TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

MACHINE LEARNING AND COMPUTATIONAL METHODS TO IDENTIFY MOLECULAR AND CLINICAL MARKERS FOR COMPLEX DISEASES

case studies in cancer and obesity

Fatemeh Seyednasrollah

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 663 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2022

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

MACHINE LEARNING AND COMPUTATIONAL METHODS TO IDENTIFY MOLECULAR AND CLINICAL MARKERS FOR COMPLEX DISEASES

case studies in cancer and obesity

Fatemeh Seyednasrollah

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 663 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2022

University of Turku

Faculty of Science
Department of Mathematics and Statistics (Applied mathematics)
Doctoral Programme in Exact Sciences

Supervised by

Professor Laura L. Elo
Turku Bioscience Centre
University of Turku and Åbo Akademi
Turku, Finland

Professor Marko M. Mäkelä
Department of Mathematics and Statistics, University of Turku
Turku, Finland

Reviewed by

Professor Merja Heinäniemi
Institute of Biomedicine, School of Medicine, University of Eastern Finland
Kuopio, Finland

Associate Professor Heidi Peterson
Institute of Computer Science
University of Tartu
Tartu, Estonia

Opponent

Assistant professor Ewa Szczurek
Institute of Informatics
University of Warsaw
Warsaw, Poland

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-8759-7 (PRINT)
ISBN 978-951-29-8760-3 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)
Painosalama, Turku, Finland 2022

To my mother, Zahra Zabihinik

UNIVERSITY OF TURKU
Faculty of Science
Department of Mathematics and Statistics (Applied mathematics)
SEYEDNASROLLAH, FATEMEH: Machine Learning and Computational Methods to Identify molecular and Clinical Markers for Complex Diseases – case studies in cancer and obesity
Doctoral Dissertation, 177 pp.
Doctoral Programme in Exact Sciences
December 2021

ABSTRACT

In biomedical research, applied machine learning and bioinformatics are the essential disciplines heavily involved in translating data-driven findings into medical practice. This task is especially accomplished by developing computational tools and algorithms assisting in detection and clarification of underlying causes of the diseases. The continuous advancements in high-throughput technologies coupled with the recently promoted data sharing policies have contributed to presence of a massive wealth of data with remarkable potential to improve human health care. In concordance with this massive boost in data production, innovative data analysis tools and methods are required to meet the growing demand. The data analyzed by bioinformaticians and computational biology experts can be broadly divided into molecular and conventional clinical data categories. The aim of this thesis was to develop novel statistical and machine learning tools and to incorporate the existing state-of-the-art methods to analyze bio-clinical data with medical applications. The findings of the studies demonstrate the impact of computational approaches in clinical decision making by improving patients risk stratification and prediction of disease outcomes.

This thesis is comprised of five studies explaining method development for 1) genomic data, 2) conventional clinical data and 3) integration of genomic and clinical data. With genomic data, the main focus is detection of differentially expressed genes as the most common task in transcriptome profiling projects. In addition to reviewing available differential expression tools, a data-adaptive statistical method called Reproducibility Optimized Test Statistic (ROTS) is proposed for detecting differential expression in RNA-sequencing studies. In order to prove the efficacy of ROTS in real biomedical applications, the method is used to identify prognostic markers in clear cell renal cell carcinoma (ccRCC). In addition to previously known markers, novel genes with potential prognostic and therapeutic role in ccRCC are detected. For conventional clinical data, ensemble based predictive models are developed to provide clinical decision support in treatment of patients with metastatic castration resistant prostate cancer (mCRPC). The proposed predictive models cover treatment and survival stratification tasks for both trial-based and real-world patient cohorts. Finally, genomic and conventional clinical data are integrated to demonstrate the importance of inclusion of genomic data in predictive ability of clinical models. Again, utilizing ensemble-based learners, a novel model is proposed to predict adulthood obesity using both genetic and social-environmental factors.

Overall, the ultimate objective of this work is to demonstrate the importance of clinical bioinformatics and machine learning for bio-clinical marker discovery in complex disease with high heterogeneity. In case of cancer, the interpretability of clinical models strongly depends on predictive markers with high reproducibility supported by validation data. The discovery of these markers would increase chance of early detection and improve prognosis assessment and treatment choice.

