, DNA methylation
Publication III	Primary tumors	BRCA [97], COAD [108]	SNP array, gene expression, mutation*, DNA methylation
	Primary tumors	GBM [106], OVCA [107]	aCGH, gene expression, mutation*, DNA methylation

Table 1: Data sets and material used in each publication. An asterisk following a data type indicates that preprocessed data were used. BRCA breast carcinoma; COAD colon adenocarcinoma; DLBCL diffuse large B-cell lymphoma; GBM glioblastoma multiforme; HNSCC head and neck squamous cell carcinoma; LUSC lung squamous cell carcinoma; OVCA ovarian adenocarcinoma.

5.2 Copy-number analysis of cancer genomes

Copy-number analysis is carried out in three steps: normalization, segmentation, and copy-number calling.

First, signal from probes is extracted from the array. The probe intensities are normalized to a mean of zero. Possible options for this normalization are platform dependent. For Agilent CGH arrays, LOWESS normalization has been shown to work well [102].

Normalized probe signals show substantial variance around genomic regional means and therefore are **segmented** to reduce noise [109]. In segmentation, a specific segmentation algorithm is employed to find loci where the regional mean of the probe signals significantly shifts. These loci are called breakpoints. After breakpoints are approximated by the segmentation algorithm, the normalized probe signal is supplanted with an averaged (segmented) signal for each region. Several segmentation algorithms have been developed and the most frequently used are circular binary segmentation (CBS) and GLAD [110, 111]. Both CBS and GLAD have been found to perform well in comparisons [112, 113].

Since segmentation reduces noise but does not tell whether a region's copy-number is altered, copy-number alterations need to be **called**. In sCNA calling, segmented regions are classified into gained (copy-number > 2), deleted (copy-number < 2) and normal (copy-number $= 2$). Sometimes high-level gains of five or more copies (amplifications) are distinguished from low-level gains for emphasis. Furthermore when measurement technology allows, two copy deletions and regions of copy-neutral LOH are separated from hemizygous deletions. The simplest calling procedure is to consider all regions with non-zero segment means as sCNA. More sophisticated methodology can take tumor purity and subclonality into account which need to be considered when studying clinical tumor samples. Impurity arises from (1) stromal admixture of non-cancerous cell types and (2) subclonal architecture of cancer cells [23]. Both cause the sCNA signal to be diluted and therefore make detecting low level sCNA more challenging. Tumor impurity can be handled in calling or preprocessing steps by focusing on high level sCNA. If required by the data, algorithms facilitating the detection of low level sCNA exist but are limited to certain types of measurement technologies [114, 115, 116].

Several sources of possible bias exist and need to be accounted for when analyzing sCNA. First, tumor samples contain an admixture of different cell types which lowers the amount of signal from cancer cells. Subclonality of the cancer cells confers the same effect as impurity and both can occur together. These were discussed in the preceding paragraph. Second, germline copy-number variation

(CNV) should be removed when studying somatic CNA [109]. In an ideal case, matched tumor and control samples are hybridized on different channels of the same array and CNVs would be automatically removed in a matched comparison. If matched controls are not available, gender-matched human reference DNA is used instead. Access to matched control data is rare in practice: the acquisition of normal reference samples can be costly, difficult, over-invasive to the patient (e.g., in brain cancer), or simply impossible to obtain (especially in retrospective studies). Therefore in studies lacking matched controls, known copy-number variable regions have to be dealt with [109], for example by removing those regions after normalization or segmentation.

5.3 Anduril

Anduril is an open-source component-based workflow development platform [43]. It includes the execution engine responsible for handling analysis execution and a simple but powerful programming language (AndurilScript) for building workflows. Anduril components are organized into bundles according to categories defined by the component's developer. Each component has a built-in support for automated testing and testing is rigorously and constantly carried out. Components can be used to execute several tasks ranging from basic filtering of tab-delimited numerical data and running statistical tests to automated querying of annotation databases such as Ensembl.

Anduril automatically parallelizes tasks within a single computing node (such as a laptop) and supports cloud-based multi-core parallelization which makes Anduril efficient for many bioinformatic tasks. Anduril is also designed to automatically parallelize execution of independent proportions of single workflows so that tasks which do not depend on each other are run simultaneously. Furthermore, Anduril increases development and analysis efficiency by automatically identifying the point of the workflow, where execution was halted, and only re-executing the portions of the analysis which were changed from the preceding workflow run.

5.4 Automatic workflow for processing and analyzing primary tumor data from TCGA

The TCGA pipeline developed with Anduril is an automatic retrieval, preprocessing, and analysis workflow for molecular primary tumor data from TCGA. The pipeline automatically analyzes gene expression, sCNA and DNA methylation data starting from raw data. This allows using different preprocessing and analysis tools than those chosen by TCGA. In addition, TCGA-processed ready-to-use mutation data

are downloaded and combined with other data levels for interpretation. The pipeline is modular and handles each data level separately. The pipeline compares tumors to controls to calculate the extent of differential expression for each gene. For sCNA and mutation data, the sCNA and mutation frequency for each gene is calculated. In addition, the pipeline calculates the univariate impact of gene expression, copy-number-alteration and DNA methylation on patient survival for each statistically significantly altered gene from each individual analysis using the log-rank test. The pipeline is scalable and allows users to implement additional data types and new cancer types. Currently, the pipeline supports BRCA, OVCA, GBM, COAD, kidney renal cell carcinoma and lung squamous cell carcinoma data. Here we focus on BRCA, OVCA, GBM and COAD.

The Cancer Genome Atlas contains two array types of gene expression microarrays (Agilent and AffyMetrix Exon array) for the four cancers. When both types of data are available for a cancer, only exon array data is used because it allows for more accurate quantification of gene expression than other available array types. This is because exon arrays probe the entire length of the gene and allow quantification of expression at the level of splice variants and individual exons [103]. The pipeline preprocesses Agilent expression data (BRCA, COAD) so that probes matching either multiple or no genes are removed and data are normalized to a mean of 0. For exon arrays (OVCA, GBM), data are normalized and gene expression values quantified with MEAP [103]. Post-normalization processing is carried out identically for both platforms. For each gene, the gene is considered up- or downregulated in a sample if the gene's expression is further than three standard deviations from the median of control samples and difference in the medians is statistically significant ($q \leq 0.001$, t-test, Benjamini-Yakutieli multiple hypothesis correction [117]). Samples are then grouped to upregulated, downregulated and unchanged groups, and univariate survival effect is assessed separately for each significantly altered gene.

Similarly to gene expression, copy-number data are available from two platforms: AffyMetrix 6.0 SNP arrays (BRCA, COAD) and Agilent CGH (GBM, OVCA). When both platforms are available for a cancer type, the pipeline uses Agilent data. AffyMetrix 6.0 SNP arrays are preprocessed with the R package crlmm [118]. Samples with signal-to-noise ratio of less than 5 and probes with confidence limit less than 0.9 are removed. Copy-number data from Agilent CGH arrays are preprocessed as described by Ovaska and colleagues [43]. Circular binary segmentation is used to segment data originating from both arrays [110]. After segmentation, copy-number calls are made in array type specific ways. For SNP arrays (BRCA, COAD), sCNA are called when the copy-number is further than 0.3 from the normalized logarithmic diploid state. For Agilent CGH arrays (GBM,

OVCA), sCNA are called as described in Publication II and Chapter 5.2. Gene specific copy-number calls are used to split samples into groups for survival analysis so that a sCNA group (deleted or amplified) is compared against a sample group with unaltered copy-number.

For DNA methylation data, the pipeline downloads beta values and transforms them into M-values [119]. This conversion transforms the measurement value distribution into a normal distribution and is performed to enable using the t-test to assess the statistical significance of changes in DNA methylation [119]. For each gene, if the genewise methylation difference between the median methylation of control samples and a tumor sample is more than 2 and statistically significant ($q < 0.05$, t-test, Benjamini-Hochberg correction [120]) then the sample is grouped into hypo- or hypermethylated sample groups. Similarly to survival analyses of copy-number and gene expression, a demethylated patient group is compared against an unaltered patient group in survival analysis for each significantly altered gene.

In addition to Publication II and Publication III, results from analyses with the TCGA pipeline have been utilized in Liu et al [121].

5.5 Moksiskaan

Moksiskaan is an integrated database connecting genes to signaling pathways and drugs [122]. Each gene is connected to other genes according to pathway information. The database has been constructed in a gene-centric way. Moksiskaan is a database of databases as it integrates signaling pathway information from five pathway databases. Pathway data have been extracted from KEGG [123], WikiPathways [124], PINA [125, 126], Gene Ontology (GO) [127] and PathwayCommons [128]. Drug data are extracted from KEGGDrug [129] and DrugBank [130].

In addition to pathway information linking genes to genes, Moksiskaan enables users to store gene-specific annotative data called studies. A study is a ranked list of genes. For each gene, annotative data include phenotype specific information such as mutation frequency in the COSMIC database and sCNA frequency in the Tumorscape database. In addition to these curated data sets, Moksiskaan contains similar annotative result data for selected TCGA cancers. For example, the amplification frequency of genes in selected TCGA data sets is stored in Moksiskaan [122]. Annotative data are specific to a Moksiskaan installation and more studies can be locally added by users. To aid in interpretation, each study has study specific cutoffs. For example, a gene is ranked in the COSMIC study only if its mutation frequency exceeds 10% and the mutation was recorded in at least 20 samples.

To illustrate the type of information that Moksiskaan contains, we take the *SLC25A32*

gene as an example. For *SLC25A32*, Moksiskaan contains annotations for 11 Gene Ontology terms, 10 studies, ten protein or gene regulatory connections to other genes, and one connection to an *SLC25A32* inhibiting drug. These numbers are larger for well known genes such *TP53* which has hundreds of regulatory connections.

5.6 Permutation test

A permutation test is a non-parametric statistical significance test. Permutation tests do not require assumptions on the background distribution of data and are therefore useful when the real distribution is unknown. Publication II and Publication III both utilize permutation testing for this reason.

In a permutation test, the significance of a statistic S is quantified by comparing the incidence of a statistic better than S when the test is repeated to randomly permuted sample of the original data. The maximum significance level of the test is determined by the number of permuted tests. For example if a test is repeated 1,000 times, the maximum significance level that can be reached is $\frac{1}{1000} = 0.001$.

5.7 Survival analysis

Many cancer studies on patient data contain clinical patient information. Of particular interest is often the time to an event such as death, relapse or metastasis. Analysis of these event data is called survival analysis [131].

Survival analysis is utilized to identify if a variant of interest — such as amplification of *ERBB2* — is predictive of patient survival in a cohort. In its simplest form, survival analysis is a comparison of two groups. The comparison assesses whether patients with the variant show significantly shorter or longer survival than patients without the variant.

Survival in a cohort is most frequently depicted using Kaplan-Meier curves. Kaplan-Meier curves are step functions which fall over time and approximate the true survival function of a cohort. The Kaplan-Meier survival estimate is the conditional probability of surviving beyond time t multiplied by the probability of survival at the previous time-point [132]. Importantly, Kaplan-Meier curves and estimates are able to handle censored data. Censoring refers to patients that drop out of a study during its course. For example, if the event in a study is defined as death to a certain type of cancer, censorings occur when patients are cured or die from a competing cause. Nonetheless, the patient has contributed to the cohort survival function and

the Kaplan-Meier estimate allows censored patients' contribution to be included in the analysis.

In univariate survival analysis, differences in Kaplan-Meier estimates of two or more groups are statistically tested. The de facto standard tests for this are the log-rank and Cox proportional hazards test [133]. As an alternative, univariate survival association can be tested with Cox survival regression [134].

In multivariate survival analysis, the de facto standard tool is the Cox regression (98% of sampled articles) [135]. The Cox regression model assumes that hazard at baseline is similar among all co-variates. This is called the proportional hazards assumption. Co-variates, whose hazards deviate from other co-variates according to the proportional hazards test, are either removed or, if categorical, stratified in the survival model [136].

6 Results

In this dissertation, I present four main results: a comparison of algorithms to integrate complementary data; how complementary integration enables finding a putative prognostic marker; an algorithm — CNAmet — integrating three levels of complementary data to improve result interpretation over methods which only use two data types; and how GOPredict — a total integration algorithm — enables prioritizing drugs for potentially sensitive subgroups of patients.

6.1 Copy-number and expression integration algorithms succeed or fail with their copy-number analysis

Identification of genes that contribute to the development, persistence and progression of cancers is one of the most important challenges for cancer research [137]. With the emergence of microarray and sequencing technologies, computational methods for data integration have become central for finding putative cancer genes [8]. One of the most successful approaches for discovering cancer genes has been the integration of genome-scale copy-number and gene expression data to find genes whose expression is driven *in cis* by an underlying sCNA [138, 139].

We wanted to characterize existing integration algorithms [140, 141, 142, 143, 144, 145, 146, 147, 148, 149]. We compared the performance of ten integration algorithms in six simulated data sets, 15 head and neck squamous cell carcinoma (HNSCC) cell lines and 129 lung squamous cell carcinoma (LUSC) primary tumors. The ten algorithms comprised methods from three method categories — black-box, controlled and hybrid methods (Table 2). Each algorithm was run using the default parameter values.

We developed a simulator framework which generated both copy-number and expression data where a varying number of genes were influenced *in cis* in varying degrees by underlying sCNA. The simulator uses a modified Willenbrock-Fridlyand-approach to model sCNA data [113]. The simulator generates baseline copy-number data from a Gaussian distribution. The variance of the distribution is constructed to simulate experimental and technological variance. The mean of the sampling distribution is generated by a function which models tumor cell admixture often found in cancer samples. Different types of sCNA are created to account for the varying sizes (broad, narrow) and magnitudes (low, medium, high) of sCNA.

In gene expression data generation, background expression values are sampled from a Gaussian distribution. To quantify sensitivity and specificity, we generated five types of ground-truth genes: three true positive and two true negative. True

Category	Algorithm	Simulation mean rank	HNSCC cell line rank	LUSC tumor data rank
Black-box	GSVD	9	5	5
	PCC	3	7	9
	pSimCCA	6	8	5
	SIM	8	1	1
	sPLS	4	8	8
Controlled	DLMM	7	8	9
	edira	5	2	4
	intCNGEan	2	5	2
	SODEGIR	10	3	5
Hybrid	S2N	1	4	3

Table 2: List of algorithms in the comparison sorted alphabetically by category and algorithm name. The simulation rank is according to the mean over six data sets. In case of ties, tied algorithms receive the same rank.

positive genes were generated to be overexpressed in 100%, 75% or 50% of sCNA samples. True negative genes were underexpressed or overexpressed in every sample irrespective of the underlying sCNA status. We generated three types of simulated data sets with different models of interdependence between sCNA and expression. The models were linear, stepwise and sigmoidal.

We compared algorithms in simulated data using sensitivity and specificity. Here, sensitivity is the ratio of correctly identified true positive genes and all positive genes. Similarly, specificity is the ratio of correctly identified true negative genes and all negative genes. To investigate the influence of sample size, we created data sets of 15 and 100 samples. Most algorithms showed higher sensitivity and lower specificity when sample size increased indicating a higher number of false positive calls with the pre-selected significance threshold. All algorithms except two performed better with a bigger sample. For the two algorithms whose performance decreased, the decrease was due to decreased specificity.

None of the integration algorithms explicitly considered the model of interdependence between copy-number and expression. Nonetheless, our results indicated that all algorithms except two handled linear, stepwise and sigmoidal dependence models with similar sensitivity and specificity.

In cell line and tumor data, we compared algorithms' performance to an expert-curated ground truth list of 30 genes which we collected from literature. Genes were chosen based on previous evidence of copy-number induced expression changes in HNSCC. The results were similar to simulations. We chose two genes for validation

with qPCR because the two genes were predicted by five algorithms and had not previously been associated with copy-number changes. Using qPCR, we validated these genes in vitro to show concordant changes in copy-number and expression levels in the same cell lines.

Due to etiology and the squamous cell type, LUSC is closely related and genomically similar to HNSCC and we therefore used the same 30 genes in our LUSC comparison. Compared to the cell line data, sensitivity increased for four and decreased for five algorithms in the substantially bigger but more noisy LUSC data set. Of note, the SIM algorithm was the most sensitive in HNSCC and LUSC data but predicted over 13,000 genes to have a significant association in HNSCC and therefore contains a large number of likely false positives.

Controlled and hybrid methods which utilized segmentation performed the best. Interestingly, the three top methods in simulated data all utilized a segmentation algorithm indicating that accurate definition and calling of sCNA regions is the essential step for integration of copy-number data. Black-box methods were consistently outperformed by controlled and hybrid methods suggesting that the choice of preprocessing tools and their optimization improve performance. Only three algorithms (DLMM, GSVD and SIM) had originally been tested on both primary tumor and simulated data. Interestingly, these three showed consistently poor performance which suggests that the simulated data the algorithms were tested on inaccurately modeled tumor data.

Our comparison did not include algorithms which only utilize SNP arrays for sCNA analysis. Two algorithms did not produce results in cell line or primary tumor data and could therefore not be included in that portion of the comparison. More thorough testing by original developers might have removed this problem. Testing is essential for software development and the lack of testing in scientific software is an enduring problem as several algorithm comparisons have reported instances of faulty software [150, 151].

To summarize, hybrid and controlled methods perform best. The most essential part of the analysis is copy-number segmentation. Our in vitro validation suggests that complementary data integration algorithms integrating copy-number and expression data are able to infer putative cancer associated genes.

6.2 Complementary data integration enables finding a putative prognostic factor in lymphoma

Simulation data provides a reproducible framework for testing integration algorithms and are completely free of noise originating from large contiguous

sCNA blocks and admixed samples. Analyses of primary tumor samples on the other hand often have to deal with sCNA blocks and varying tumor cell subpopulations with different sCNA frequencies when identifying putative cancer genes. Complementary integration approaches provide an efficient way to enhance cancer gene characterization when complementary data are available.

Diffuse large B-cell lymphomas (DLBCL) are hematological malignancies that carry several frequent sCNA but effects of sCNA on DLBCL transcriptome have been reported scantly [96]. We analyzed copy-number profiles of 51 DLBCL primary tumors. These samples comprised both ABC and GCB subtypes. The samples originated from a prospective Nordic lymphoma trial and all patients had undergone a standardized treatment protocol. Complementary copy-number and expression data were available for 38 samples for which we analyzed the impact of sCNA on gene expression.

We found frequent amplifications (occurring in five samples or more) in chromosomes 1q (two hot spots), 2p15-16, 14q, 18q and 20q as well as a deletion in 9p21.3. Gene dosage significantly altered the expression of 29 amplified and 2 deleted genes (Benjamini-Hochberg false discovery rate $q < 0.05$). The amplified and overexpressed genes included known DLBCL oncogenes *BCL2* and *REL* as well as previously uncharacterized *in cis* altered genes. The two deleted and underexpressed genes, *CDKN2A* and *MTAP*, both reside on 9p21.3 and due to their physical proximity are frequently co-deleted in multiple different cancers [29].

To further characterize the most important sCNA loci, we analyzed the survival association of loci that showed concomitant genomic and transcriptomic alterations. For survival analysis, we used the complete set of 51 samples for which copy-number data existed and compared survival between copy-number altered and copy-number normal patients. We tested both progression free and lymphoma-specific overall survival using the log-rank test. In lymphoma-specific overall survival, the cause of death of the patient has been recorded to be lymphoma. As patients can expire due to competing causes during the study period, focusing on lymphoma-specific events enables a more accurate analysis of survival association.

Somatic CNA in chromosomes 2p15 and 18q12 were associated with progression free survival and the sCNA in 18q12 was furthermore associated with decrease in lymphoma-specific overall survival. The amplification region in chromosome 2p15 contained five genes. Of these five genes, we validated the association between gene dosage and expression for *XPO1* and *COMMD1* with qPCR. Furthermore, the qPCR expression of *COMMD1* was also significantly associated with survival.