KEYWORDS: differential expression testing, machine learning, boosting algorithms, survival analysis, next generation sequencing, transcriptomics, clear cell renal cell carcinoma, metastatic-castration resistance prostate cancer, obesity

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Matematiikan ja tilastotieteen laitos
SEYEDNASROLLAH, FATEMEH: Machine Learning and Computational Methods to Identify molecular and Clinical Markers for Complex Diseases – case studies in cancer and obesity
Väitöskirja, 177 s.
Eksaktien tieteiden tohtoriohjelma
Joulukuu 2021

TIIVISTELMÄ

Sovellettua koneoppimista ja bioinformatiikkaa käytetään biolääketieteessä tietoaineistojen analysointiin ja uusien analyysimenetelmien kehittämiseen. Analyysin tuloksista johdetaan uusia lääketieteellisiä käytäntöjä ja hoitoja sekä etsitään sairauksien syitä. Tässä väitöskirjassa kehitetään uusia laskennallisia ja tilastollisia koneoppimismenetelmiä, ja sovelletaan niitä biolääketieteellisiin ja kliinisiin aineistoihin. Saadut tulokset osoittavat, että kehitetyt menetelmät parantavat kliinisiä päätöksiä mm. potilaiden riskiluokituksen ja sairauksien vakavuuden arvioinnissa.

Tämä väitöstutkimus sisältää viisi osatyötä, joissa käsitellään menetelmiä 1) genomiikalle, 2) kliiniselle datalle ja 3) näiden yhdistelmälle. Kehitetyt menetelmät hyödyntävät useita erilaisia lähestymistapoja kuten toistettavuuden optimointia, elinaika-analyysia ja ensemble-oppimista. Sovelluksina muun muassa tunnistetaan markkereita munuaiskarsinooman alatyyppiin, tuetaan kliinistä päätöksentekoa eturauhassyövän hoidossa ja ennustetaan aikuisiän ylipainoa.

Väitöstyön päätavoite on esitellä bioinformatiikan ja koneoppimisen soveltuvuutta markkereiden tunnistamiseen heterogeenisestä datasta. Syöpätutkimuksen tapauksessa kliinisten mallien hyödyntäminen edellyttää robusteja markkereita, joiden luotettavuus on validoitu usealla riippumattomalla datalla. Tällaisten markkereiden löytäminen voi edesauttaa syövän varhaista diagnosointia, etenemisen ennustamista ja oikean hoitomuodon valintaa.

AVAINSANAT: erilainen ekspressio, koneoppiminen, oppimisen tehostaminen, elinaika-analyysi, uuden sukupolven sekvensointi, transkriptomiikka, munuaiskarsinooma, eturauhassyöpä, ylipaino

Table of Contents

Abbreviations

Adaboost	Adaptive boosting
ADT	Androgen depletion therapy
ALB	Albumin
ALP	Alkaline phosphatase
AST	Aspartate aminotransferase
AUC	Area under the curve
AUPRC	Area under the precision-recall curve
BMI	Body mass index
ccRCC	Clear cell renal carcinoma
cDNA	Complementary deoxyribonucleic acid
CRP	C-reactive protein
DAVID	Database for Annotation, Visualization an Integrated Discovery
DNA	Deoxyribonucleic acid
DREAM	Dialogue for Reverse Engineering Assessments and Methods
ECOG	Eastern Cooperative Oncology Group
EDA	Exploratory data analysis
EGA	European Genome-phenome Archive
EMRs	Electronic medical records
ERCC	External Ribonucleic acid molecules Control Consortium
FN	False negative
FP	False positive
FSAM	Forward stagewise additive modeling
GBM	Gradient boosting machines
GC	Guanine-cytosine
GO	Gene Ontology
GRS	Genetic risk score
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association studies
HB	Hemoglobin
iAUC	Integrated are under the curve
IMG	Integrated Microbial Genomes