To further corroborate our expression level findings, we quantified protein levels of COMMD1 from IHC images in two tissue microarray sample sets. The first set

of 70 samples comprised additional samples from the same lymphoma trial and the second set was an independent validation cohort of 146 samples. As expected, COMMD1 protein levels were associated with progression free survival in both data sets. Incorporating these parallel data strengthens our results.

To summarize, complementary integration of copy-number and gene expression data enabled us to find genes whose expression was altered by sCNA. We showed a survival association of one of these genes, *COMMD1*, in array based data, qPCR expression data, and protein data from IHC tissue microarrays in two cohorts. COMMD1 is protein which has been shown to have both cancer promoting and opposing functions [152, 153]. This is a known phenomenon and several genes, including *TP53*, have been shown to context dependently either promote or suppress tumorigenesis [154]. Our results suggest that COMMD1 activity is higher in more aggressive DLBCLs.

6.3 Integration of additional levels of complementary data enhances driver gene characterization

Integration of copy-number and gene expression data shows that approximately half of the genes in sCNA regions respond to sCNA with a change in expression levels [27]. This low number is not surprising, however, considering that several processes regulate gene expression in addition to copy-numbers. For example, DNA methylation of promoter regions functions *in cis* almost identically to copy-numbers. We therefore wanted to investigate whether adding DNA methylation data to the integration could enhance predicting expression levels of genes and thereby aid in interpretation.

To integrate copy-number, DNA methylation and gene expression data, we developed the CNAmet algorithm. CNAmet extends a previous controlled approach for integration [143]. CNAmet combines categorization of explanatory variables (DNA methylation, copy-number) with gene expression using a signal-to-noise ratio test (Equation 1). The test quantifies the magnitude of the association between the explanatory and dependent variables while accounting for the variance in the expression values:

$$W = \frac{\mu_1 - \mu_0}{\sigma_1 + \sigma_0}, \quad \sigma_1 > 0, \quad \sigma_0 > 0 \tag{1}$$

where μ_1 is the mean and σ_1 the standard deviation of expression values of samples with an alteration. For samples without an alteration μ_0 and σ_0 are calculated similarly. The intermediate test statistics W are called weights in

CNAmet. A separate weight is calculated for methylation-expression and copy-number-expression complements.

In the second step, the CNAmet *S*-statistic is the sum of the two constituent weights multiplied by the percentage ε of samples sharing both methylation and copy-number alterations:

$$S = (W_M + W_C)\varepsilon, \quad W_M > 0, \quad W_C > 0 \tag{2}$$

where W_M is the weight for the methylation-expression and W_C the weight for the copy-number-expression complement. The ε term favors genes which are simultaneously demethylated and copy-number altered and therefore are more likely to exhibit synergistic effects of methylation and sCNA. The *S*-statistic is calculated separately for deleted and hypermethylated and amplified and hypomethylated cases.

We tested CNAmet with 181 samples of the TCGA glioblastoma brain cancer cohort for which all three data levels were available. Top significant genes according to CNAmet included the epidermal growth factor receptor (*EGFR*) as well as other known oncogenes. *EGFR* is a well known glioblastoma oncogene and amplified in 40% to 50% of glioblastoma tumors [43, 155]. Our results indicate that amplified and hypomethylated samples exhibit significantly higher *EGFR* expression than samples which are only amplified (t-test $P = 3.8 \times 10^{-8}$). In addition, we analyzed 188 TCGA OVCA samples and similarly found synergistically altered genes in these parallel data.

The utility of *EGFR* copy-number amplification as a prognostic marker in glioblastoma is debated [156]. We compared the survival between hypomethylated and amplified *EGFR* (group 1) and hypomethylated *EGFR* (group 2) but not amplified patients. The difference was small but significant (log-rank test $P = 0.0584$). However, the small number of patients in the comparison (group one $n = 15$, group two $n = 18$) decreased the reliability of the prediction.

Our main conclusion was that aberrant DNA methylation and sCNA work synergistically to deregulate gene expression in glioblastoma and ovarian cancer. Our computational findings in these two cancers have gained additional support in breast cancer from Aure and colleagues who similarly integrated copy-number and expression data for microRNA in two large, independent cohorts [157]. They activated candidate microRNAs — microRNAs which showed expression changes due to sCNA, DNA methylation alterations or both — from their integrated analysis in vitro. The results showed that candidate microRNAs were connected to cell proliferation, cell viability and apoptosis. In conclusion, complementary integration

of these three measurements can be used to find putative prognostic markers and characterize cancer related genes.

6.4 Total integration improves drug prioritization and patient stratification for treatment

Complementary data integration can be employed to find and characterize putative cancer genes and it can therefore be used to prioritize genes for further computational analysis. On the other hand, parallel data integration enables further characterization of candidate genes in independent data and can strengthen results when used to validate initial results as in Related Publication I. In Publication III, we combined complementary and parallel integration and developed a total integration approach for molecular data driven drug prioritization.

The molecular landscape of tumors determines the efficacy of drug treatments [158]. Patient stratification with molecular data allows identification of patients who are likely to respond to targeted therapy [159]. Since molecular alterations occur in multiple cancer types, molecular alterations can be used to guide drug repositioning from one cancer to another [160]. Furthermore, prioritizing drugs and drug targets enhances the efficacy of drugs and maximizes the number of responding patients [159].

Improving stratification and drug repositioning with multiple types of molecular data requires computational data integration methods [8]. We identified molecular results data in public databases — such as COSMIC [161] and Tumorscape [30] — and open-access cancer genomic studies — such as the Cancer Gene Census [162] — as a large untapped parallel data resource which could be utilized by integrative methods. To integrate several levels of molecular data with drug-target information, signaling pathways and existing public result databases required employing total integration.

Our total integration approach for drug prioritization has two main building blocks: (1) a knowledge-base to store data from these varied sources and (2) an algorithm — GOPredict — to mine the knowledge-base (Figure 3). GOPredict prioritizes drugs and stratifies patients for the drugs and thereby facilitates drug repositioning.

We constructed the knowledge-base as an instance of Moksiskaan. The knowledge-base contained analysis results from five curated public databases (Table 3), in-house analysis results from four cancers (TCGA BRCA, COAD, OVCA and GBM) as well as drug-target information from KEGGDrug and DrugBank databases [129, 130]. The in-house data had been processed with the TCGA pipeline and comprised gene expression, sCNA, mutation and DNA methylation data. Both curated and in-house

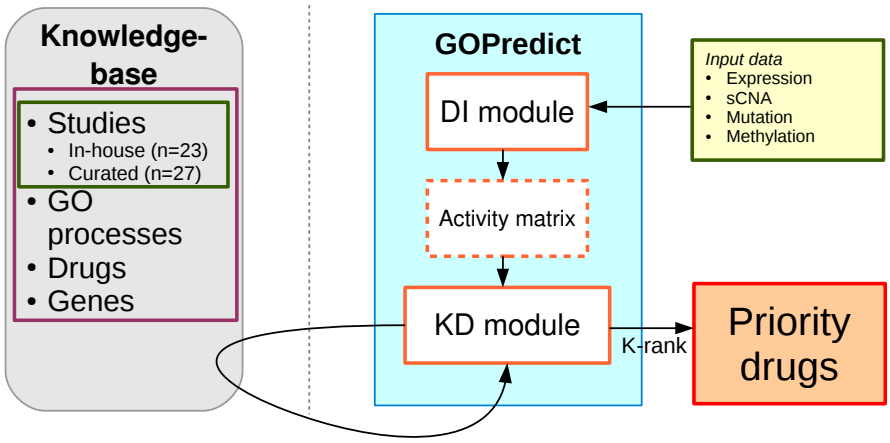

Figure 3: Overview of the GOPredict total integration approach. GOPredict's data integration (DI) module is used to create an activity matrix from multilevel input data (yellow box). The knowledge discovery (KD) module utilizes information stored in the knowledge-base (gray box) to prioritize drugs and stratify patients. Complementary data is denoted with a green border. Parallel data is denoted with a purple border. Studies contain both complementary and parallel data.

data were stored as annotative data (studies). Each gene is ranked for each study according to study specific rules (full details of rules are given in Publication III). For example, *EGFR* is the most frequently amplified gene in TCGA glioblastomas and therefore is ranked first in the *TCGA, copy-number alteration*-study. A full list of studies and data sets used to construct these studies is in Table 3.

Gene Ontology biological process (GO processes) are connected to genes to define the pathway context for each gene. The Gene Ontology is a flexible data source for gene signaling pathways because (1) it utilizes a stable naming system and (2) contains both general processes (e.g., 'cell proliferation') and specific processes (e.g., 'fibroblast growth factor receptor signaling pathway') which are in contrast to other pathway databases [163, 164].

The GOPredict algorithm has two modules: the knowledge discovery module and the data integration module (Figure 3). The knowledge-discovery module comprises the data mining algorithm which prioritizes drugs for input data sets. The data integration module preprocesses complementary input data.

The full formal description of the knowledge discovery module is in Publication III. Briefly, the knowledge discovery module works in four steps. In the first step, ranks are fetched from the knowledge-base and normalized to calculate for each gene its K-rank. In the second step, genes are connected to GO processes. GO processes are ranked by summing up K-ranks of genes regulating a GO process. Statistical significance is calculated by a permutation test so that K-ranks for the same number of genes are randomly sampled from the knowledge-base and the GO process rank

recalculated. K-ranks of genes are recalibrated in the third step based on the ranks of GO process, that the gene regulates, from step two. In the fourth step, drugs are prioritized based on the input activity matrix and the recalibrated K-ranks from step three.

Input data for GOPredict need to be preprocessed. GOPredict contains an abstraction based data integration module which is used to fuse multiple levels of molecular input data using biologically motivated rules. The minimum requirement is one level of either gene expression or copy number data. For each gene, the gene's activity status in the activity matrix is defined by (1) the gene's expression state if the gene is genomically normal (no mutation or sCNA) or (2) otherwise by the gene's genomic alteration status. This integration step results in a gene-by-sample gene activity matrix.

GOPredict and the knowledge-base constitute a total integration approach. GOPredict is computationally efficient and scales to input data sets with thousands of samples. The knowledge-base comprises parallel data from multiple cancers and cancer cohorts whereas the activity matrix is constructed from complementary data.

GOPredict prioritizes FGFR inhibitors for breast carcinoma and CDK inhibitors for ovarian carcinoma

We prioritized drugs with GOPredict in 497 breast (BRCA) and 390 ovarian carcinoma (OVCA) samples from the Cancer Genome Atlas. We utilized GOPredict's abstraction based data integration module to fuse gene expression, point mutation and sCNA data from these input data sets into two activity matrices — one for each cohort.

We first tested GOPredict's ability to prioritize subtype specific drugs. Accordingly, we analyzed all BRCA samples with a immunohistochemically verified *ERBB2* amplification according to TCGA provided clinical information. This represents the HER2 activated breast cancer subtype according to the PAM50 breast cancer intrinsic molecular subtyping [77]. As expected, HER2 inhibitors were among the top priority drugs with 4 inhibitors in the top 10 drugs.

We next analyzed the entire TCGA BRCA data set. In the entire data set, five of ten top prioritized drugs were multi-kinase inhibitors. Notably, the five inhibitors target members of the fibroblast growth factor receptor (FGFR) family. The top four inhibitors were prioritized nearly exclusively based on the activity of *FGFR3* (97% of sensitive samples for dovitinib, lenvatinib and ponatinib). Over one-third of TCGA BRCA samples ($35-42\%$) were potentially amenable to treatment with these FGFR inhibitors.

Type	Cancer	Studies	Number of studies
In-house	BRCA	somatic CNA (amp, del, survival) methylation (survival) expression (survival, fold-change)	6
	COAD	somatic CNA (amp, del) methylation (survival) expression (survival, fold-change)	5
	GBM	somatic CNA (amp, del, survival) methylation (survival) expression (survival, fold-change)	6
	OVCA	somatic CNA (amp, del, survival) methylation (survival) expression (survival, fold-change)	6
Curated	Multiple	Amplified and overexpressed cancer genes	1
	Breast	Brain metastasis genes	1
	Multiple	Cancer Gene Census (act, inact)	2
	Multiple	COSMIC (primary, recurrent, metastasis)	3
	Multiple	Tumorscape (amp, del)	20

Table 3: List of studies stored in the database ordered by source type. Study descriptions in parenthesis indicate the type of data stored: amp=amplification frequency, del=deletion frequency, survival=log-rank P-value, fold-change=continuous expression difference between medians of tumor and control samples, act=activating mutations, inact=inactivating mutations. For COSMIC, the type denotes the mutation frequencies in primary samples, recurrent samples or metastatic samples.

In addition to breast cancer, we analyzed a parallel ovarian cancer cohort from TCGA. In OVCA, seven of ten top drugs were CDK inhibitors. Interestingly, 91% of patients were potentially sensitive to the highest ranked CDK inhibitor, dinaciclib, which sensitizes ovarian cancer cells to chemotherapy. This suggests that a high number of OVCA patients could benefit from treatments combining chemotherapy with CDK inhibitors [165].

Sample stratification showed breast cancer subtype specific patterns of FGFR3 inhibitor sensitivity. We analyzed the impact of possible confounding alterations in selected BRCA cancer genes and found them unrelated to *FGFR3* status. FGFR3 inhibitor sensitive samples were enriched for luminal A and B samples. In addition, luminal B samples were substantially more prominent than luminal A among the sensitive samples.

To test GOPredict's predictions in vitro, we tested the efficacy of four drugs (three FGFR3 inhibitors and one proteosome inhibitor) in a panel of breast cancer cell lines with varying levels of FGFR3 protein. The in vitro results were concordant

with our prediction: three out of four FGFR3 expressing cell lines were sensitive to at least one targeted FGFR3 inhibitor whereas non-expressing cell lines showed no sensitivity. In line with our stratification, sensitive cell lines were all of the luminal and insensitive of the basal subtype. The single FGFR3 expressing cell line, which did not show any FGFR inhibitor sensitivity, harbors a KRAS mutations which is known to confer resistance to multiple targeted drugs in cell lines, xenografts and patient samples. Since sensitive cell lines carried a wild type KRAS, the KRAS mutation can explain the lack of sensitivity for FGFR inhibitors in one cell line.

Finally, to test if the K-rank can be used to find putative prognostic markers, we selected genes which received high K-ranks in both ovarian and breast cancer studies. Out of three genes which fulfilled this criterion, the expression status of *SLC25A32* was one of two significant independent prognostic markers in TCGA ovarian cancer in multivariate Cox regression ($P = 0.003$). Other co-variates in the model were FIGO stage, tumor grade and post-operative residual tumor size, of which tumor size was the second significant predictor ($P = 0.02$). This survival result suggests that parallel data integration with K-ranks can be utilized to find putative cancer genes. Interestingly, SLC25A32 is a folate transporter and it is expressed in ovarian cancer tissue. This in combination with our multivariate model suggests that antifolate drugs could be useful in treatment of ovarian cancers.

To summarize, total integration enables combining information from local genomic data with signaling pathways and public genomic analysis results, facilitates drug repositioning, and aids in finding putative prognostic markers.

7 Discussion

Cancer is a complex group of diseases characterized by neoplastic growth of cells in the body. Deciphering the molecular changes that drive tumorigenic growth requires analyzing multiple different types of measurement data. These data are increasingly produced with technologies that efficiently and reliably generate gigabytes of raw data from thousands of samples. The management, processing, analysis and interpretation of these data are facilitated by computational data integration.

Utilizing data integration offers a means to accurately interpret and draw inference from measurement data. I have in this book outlined my theses for the three different forms of data integration: complementary, parallel and total. In complementary integration, dependent measurements of different aspects of the same cohort and genes are integrated. In parallel integration, the same aspects of independent samples or independent aspects of the same samples are integrated. In addition, total integration refers to combining complementary and parallel approaches. In this dissertation, I have shown examples of how complementary, parallel and total approaches can be utilized to infer putative prognostic cancer markers, improve interpretation of multilevel genome-wide data, and stratify patients for efficient drug therapies.

In the first part, we built a simulator to compare complementary integration algorithms that integrate sCNA and gene expression data. Our results indicate that good performance was a result of appropriate processing of copy-number data. Of note, in a later comparison from Lahti and colleagues (comparing almost the same set of algorithms), the relative performance of the algorithms was different but the authors agree on the importance of sCNA analysis [166]. Interestingly, the comparison used a different set of test data and showed that each algorithm worked best on the data the algorithms were tested with in original publications. Results from these two parallel comparisons illustrate the tendency of algorithms to be optimized to suit certain data and brings to light the need for independent, community-based ground-truth data sets including both molecular and simulated data, which notably is one of the strengths of the second comparison from Lahti and colleagues. Indeed, DREAM challenges have used such a crowdsourcing approach [167]. In the challenges, participating algorithms compete for accuracy in iterative rounds where each participant can improve their performance. The DREAM challenges so far have sought to improve and build prediction and classification algorithms tackling diverse biomedical problems such as disease progression [168], drug sensitivity [169] and gene network inference [170]. Increased community-based development offers exciting opportunities for improving the development process for scientific algorithms.

The two methodological contributions in this dissertation provide tools for interpretation of multilevel molecular data. The first, CNAmet, is an algorithm for integrating complementary sCNA, DNA methylation and gene expression data to find genes whose expression is synergistically altered by genomic and epigenomic changes. Of note, similar algorithmic methods integrating more than two data types focus either on patient subtyping [57] or on automatically detecting concomitant but non-synergistically altered genes [60, 157]. Interestingly, a chemokine coding gene cluster in chromosome 4q21 was shown to be synergistically upregulated by DNA methylation and sCNA in two small esophageal carcinoma primary tumor cohorts using microarrays, fluorescence in situ hybridization, and qPCR [61]. Reports of these synergistic effects are still limited and more work is needed to characterize whether simultaneous hypomethylation and copy-number alteration is a driver event for oncogene activation.

The second contribution, GOPredict, is a total integration method which combines a knowledge-base and a machine learning algorithm to suggest drug sensitive subgroups of patients. GOPredict could be improved in the future in three ways: (1) including binding affinity of drug-target pairs to weight each drug target gene and; (2) including data on drug combinations and synthetic lethal interactions. Data on both of these are currently scattered and beyond automated retrieval. Interestingly, a second GOPredict-like "in silico prescription" approach, CDAD, also suggested that a high proportion of patients are potentially amenable to treatment with compounds already approved for a cancer [171]. The emergence of approaches such as GOPredict and CDAD suggests that efforts to retarget drugs from one cancer to another will increase in the near future. Although promising, results of in silico prescription algorithms will not transition to the clinic directly but serve to guide future drug experiments and experimental design.

We are already in the era of total integration in biomedicine as platforms and systems enabling and embracing automated total integration are emerging. Workflow systems such as Anduril enable computational biologists to implement programmatic interfaces and rapidly access a constantly growing ecosystem of open data- and knowledge-bases. Unlike GOPredict, databases such as intOGen and cBioPortal currently provide user-friendly graphical interfaces which are used from the web browser [172, 65]. These interfaces enable lightweight statistical analysis and data integration online for researchers lacking in programming skills. As computing moves back to large server farms (called the cloud), the number of databases and their computing capability will sharply increase soon and further drive the usage of statistical and computational tools as well as data integration by opening these possibilities to more scientists.

All methods presented here were limited to finding, characterizing and quantifying

in cis effects. For example, Curtis and colleagues looked at *in trans* effects of sCNA on gene expression and found on multiple chromosomes several trans-acting sCNA hotspots each of which was associated with expression changes of more than thirty genes [26]. Furthermore. results presented in this thesis are based on analysis of microarray measurement data which are currently being replaced by sequencing based technologies. With the advent whole-genome and whole exome (referring to the entire coding portion of the genome) sequencing, copy-number analysis pipelines and data integration methods have to be redesigned to handle these novel kinds of data [150].