IPA	Ingenuity pathway analysis
KEGG	Kyoto Encyclopedia of genes and genomes
KM	Kaplan-Meier
Lasso	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LDH	Lactate dehydrogenase
log-cpm	Log-counts per million
MAR	Missing at random
mCRPC	Metastatic castration-resistant prostate cancer
mRNA	Messenger ribonucleic acid molecules
NGS	Next generation sequencing
PCA	Principal component analysis
PCR	Polymerase chain reaction
PSA	Prostate-specific antigen
RCC	Renal cell carcinoma
RCT	Randomized clinical trial
RNA	Ribonucleic acid molecules
RNA-seq	RNA-sequencing
ROC	Receiver operating characteristic
ROTS	Reproducibility Optimized Test Statistic
RPKM	Reads Per Kilobase per Million mapped reads
SLC	Solute carrier
SNPs	Single nucleotide polymorphisms
TCGA	The Cancer Genome Atlas
TMM	Trimmed Mean of M-values
TN	True negative
TP	True positive
Voom	Variance modeling at the observational level
WGS	Whole-genome sequencing
YFS	Young Finns Study

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I **Seyednasrollah, F.**, Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief. Bioinform. 16, 59–70 (2015).

II **Seyednasrollah, F.**, Rantanen, K., Jaakkola, P. & Elo, L. L. ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. Nucleic Acids Res. gkv806 (2015).

III **Seyednasrollah, F.**, Koestler, C. D., Wang T., Piccolo, S. R., Vega, R., Greiner, R., Fuchs, C., Gofer, E., Kumar, L., Wolfinger, R. D., Kanigel, Winner K., Bare, C., Neto E. C., Yu, T., Shen, L., Abdallah, K., Norman, T., Stolovitzky, G., Soule, H. R., Sweeney, C. J., Ryan, C. J., Scher, H. I., Sartor, O., Elo, L. L., Zhou, F. L., Guinney, J., Costello, J. C., and Prostate Cancer DREAM Challenge Community A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. JCO Clin. Cancer Inform. 1–15 (2017).

IV **Seyednasrollah, F.**, Mahmoudian, M., Rautakorpi, L., Hirvonen, O., Laitinen, T., Jyrkkiö, S., Elo, L. L. How Reliable are Trial-based Prognostic Models in Real-world Patients with Metastatic Castration-resistant Prostate Cancer? Eur. Urol. 71, 838–840 (2017).

V **Seyednasrollah, F.**, Mäkelä, J., Pitkänen, N., Juonala, M., Hutri-Kähönen, N., Lehtimäki, T., Viikari, J., Kelly, T., Li, C., Bazzano, L., Elo, L. L., Raitakari, O. T. Prediction of Adulthood Obesity Using Genetic and Childhood Clinical Risk Factors in the Cardiovascular Risk in Young Finns Study. Circ. Cardiovasc. Genet. 10, e001554 (2017).

The original publications have been reproduced with the permission of the copyright holders.

1 Introduction

1.1 Background and motivation

Over the last two decades, the rate of biological data generation has been growing dramatically. The first consequence of encountering such a massive and complex data was an urgent demand for informatics facilities and specialists to optimally extract knowledge from the available raw data. Developing a new interdisciplinary filed, bioinformatics, was the biomedical community's response to address their needs [1]. When first introduced, bioinformatics was defined as utilizing computational tools and hardware resources to store, manage and ultimately understand the information encoded in biological data. The advent of high-throughput technologies was a turning point from traditional bioinformatics to the advanced computational biology and systems biology with deeper biomedical insight. It shifted the focus of bioinformatics tools from database management and query optimization algorithms to more sophisticated statistical learning and mathematical modeling approaches. Despite the remarkable theoretical and practical achievements, the field still lacks consensus in various aspects specially in development of rigorous data analysis tools and methods. In this thesis, I intend to elaborate some of the interesting while critical challenges of the field and the corresponding novel solutions to them I enjoyed dealing with during my PhD studies.