In Related Publication I, Publication II and Publication III, highly ranked genes from integrative analysis showed significant prognostic impact. In all instances, survival analysis was carried out by testing a single molecular marker at a time, both in univariate and multivariate analysis. Evidence is mounting, however, that specific patterns of molecular alterations (and not single alterations alone even when they are in strongly cancer associated genes) are driving carcinogenesis — which is as expected. For example, in DLBCL double-mutant patients with a chromosomal alteration in both *BCL2* and *MYC* show the worst prognosis whereas patients with mutant *MYC* but normal *BCL2* fare best [173]. Some approaches have already tackled multi-gene combinatorial survival analysis for germline variants [174, 175], and in the near future similar algorithmic approaches will likely be utilized to analyze somatic tumor data.

Cancer genomic approaches have their limits. The last decade has witnessed the emergence of a new understanding regarding the role of tumor microenvironment in cancer [23]. Tumors are increasingly recognized as functional entities comprising a multitude of different cells in addition to cancer cells. In Related Publication I, we validated our complementary integration finding via cell imaging. Image analysis computationally from large cohorts opens up possibilities to integrate microscopy imaging data with molecular data to characterize cellular admixture in tumors. Since microscopy images allow for automated identification of different cell types [12], these data could be utilized to decipher the interplay between cancer cells and the stromal structure of a tumor. Efforts to characterize the cellular structure and interactions occurring in the tumor microenvironment and their connection to molecular alterations in cancer cells is an important and central research question in cancer for the next decade. Describing and explaining the role of the microenvironment will require data integration.

Acknowledgements

This work was carried out in the Systems Biology Laboratory at the Faculty of Medicine, University of Helsinki during 2008-2015. I thank Prof Sampsa Hautaniemi for leading the laboratory in a jovial but demanding manner. Without your commitment, guidance and dedication I could not have achieved the way I did.

I want to acknowledge the following bodies for funding my work: Helsinki Biomedical Graduate Progam (currently known as the Doctoral Program in Biomedicine), Biomedicum Helsinki Foundation, Ida Montinin Säätiö, and Emil Aaltonen Foundation.

I am grateful to my thesis committee members Dr Jari Haukka and Prof Olli Kallioniemi for their encouragement and advice during our annual meetings.

I thank the reviewers of this thesis Dr Merja Heinäniemi and Dr Päivi Onkamo. Your comments improved the thesis and its presentation considerably.

When first starting in the lab, I was thoroughly humbled by the intelligence of my colleagues — and still am. Marko Laakso was and continues to be my mentor in all things computational and a co-conspirator in research projects too numerous to list. Ville Rantanen, in addition to being cooler than the Dude, has taught me all I know about unix and is always fun to talk to on topics ranging from machine learning or the English language to building a dry-wall. Our lab is a close-knit environment with tight interpersonal relations and an unreserved, conversational atmosphere (that is, the room is small and we like it that way). My sincere thanks to the friendly and helpful members of our lab (past and present): Alejandra, Amjad, Anna-Maria, Chengyu, Chiara, Emilia, Gabriele, Javier, Julia, Katherine, Kristian, Lauri, Lilli, Markku, Mikko, Ping, Rainer, Sirkku, Tiia, Viljami and Vladimir. Special thanks go to Rainer, Amjad and Ping for reading and commenting on my dissertation, and to Vladimir for creating the TCGA infrastructure which most of my work relies on. Finally, I want to thank everyone who participated in the Significo experiment, especially Sirkku, Marko and Kristian. I know we all learned a lot.

Computational biology is a discipline which cannot thrive without the work of molecular biologists and physicians. My sincere thanks to Tanya Lepikhova and Outi Monni for their big contributions to the Nature Methods paper and many other projects; Sirpa Leppä, Minna Taskinen and other members of the Leppä Group for the ever-expanding lymphoma project; Denis Belitskin and Juha Klefström for their big effort with the GOPredict manuscript; Vessela Kristensen and members of the Kristensen lab in Oslo for showing me that our work can lead to biological discoveries; Kaisa Lehti and her research group; Heikki Joensuu, Tero Aittokallio, Piia-Riitta Karhemo and Harri Sihto for the journey into drug sensitivity analysis;

Tuomas Heikkinen and Dario Greco from Heli Nevanlinna's group for introducing me to medical genetics and survival analysis; and Ali Faisal from Samuel Kaski's group at Aalto University. I am fortunate and grateful for having such committed, enthusiastic and approachable collaborators.

During the course of my thesis work, I also spent three months at the Cancer Research UK Cambridge Institute in England. My sincere thanks to Dr Florian Markowetz, Dr Yinyin Yuan and the members of Florian's lab, especially Andy, Ines, Ke, Anne, Edith, Inga and Leon, for making my visit educational, efficient and enormous fun.

In biomedical journal articles, the most important persons are mentioned last in the author list and so it is here, too. My mother, Eeva, I thank for her unwavering support. Despite the bad set of cards we were dealt early on, I salute your courage and determination in pushing me to fulfill my potential. I think we both agree that your brainwashing with my god mother worked better than you thought it would. Lastly, my beloved wife and best friend Sara: I thank you for your support, understanding and companionship in all things scientific and unscientific. Neither I nor this thesis would be the same without you.

Riku Louhimo
Helsinki, July 2015

References

		Page(s)
[1]	McAfee, Andrew and Brynjolfsson, Erik. (2012) Big data: the management revolution. *Harv Bus Rev* pp. 60–66.	1
[2]	Biesdorf, S, Court, D, & Willmott, P. (2013) Big data: What's your plan? *McKinsey Quarterly*. March 2013.	1
[3]	Mauch, M, MacCallum, R. M, Levy, M, & Leroi, A. M. (2015) The Evolution of Popular Music: USA 1960-2010. *arXiv preprint arXiv:1502.05417 [physics.soc-ph]*.	1
[4]	Li, W & Radke, J. D. (2012) Geospatial data integration and modeling for the investigation of urban neighborhood crime. *Ann GIS* **18**, 185–205.	1
[5]	Chin, L, Andersen, J. N, & Futreal, P. A. (2011) Cancer genomics: from discovery science to personalized medicine. *Nat Med* **17**, 297–303.	1
[6]	Schultz, J. (1959) Integrative Mechanisms in Biology: Introductory Remarks. *American Naturalist* **93**, 209–211. Meeting of the American Society of Naturalists, Washington D.C., USA, 29 Dec 1958.	1
[7]	Jiang, P & Liu, X. S. (2015) Big data mining yields novel insights on cancer. *Nat Genet* **47**, 103–104.	2, 3, 4, 15
[8]	Kristensen, V. N, Lingjærde, O. C, Russnes, H. G, Vollan, H. K. M, Frigessi, A, & Børresen-Dale, A.-L. (2014) Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* **14**, 299–313.	2, 4, 10, 29, 35
[9]	Feinberg, A. P, Vogelstein, B, et al. (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89–92.	2
[10]	Galperin, M. Y, Rigden, D. J, & Fernández-Suárez, X. M. (2015) The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic Acids Res* **43**, D1–D5.	2
[11]	Lazebnik, Y. (2002) Can a biologist fix a radio?–Or, what I learned while studying apoptosis. *Cancer Cell* **2**, 179–182.	3
[12]	Yuan, Y, Failmezger, H, Rueda, O. M, Ali, H. R, Gräf, S, Chin, S.-F, Schwarz, R. F, Curtis, C, Dunning, M. J, Bardwell, H, Johnson, N, Doyle, S, Turashvili, G, Provenzano, E, Aparicio, S, Caldas, C, & Markowetz, F. (2012) Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling. *Sci Transl Med* **4**, 157ra143.	5, 42
[13]	Yuan, Y. (2015) Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J R Soc Interface* **12**, 20141153.	5, 16
[14]	Hanash, S. (2004) Integrated global profiling of cancer. *Nat Rev Cancer* **4**, 638–644.	7
[15]	Weinberg, R. (2013) *The biology of cancer*. (Garland Science), 2nd edition.	9

[16] Tomasetti, C & Vogelstein, B. (2015) Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81. 9

[17] Carbone, D. (1992) Smoking and cancer. *Am Journal Med* **93**, S13–S17. 9

[18] Bosch, F. X, Manos, M. M, Muñoz, N, Sherman, M, Jansen, A. M, Peto, J, Schiffman, M. H, Moreno, V, Kurman, R, Shan, K. V, et al. (1995) Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. *J Natl Cancer Inst* **87**, 796–802. 9

[19] Gillison, M. L. (2004) Human papillomavirus-associated head and neck cancer is a distinct epidemiologic, clinical, and molecular entity. *Semin Oncol* **31**, 744–754. 9

[20] Jemal, A, Bray, F, Center, M. M, Ferlay, J, Ward, E, & Forman, D. (2011) Global cancer statistics. *CA Cancer J Clin* **61**, 69–90. 9

[21] Ott, J, Ullrich, A, Mascarenhas, M, & Stevens, G. (2011) Global cancer incidence and mortality caused by behavior and infection. *J Public Health (Oxf)* **33**, 223–233. 9

[22] Negrini, S, Gorgoulis, V. G, & Halazonetis, T. D. (2010) Genomic instability—an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220–228. 9, 11

[23] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674. 9, 11, 14, 23, 42

[24] Timp, W & Feinberg, A. P. (2013) Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat Rev Cancer* **13**, 497–510. 9, 12

[25] Strachan, T & Read, A. (2011) *Human Molecular Genetics*. (Garland Science/Taylor & Francis Group), 4th edition. 10

[26] Curtis, C, Shah, S. P, Chin, S.-F, Turashvili, G, Rueda, O. M, Dunning, M. J, Speed, D, Lynch, A. G, Samarajiwa, S, Yuan, Y, Graf, S, Ha, G, Haffari, G, Bashashati, A, Russell, R, McKinney, S, METABRIC Group, Langerod, A, Green, A, Provenzano, E, Wishart, G, Pinder, S, Watson, P, Markowetz, F, Murphy, L, Ellis, I, Purushotham, A, Børresen-Dale, A.-L, Brenton, J. D, Tavare, S, Caldas, C, & Aparicio, S. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352. 10, 15, 17, 42