In the field of computational biology, a large proportion of projects involve medical applications. These projects utilize applied mathematics to improve current level of knowledge about mechanism of complex diseases such as cancer. The information provided by prediction models assist healthcare experts and patients in deciding which treatment strategy to take or avoid. For instance, the developed models facilitate assessment of patient prognosis and overall survival, prediction of disease recurrence or evaluation of safety and efficacy of therapeutic strategies. More importantly, these models serve as an aid to early diagnosis of the disease which is a key point in treatment outcome. To achieve this insight, machine learning algorithms learn from a training patient cohort for whom clinically relevant information as well as outcome of interest are available. Traditional prediction models have typically relied solely on clinical features such as demographic data, laboratory tests, pathology and medical history, which are mainly collected under two main categories: real-world electronic medical records (EMRs) and formal randomized clinical trials (RCTs). There are controversial opinions about usage of trial-based prediction models in everyday clinical practice. This is mainly because RCTs involve more homogenous cohort of patients compared to real-world situations [2]. Hence, assessing the reliability of trial-based prediction models in real-world cases is crucial.

In addition to clinical data, accurate and cost-effective high throughput technologies have enabled us to incorporate molecular or genetic information into prediction models. While molecular data may cover a large set of different omics, this thesis uses information from gene expression and genetic variants to predict clinical outcomes. In case of gene expression data, a common research task is to identify genes which are differentially expressed across distinct biological conditions (e.g. healthy versus diseased). So far, various statistical tools have been developed to perform differential expression testing; however, there is no clear consensus about the choice of optimal tool; as available tools behave discordant under different experimental designs. Since the identified genes may further serve as potential biomarkers, development of an accurate differential expression analysis tool is an important challenge to be addressed. In this domain, efficient statistical tools require robust approach to precisely estimate expression variability between and within experimental samples. When studying complex diseases, data heterogeneity is a real threat to analysis reproducibility. One approach to dealing with this problem is to combine genetic and clinical data to provide deeper understanding of the disease mechanism. This way, we expect to improve model performance and generalizability.

When the desired variables are collected, the next step is data preprocessing and standardization irrespective of the data type. This can be a challenging and time-consuming step as often there is no domain-specific standard workflow for data curation [3] [4]. A basic preprocessing approach standardizes variables by curing noisy and outlier incidences and imputing missing data when necessary. In case of genetic data, biological and systematic technical biases need to get minimized. After preprocessing the data, the next step is to build the prediction model. So far, numerous machine learning approaches have been proposed to predict healthcare outcomes. To account for clinical data complexity, ensemble-based methods with ability to capture interactions and non-linear associations between the variables are attracting great interest [5]. Among the many ensemble-based methods, gradient boosting is one of the most powerful techniques with common utility in medical applications [6] [7]. The primary goal of gradient boosting is to combine many weak learners to produce a powerful model with optimally enhanced performance [8]. Internally, gradient boosting strategies perform feature selection and are capable of handling missing data. When a model is built, selected variables are reported in an ordered format according to their relative influence to facilitate model interpretation. An efficient learning algorithm should establish a balanced approach between the training and the test error rates to avoid model overfitting. Various methods such as cross-validation techniques are used to control model complexity and stop the learning procedure before overfitting occurrence. Once a model is trained, it is necessary to validate the predictive performance using suitable metrics (e.g. accuracy, ROC curve-AUC score), more preferably by an independent external data.

The general aim of clinical predictive models is to improve the patient quality of care; and today, medical bioinformatics is becoming a main part of every clinical study [9]. The risk stratification methods are used to elucidate the causal factors of the disease, prognosis and possible treatment outcome. The studies presented in the following sections demonstrate the practical utility of applied mathematics and computer science in medical applications.

1.2 Objectives and outline of this thesis

Given the significant role of computational techniques in medical applications advancements, this thesis aims to introduce novel tools and algorithms developed to extract clinically relevant information from the biomedical raw material. More specifically, it provides robust methods to identify bio-clinical markers with substantial role in disease development, prognosis and treatment outcome. As case studies, the developed methods are used to identify:

a. prognostic biomarkers in clear cell renal cell carcinoma (ccRCC),

b. clinical markers predicting treatment outcome in metastatic castration-resistant prostate cancer (mCRPC),

c. bio-clinical markers predicting adulthood obesity.