[27] Hyman, E, Kauraniemi, P, Hautaniemi, S, Wolf, M, Mousses, S, Rozenblum, E, Ringnér, M, Sauter, G, Monni, O, Elkahloun, A, Kallioniemi, O.-P, & Kallioniemi, A. (2002) Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. *Cancer Res* **62**, 6240–6245. 10, 14, 33

[28] Maier, T, Güell, M, & Serrano, L. (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett* **583**, 3966–3973. 10

[29] Zack, T. I, Schumacher, S. E, Carter, S. L, Cherniack, A. D, Saksena, G, Tabak, B, Lawrence, M. S, Zhang, C.-Z, Wala, J, Mermel, C. H, et al. (2013) Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134–1140. 11, 15, 18, 32

[30] Beroukhim, R, Mermel, C. H, Porter, D, Wei, G, Raychaudhuri, S, Donovan, J, Barretina, J, Boehm, J. S, Dobson, J, Urashima, M, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905. 11, 13, 35

[31] Huang, J, Wei, W, Zhang, J, Liu, G, Bignell, G. R, Stratton, M. R, Futreal, P. A, Wooster, R, Jones, K. W, & Shapero, M. H. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* **1**, 287. 11

[32] Kuga, D, Mizoguchi, M, Guan, Y, Hata, N, Yoshimoto, K, Shono, T, Suzuki, S. O, Kukita, Y, Tahira, T, Nagata, S, et al. (2008) Prevalence of copy-number neutral LOH in glioblastomas revealed by genomewide analysis of laser-microdissected tissues. *Neuro Oncol* **10**, 995–1003. 11

[33] Stephens, P. J, Greenman, C. D, Fu, B, Yang, F, Bignell, G. R, Mudie, L. J, Pleasance, E. D, Lau, K. W, Beare, D, Stebbings, L. A, et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40. 11

[34] Baca, S. C, Prandi, D, Lawrence, M. S, Mosquera, J. M, Romanel, A, Drier, Y, Park, K, Kitabayashi, N, MacDonald, T. Y, Ghandi, M, et al. (2013) Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677. 11

[35] Ciriello, G, Miller, M. L, Aksoy, B. A, Senbabaoglu, Y, Schultz, N, & Sander, C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127–1133. 11, 17, 18

[36] Jones, P. A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484–492. 12

[37] Heyn, H & Esteller, M. (2012) DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* **13**, 679–692. 12

[38] Sproul, D, Kitchen, R. R, Nestor, C. E, Dixon, J. M, Sims, A. H, Harrison, D. J, Ramsahoye, B. H, & Meehan, R. R. (2012) Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol* **13**, R84. 12

[39] Glasspool, R, Teodoridis, J. M, & Brown, R. (2006) Epigenetics as a mechanism driving polygenic clinical drug resistance. *Br J Cancer* **94**, 1087–1092. 12

[40] Stratton, M. R, Campbell, P. J, & Futreal, P. A. (2009) The cancer genome. *Nature* **458**, 719–724. 12, 13

[41] Davies, H, Bignell, G. R, Cox, C, Stephens, P, Edkins, S, Clegg, S, Teague, J, Woffendin, H, Garnett, M. J, Bottomley, W, et al. (2002) Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954. 13

[42] Wong, A. J, Bigner, S. H, Bigner, D. D, Kinzler, K. W, Hamilton, S. R, & Vogelstein, B. (1987) Increased expression of the epidermal growth factor receptor gene in malignant gliomas is invariably associated with gene amplification. *Proc Natl Acad Sci USA* **84**, 6899–6903. 13

[43] Ovaska, K, Laakso, M, Haapa-Paananen, S, Louhimo, R, Chen, P, Aittomäki, V, Valo, E, Núñez-Fontarnau, J, Rantanen, V, Karinen, S, Nousiainen, K, Lahesmaa-Korpinen, A.-M, Miettinen, M, Saarinen, L, Kohonen, P, Wu, J, Westermarck, J, & Hautaniemi, S. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med* **2**, 65. 13, 20, 24, 25, 34

[44] Hoeflich, K. P, Gray, D. C, Eby, M. T, Tien, J. Y, Wong, L, Bower, J, Gogineni, A, Zha, J, Cole, M. J, Stern, H. M, et al. (2006) Oncogenic BRAF is required for tumor growth and maintenance in melanoma models. *Cancer Res* **66**, 999–1006. 13

[45] Sarkaria, J. N, Carlson, B. L, Schroeder, M. A, Grogan, P, Brown, P. D, Giannini, C, Ballman, K. V, Kitange, G. J, Guha, A, Pandita, A, et al. (2006) Use of an orthotopic xenograft model for assessing the effect of epidermal growth factor receptor amplification on glioblastoma radiation response. *Clin Cancer Res* **12**, 2264–2271. 13

[46] The Cancer Genome Atlas Network. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582. 13, 17

[47] Seiwert, T. Y, Jagadeeswaran, R, Faoro, L, Janamanchi, V, Nallasura, V, El Dinali, M, Yala, S, Kanteti, R, Cohen, E. E, Lingen, M. W, et al. (2009) The MET receptor tyrosine kinase is a potential novel therapeutic target for head and neck squamous cell carcinoma. *Cancer Res* **69**, 3021–3031. 13

[48] Garraway, L. A, Widlund, H. R, Rubin, M. A, Getz, G, Berger, A. J, Ramaswamy, S, Beroukhim, R, Milner, D. A, Granter, S. R, Du, J, et al. (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122. 13

[49] Bass, A. J, Watanabe, H, Mermel, C. H, Yu, S, Perner, S, Verhaak, R. G, Kim, S. Y, Wardwell, L, Tamayo, P, Gat-Viks, I, et al. (2009) SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat Genet* **41**, 1238–1242. 13

[50] Santarius, T, Shipley, J, Brewer, D, Stratton, M. R, & Cooper, C. S. (2010) A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* **10**, 59–64. 13, 17, 18

[51] Zhao, M, Sun, J, & Zhao, Z. (2013) TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res* **41**, D970–D976. 13

[52] Schreiber, R. D, Old, L. J, & Smyth, M. J. (2011) Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570. 14

[53] Yoshihara, K, Shahmoradgoli, M, Martínez, E, Vegesna, R, Kim, H, Torres-Garcia, W, Treviño, V, Shen, H, Laird, P. W, Levine, D. A, et al. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**. 14

[54] Polyak, K, Haviv, I, & Campbell, I. G. (2009) Co-evolution of tumor cells and their microenvironment. *Trends Genet* **25**, 30–38. 15

[55] https://tcga data.nci.nih.gov/tcga/. (2015) The Cancer Genome Atlas website. Accessed 6 March 2015. 15

[56] Edgar, R, Domrachev, M, & Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210. 15

[57] Shen, R, Mo, Q, Schultz, N, Seshan, V. E, Olshen, A. B, Huse, J, Ladanyi, M, & Sander, C. (2012) Integrative subtype discovery in glioblastoma using icluster. *PLoS ONE* **7**, e35236. 15, 41

[58] Mo, Q, Wang, S, Seshan, V. E, Olshen, A. B, Schultz, N, Sander, C, Powers, R. S, Ladanyi, M, & Shen, R. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA* **110**, 4245–4250. 15

[59] Yuan, Y, Savage, R. S, & Markowetz, F. (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comp Biol* **7**, e1002227. 15

[60] Tong, P & Coombes, K. R. (2012) integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory. *Bioinformatics* **28**, 2861–2869. 16, 41

[61] Alvarez, H, Opalinska, J, Zhou, L, Sohal, D, Fazzari, M. J, Yu, Y, Montagna, C, Montgomery, E. A, Canto, M, Dunbar, K. B, et al. (2011) Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis. *PLoS genetics* **7**, e1001356. 16, 41

[62] Yang, D, Sun, Y, Hu, L, Zheng, H, Ji, P, Pecot, C, Zhao, Y, Reynolds, S, Cheng, H, Rupaimoole, R, Cogdell, D, Nykter, M, Broaddus, R, Rodriguez-Aguayo, C, Lopez-Berestein, G, Liu, J, Shmulevich, I, Sood, A, Chen, K, & Zhang, W. (2013) Integrated Analyses Identify a Master MicroRNA Regulatory Network for the Mesenchymal Subtype in Serous Ovarian Cancer. *Cancer Cell* **23**, 186–199. 16

[63] Jörnsten, R, Abenius, T, Kling, T, Schmidt, L, Johansson, E, Nordling, T. E, Nordlander, B, Sander, C, Gennemark, P, Funa, K, et al. (2011) Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol Syst Biol* **7**. 16

[64] Tyekucheva, S, Marchionni, L, Karchin, R, & Parmigiani, G. (2011) Integrating diverse genomic data using gene sets. *Genome Biol* **12**, R105. 16

[65] Gundem, G, Perez-Llamas, C, Jene-Sanz, A, Kedzierska, A, Islam, A, Deu-Pons, J, Furney, S. J, & Lopez-Bigas, N. (2010) IntOGen: integration and data mining of multidimensional oncogenomic data. *Nat Methods* **7**, 92–93. 16, 41

[66] Ciriello, G, Cerami, E, Sander, C, & Schultz, N. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* **22**, 398–406. 16

[67] The Cancer Genome Atlas Research Network, Weinstein, J. N, Collisson, E. A, Mills, G. B, Shaw, K. R. M, Ozenberger, B. A, Ellrott, K, Shmulevich, I, Sander, C, & Stuart, J. M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120. 16

[68] Yang, Y, Han, L, Yuan, Y, Li, J, Hei, N, & Liang, H. (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun* **5**. 16

[69] Shi, Z, Wang, J, & Zhang, B. (2013) NetGestalt: integrating multidimensional omics data over biological networks. *Nat Methods* **10**, 597–598. 16

[70] Ali, H. R, Rueda, O. M, Chin, S.-F, Curtis, C, Dunning, M. J, Aparicio, S, & Caldas, C. (2014) Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol* **15**, 431. 16, 17

[71] Sanchez-Garcia, F, Villagrasa, P, Matsui, J, Kotliar, D, Castro, V, Akavia, U.-D, Chen, B.-J, Saucedo-Cuevas, L, Barrueco, R. R, Llobet-Navas, D, Silva, J. M, & Pe'er, D. (2014) Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers. *Cell* **159**, 1461–1475. 16

[72] Osmanbeyoglu, H. U, Pelossof, R, Bromberg, J. F, & Leslie, C. S. (2014) Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res* **24**, 1869–1880. 16

[73] Zhao, Q, Shi, X, Xie, Y, Huang, J, Shia, B, & Ma, S. (2015) Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinformatics* **16**, 291–303. 16

[74] Yuan, Y, Van Allen, E. M, Omberg, L, Wagle, N, Amin-Mansour, A, Sokolov, A, Byers, L. A, Xu, Y, Hess, K. R, Diao, L, et al. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* **32**, 644–652. 16

[75] Venet, D, Dumont, J. E, & Detours, V. (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comp Biol* **7**, e1002240. 17

[76] Kim, D, Shin, H, Song, Y. S, & Kim, J. H. (2012) Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform* **45**, 1191–1198. 17

[77] Sørlie, T, Perou, C. M, Tibshirani, R, Aas, T, Geisler, S, Johnsen, H, Hastie, T, Eisen, M. B, van de Rijn, M, Jeffrey, S. S, et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**, 10869–10874. 17, 18, 37

[78] Furnari, F. B, Fenton, T, Bachoo, R. M, Mukasa, A, Stommel, J. M, Stegh, A, Hahn, W. C, Ligon, K. L, Louis, D. N, Brennan, C, et al. (2007) Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev* **21**, 2683–2710. 17

[79] Mischel, P. S, Shai, R, Shi, T, Horvath, S, Lu, K. V, Choe, G, Seligson, D, Kremen, T. J, Palotie, A, Liau, L. M, et al. (2003) Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene* **22**, 2361–2373. 17

[80] Verhaak, R. G, Hoadley, K. A, Purdom, E, Wang, V, Qi, Y, Wilkerson, M. D, Miller, C. R, Ding, L, Golub, T, Mesirov, J. P, et al. (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110. 17

[81] Ferlay, J, Shin, H.-R, Bray, F, Forman, D, Mathers, C, & Parkin, D. M. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* **127**, 2893–2917. 17

[82] The Cancer Genome Atlas Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525. 17, 22

[83] Vermorken, J. B & Specenier, P. (2010) Optimal treatment for recurrent/metastatic head and neck cancer. *Ann Oncol* **21**, vii252–vii261. 17

[84] Ang, K. K, Harris, J, Wheeler, R, Weber, R, Rosenthal, D. I, Nguyen-Tân, P. F, Westra, W. H, Chung, C. H, Jordan, R. C, Lu, C, et al. (2010) Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med* **363**, 24–35. 17

[85] Leemans, C. R, Braakhuis, B. J, & Brakenhoff, R. H. (2011) The molecular biology of head and neck cancer. *Nat Rev Cancer* **11**, 9–22. 17

[86] Chung, C. H, Parker, J. S, Karaca, G, Wu, J, Funkhouser, W. K, Moore, D, Butterfoss, D, Xiang, D, Zanation, A, Yin, X, et al. (2004) Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* **5**, 489–500. 17

[87] Walter, V, Yin, X, Wilkerson, M. D, Cabanski, C. R, Zhao, N, Du, Y, Ang, M. K, Hayward, M. C, Salazar, A. H, Hoadley, K. A, et al. (2013) Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PLoS ONE* **8**, e56823. 17

[88] Suh, Y, Amelio, I, Urbano, T. G, & Tavassoli, M. (2014) Clinical update on cancer: molecular oncology of head and neck cancer. *Cell Death Dis* **5**, e1018. 18

[89] Suomen Rintasyöpäryhmä ry (Finnish Breast Cancer Group). (2013) Rintasyövän valtakunnallinen diagnostiikka- ja hoitosuositus (http://rintasyoparyhma.yhdistysavain.fi/hoitosuositus/). In Finnish. 18

[90] Holm, K, Staaf, J, Jönsson, G, Vallon-Christersson, J, Gunnarsson, H, Arason, A, Magnusson, L, Barkardottir, R. B, Hegardt, C, Ringnér, M, et al. (2012) Characterisation of amplification patterns and target genes at chromosome 11q13 in CCND1-amplified sporadic and familial breast tumours. *Breast Cancer Res Treat* **133**, 583–594. 18

[91] Morton, L. M, Wang, S. S, Devesa, S. S, Hartge, P, Weisenburger, D. D, & Linet, M. S. (2006) Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. *Blood* **107**, 265–276. 18

[92] Lenz, G & Staudt, L. M. (2010) Aggressive lymphomas. *N Engl J Med* **362**, 1417–1429. 18

[93] Wright, G, Tan, B, Rosenwald, A, Hurt, E. H, Wiestner, A, & Staudt, L. M. (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci USA* **100**, 9991–9996. 18

[94] Pasqualucci, L, Trifonov, V, Fabbri, G, Ma, J, Rossi, D, Chiarenza, A, Wells, V. A, Grunn, A, Messina, M, Elliot, O, et al. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat Genet* **43**, 830–837. 18

[95] Morin, R. D, Mungall, K, Pleasance, E, Mungall, A. J, Goya, R, Huff, R. D, Scott, D. W, Ding, J, Roth, A, Chiu, R, et al. (2013) Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265. 18

[96] Monti, S, Chapuy, B, Takeyama, K, Rodig, S. J, Hao, Y, Yeda, K. T, Inguilizian, H, Mermel, C, Currie, T, Dogan, A, et al. (2012) Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* **22**, 359–372. 18, 32

[97] The Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70. 18, 22

[98] Nik-Zainal, S, Alexandrov, L. B, Wedge, D. C, Van Loo, P, Greenman, C. D, Raine, K, Jones, D, Hinton, J, Marshall, J, Stebbings, L. A, et al. (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993. 19

[99] Alexandrov, L. B, Nik-Zainal, S, Wedge, D. C, Aparicio, S. A, Behjati, S, Biankin, A. V, Bignell, G. R, Bolli, N, Borg, A, Børresen-Dale, A.-L, et al. (2013) Signatures of mutational processes in human cancer. *Nature* **500**, 415–421. 19

[100] Trevino, V, Falciani, F, & Barrera-Saldaña, H. A. (2007) DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med* **13**, 527. 19

[101] Irizarry, R. A, Hobbs, B, Collin, F, Beazer-Barclay, Y. D, Antonellis, K. J, Scherf, U, & Speed, T. P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264. 20

[102] Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat Genet* **32**, 496–501. 20, 23

[103] Chen, P, Lepikhova, T, Hu, Y, Monni, O, & Hautaniemi, S. (2011) Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. *Nucleic Acids Res* **39**, e123. 20, 25

[104] Laird, P. W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11**, 191–203. 20

[105] Feber, A, Guilhamon, P, Lechner, M, Fenton, T, Wilson, G. A, Thirlwell, C, Morris, T. J, Flanagan, A. M, Teschendorff, A. E, Kelly, J. D, et al. (2014) Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol* **15**, R30. 20

[106] The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068. 22

[107] The Cancer Genome Atlas Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615. 22

[108] The Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337. 22

[109] Van de Wiel, M. A, Picard, F, Van Wieringen, W. N, & Ylstra, B. (2011) Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform* **12**, 10–21. 23, 24

[110] Olshen, A. B, Venkatraman, E, Lucito, R, & Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572. 23, 25

[111] Hupé, P, Stransky, N, Thiery, J.-P, Radvanyi, F, & Barillot, E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422. 23

[112] Lai, W. R, Johnson, M. D, Kucherlapati, R, & Park, P. J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770. 23

[113] Willenbrock, H & Fridlyand, J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091. 23, 29

[114] Van Loo, P, Nordgard, S. H, Lingjærde, O. C, Russnes, H. G, Rye, I. H, Sun, W, Weigman, V. J, Marynen, P, Zetterberg, A, Naume, B, Perou, C. M, Børresen-Dale, A.-L, & Kristensen, V. N. (2010) Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* **107**, 16910–16915. 23

[115] Rasmussen, M, Sundstrom, M, Goransson Kultima, H, Botling, J, Micke, P, Birgisson, H, Glimelius, B, & Isaksson, A. (2011) Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* **12**, R108–R108. 23

[116] Greenman, C. D, Bignell, G, Butler, A, Edkins, S, Hinton, J, Beare, D, Swamy, S, Santarius, T, Chen, L, Widaa, S, et al. (2010) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164–175. 23

[117] Benjamini, Y & Yekutieli, D. (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann Stat* **29**, 1165–1188. 25

[118] Carvalho, B, Bengtsson, H, Speed, T. P, & Irizarry, R. A. (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* **8**, 485–499. 25

[119] Du, P, Zhang, X, Huang, C, Jafari, N, Kibbe, W, Hou, L, & Lin, S. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587. 26

[120] Benjamini, Y & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* pp. 289–300. 26

[121] Liu, C, Louhimo, R, Laakso, M, Lehtonen, R, & Hautaniemi, S. (2015) Identification of sample-specific regulations using integrative network level analysis. *BMC Cancer* **15**, 319. 26

[122] Laakso, M & Hautaniemi, S. (2010) Integrative platform to translate gene sets to networks. *Bioinformatics* **26**, 1802–1803. 26

[123] Kanehisa, M, Goto, S, Sato, Y, Furumichi, M, & Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–D114. 26

[124] Kelder, T, van Iersel, M, Hanspers, K, Kutmon, M, Conklin, B, Evelo, C, & Pico, A. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res* **40**, D1301–D1307. 26

[125] Wu, J, Vallenius, T, Ovaska, K, Westermarck, J, Mäkelä, T. P, & Hautaniemi, S. (2009) Integrated network analysis platform for protein-protein interactions. *Nat Methods* **6**, 75–77. 26

[126] Cowley, M, Pinese, M, Kassahn, K, Waddell, N, Pearson, J, Grimmond, S, Biankin, A, Hautaniemi, S, & Wu, J. (2012) PINA v2.0: mining interactome modules. *Nucleic Acids Res* **40**, D862–D865. 26

[127] Ashburner, M, Ball, C. A, Blake, J. A, Botstein, D, Butler, H, Cherry, J. M, Davis, A. P, Dolinski, K, Dwight, S. S, & Eppig, J. T. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29. 26

[128] Cerami, E. G, Gross, B. E, Demir, E, Rodchenkov, I, Babur, O, Anwar, N, Schultz, N, Bader, G. D, & Sander, C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685–D690. 26

[129] Kanehisa, M, Goto, S, Furumichi, M, Tanabe, M, & Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355–D360. 26, 35

[130] Knox, C, Law, V, Jewison, T, Liu, P, Ly, S, Frolkis, A, Pon, A, Banco, K, Mak, C, Neveu, V, Djoumbou, Y, Eisner, R, Guo, A. C, & Wishart, D. S. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **39**, D1035–D1041. 26, 35

[131] Altman, D. G & Bland, J. M. (1998) Time to event (survival) data. *BMJ* **317**, 468–469. 27

[132] Kaplan, E. L & Meier, P. (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* **53**, 457–481. 27

[133] Bland, J. M & Altman, D. G. (2004) The logrank test. *BMJ* **328**, 1073. 28

[134] Andersen, P. K & Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *Ann Stat* pp. 1100–1120. 28

[135] Burton, A & Altman, D. (2004) Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* **91**, 4–8. 28

[136] Grambsch, P. M & Therneau, T. M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526. 28

[137] Garraway, L. A & Lander, E. S. (2013) Lessons from the cancer genome. *Cell* **153**, 17–37. 29

[138] Fröhling, S & Döhner, H. (2008) Chromosomal abnormalities in cancer. *N Engl J Med* **359**, 722–734. 29

[139] Tang, Y.-C & Amon, A. (2013) Gene copy-number alterations: a cost-benefit analysis. *Cell* **152**, 394–405. 29

[140] Berger, J. A, Hautaniemi, S, Mitra, S. K, & Astola, J. (2006) Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans Comput Biol Bioinformatics* **3**, 2. 29

[141] Bicciato, S, Spinelli, R, Zampieri, M, Mangano, E, Ferrari, F, Beltrame, L, Cifola, I, Peano, C, Solari, A, & Battaglia, C. (2009) A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Res* **37**, 5057–5070. 29

[142] Choi, H, Qin, Z. S, & Ghosh, D. (2010) A Double-Layered Mixture Model for the Joint Analysis of DNA Copy Number and Gene Expression Data. *J Comput Biol* **17**, 121–137. 29

[143] Hautaniemi, S, Ringnér, M, Kauraniemi, P, Autio, R, Edgren, H, Yli-Harja, O, Astola, J, Kallioniemi, A, & Kallioniemi, O.-P. (2004) A strategy for identifying putative causes of gene expression variation in human cancers. *J Franklin Inst* **341**, 77–88. 29, 33

[144] Lahti, L, Myllykangas, S, Knuutila, S, & Kaski, S. (2009) *Dependency detection with similarity constraints*, Proceedings of the 2009 IEEE International Workshop on Machine Learning for Signal Processing XIX. (IEEE, Piscataway, NJ, USA), pp. 89–94. 29

[145] Lê Cao, K.-A, González, I, & Déjean, S. (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **25**, 2855–2856. 29

[146] Menezes, R, Boetzer, M, Sieswerda, M, van Ommen, G.-J, & Boer, J. (2009) Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics* **10**, 203. 29

[147] Salari, K, Tibshirani, R, & Pollack, J. R. (2010) DR–Integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics* **26**, 414–416. 29

[148] Schäfer, M, Schwender, H, Merk, S, Haferlach, C, Ickstadt, K, & Dugas, M. (2009) Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics* **25**, 3228–3235. 29

[149] van Wieringen, W. N & van de Wiel, M. A. (2009) Nonparametric Testing for DNA Copy Number Induced Differential mRNA Gene Expression. *Biometrics* **65**, 19–29. 29

[150] Alkodsi, A, Louhimo, R, & Hautaniemi, S. (2015) Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief Bioinform* **16**, 242–254. 31, 42

[151] Duan, J, Zhang, J.-G, Deng, H.-W, & Wang, Y.-P. (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS ONE* **8**, e59128. 31

[152] Kwiecinska, A, Ichimura, K, Berglund, M, Dinets, A, Sulaiman, L, Collins, V. P, Larsson, C, Porwit, A, & Lagercrantz, S. B. (2014) Amplification of 2p as a genomic marker for transformation in lymphoma. *Genes Chromosomes Cancer* **53**, 750–768. 33

[153] van de Sluis, B, Mao, X, Zhai, Y, Groot, A. J, Vermeulen, J. F, van der Wall, E, van Diest, P. J, Hofker, M. H, Wijmenga, C, Klomp, L. W, et al. (2010) COMMD1 disrupts HIF-1α/β dimerization and inhibits human tumor cell invasion. *J Clin Invest* **120**, 2119. 33

[154] Muller, P. A & Vousden, K. H. (2013) p53 mutations in cancer. *Nat Cell Biol* **15**, 2–8. 33

[155] Tanaka, S, Louis, D. N, Curry, W. T, Batchelor, T. T, & Dietrich, J. (2013) Diagnostic and therapeutic avenues for glioblastoma: no longer a dead end? *Nat Rev Clin Oncol* **10**, 14–26. 34

[156] Hobbs, J, Nikiforova, M. N, Fardo, D. W, Bortoluzzi, S, Cieply, K, Hamilton, R. L, & Horbinski, C. (2012) Paradoxical relationship between degree of EGFR amplification and outcome in glioblastomas. *Am J Surg Pathol* **36**, 1186. 34

[157] Aure, M, Leivonen, S.-K, Fleischer, T, Zhu, Q, Overgaard, J, Alsner, J, Tramm, T, Louhimo, R, Alnaes, G. I, Perälä, M, Busato, F, Touleimat, N, Tost, J, Børresen-Dale, A.-L, Hautaniemi, S, Troyanskaya, O, Lingjaerde, O, Sahlberg, K, & Kristensen, V. (2013) Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. *Genome Biol* **14**, R126. 34, 41

[158] McDermott, U, Downing, J. R, & Stratton, M. R. (2011) Genomics and the continuum of cancer care. *N Engl J Med* **364**, 340–350. 35

[159] Li, Y. Y & Jones, S. (2012) Drug repositioning for personalized medicine. *Genome Med* **4**, 27. 35

[160] Haber, D. A, Gray, N. S, & Baselga, J. (2011) The evolving war on cancer. *Cell* **145**, 19–24. 35

[161] Forbes, S, Bindal, N, Bamford, S, Cole, C, Kok, C, Beare, D, Jia, M, Shepherd, R, Leung, K, Menzies, A, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945. 35

[162] Futreal, P. A, Coin, L, Marshall, M, Down, T, Hubbard, T, Wooster, R, Rahman, N, & Stratton, M. R. (2004) A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183. 35

[163] Chowdhury, S & Sarkar, R. R. (2015) Comparison of human cell signaling pathway databases — evolution, drawbacks and challenges. *Database* **2015**, bau126. 36

[164] Nguyen, N. T, Lindsey, M. L, & Jin, Y.-F. (2015) Systems analysis of gene ontology and biological pathways involved in post-myocardial infarction responses. *BMC Genomics* **16**, S18. 36

[165] Wiedemeyer, W. R, Beach, J. A, & Karlan, B. Y. (2014) Reversing platinum resistance in high-grade serous ovarian carcinoma: targeting BRCA and the homologous recombination system. *Front Oncol* **4**. 38

[166] Lahti, L, Schäfer, M, Klein, H.-U, Bicciato, S, & Dugas, M. (2013) Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. *Brief Bioinform* **14**, 27–35. 40

[167] Stolovitzky, G, Monroe, D, & Califano, A. (2007) Dialogue on Reverse-Engineering Assessment and Methods (DREAM). *Ann NY Acad Sci* **1115**, 1–22. 40

[168] Küffner, R, Zach, N, Norel, R, Hawe, J, Schoenfeld, D, Wang, L, Li, G, Fang, L, Mackey, L, Hardiman, O, et al. (2015) Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotechnol* **33**, 51–57. 40

[169] Costello, J. C, Heiser, L. M, Georgii, E, Gönen, M, Menden, M. P, Wang, N. J, Bansal, M, Hintsanen, P, Khan, S. A, Mpindi, J.-P, et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* **32**, 1202–1212. 40

[170] Marbach, D, Costello, J. C, Küffner, R, Vega, N. M, Prill, R. J, Camacho, D. M, Allison, K. R, Kellis, M, Collins, J. J, Stolovitzky, G, et al. (2012) Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796–804. 40

[171] Rubio-Perez, C, Tamborero, D, Schroeder, M. P, Antolín, A. A, Deu-Pons, J, Perez-Llamas, C, Mestres, J, Gonzalez-Perez, A, & Lopez-Bigas, N. (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**, 382–396. 41

[172] Cerami, E, Gao, J, Dogrusoz, U, Gross, B. E, Sumer, S. O, Aksoy, B. A, Jacobsen, A, Byrne, C. J, Heuer, M. L, Larsson, E, et al. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404. 41

[173] Caponetti, G. C, Dave, B. J, Perry, A. M, Smith, L. M, Jain, S, Meyer, P. N, Bast, M, Bierman, P. J, Bociek, R. G, Vose, J. M, et al. (2015) Isolated MYC cytogenetic abnormalities in diffuse large B-cell lymphoma do not predict an adverse clinical outcome. *Leuk Lymphoma* **Epub ahead of print**. 42

[174] Louhimo, R, Laakso, M, Heikkinen, T, Laitinen, S, Manninen, P, Rogojin, V, Miettinen, M, Blomqvist, C, Liu, J, Nevanlinna, H, & Hautaniemi, S. (2013) Identification of genetic markers with synergistic survival effect in cancer. *BMC Syst Biol* **7**, S2. 42

[175] Onay, V. Ü, Briollais, L, Knight, J. A, Shi, E, Wang, Y, Wells, S, Li, H, Rajendram, I, Andrulis, I. L, & Ozcelik, H. (2006) SNP-SNP interactions in breast cancer susceptibility. *BMC Cancer* **6**, 114. 42