Additionally, this thesis assesses the reliability of trial-based prognostic models in read-world patients with mCRPC. This is a clinically critical question as patient prognosis is one of the main indicators of treatment plan for these patients.

My PhD studies started by investigating computational tools developed for detection of differentially expressed genes in transcriptomics studies. Considering the RNA-sequencing (RNA-seq) technology as the current standard protocol in transcriptome profiling, my first publication provides a practical guideline to assist researchers in choosing suitable method for different RNA-seq experimental designs. In particular, a comprehensive comparative study was launched to assess the performance of eight widely used methods in detection of differentially expressed genes in real-world datasets with relatively large sample sizes. The results of this study revealed a significant inconsistency among methods in case of their differential expression detections and served as a proved demand for a robust method in the field. Inspired by the results from the first publication, in publication II, a new data-adaptive statistical method to detect differentially expressed features in RNA-seq studies is introduced. The proposed method is Reproducibility Optimized Test Statistic (ROTS) which aims to maximize the reproducibility of detections by optimizing a data-driven modified t-test statistic. The performance of the method in terms of sensitivity and specificity was first verified using a spike-in dataset with controlled mixes of synthesized transcripts. Next, this method was utilized to detect novel prognostic biomarkers for ccRCC. As a result, I was able to stratify ccRCC patients into distinct groups of poor, moderate and better prognosis by feeding ROTS detections into tailored clustering and survival techniques. Further enrichment pathway analysis confirmed the potential possibility of employing some of the proposed novel biomarkers as drug target candidates. Throughout this thesis, the term “detection” refers to differentially expressed genomic features (most often genes) detected by the utilized tools.

Despite the significant role of molecular markers in understanding and treatment of cancer, conventional clinical data is still the gold standard utilized in everyday clinical practice. Therefore, publications III and IV in this thesis focus on development of predictive models using clinical features. In publication III, a novel modeling approach to predict short-term discontinuation of chemotherapy in mCRPC was developed. This paper is a collaborative project between seven international teams and we successfully show that routinely collected clinical measurements can be used in early prediction of docetaxel-based treatment discontinuation due to adverse events in patients with mCRPC. From clinical point of view, these results can assist oncologists in clinical practice decision making and also future clinical trials design.

In publication IV, the aim is to address a fundamental oncological question: how reliable are the trial-based prognostic models in mCRPC everyday practice? Generally, real-world patients are older with more comorbidities compared to trial eligible patients. Accordingly, it is a challenging oncology task to predict the patient overall survival on the basis of trial-based prognostic models. Here, the investigated prognostic models resembled the state-of-the-art methods of the field published in two separate studies by Guinney et al. [10] and Halabi et al. [11]. The performance of the models was evaluated using a real-world mCRPC cohort from Turku University Hospital. The result of this paper confirms the applicability and generalizability of trial-tailored prognostic models in mCRPC routine practice.

In addition to developing methods for both molecular and clinical data in separate manner, I investigated the combined use of them in developing clinical predictive models as the final stage of my PhD studies. A complex disorder which has been heavily studied at both molecular and clinical levels is obesity. It is a major public health issue that has been proved to be influenced by both genetic and social-environmental factors. In publication V of this thesis, a novel data modeling strategy is proposed to predict adulthood obesity using both clinical and genetic risk factors. The results of this study confirm that combination of childhood clinical factors and genetic risk factors from the genome-wide association studies (GWAS) can significantly improve the prediction of adulthood obesity when compared to clinical factors alone.

The current thesis is organized in five parts: the first part is the introduction and study background. The second part (chapter 2) describes the methods and algorithms developed or used for publications I and II focusing on computational challenges at molecular level. Specifically, it gives an overview about transcriptomics data and the proposed novel computational approaches to analyze this type of data. The third part (chapter 3) covers our machine learning and statistical learning solutions to predict crucial events in metastatic prostate cancer. In contrast to previous part, purely clinical variables rather than molecular variables were available for these analyses. Chapter 3 includes literature review, details of proposed predictive models, as well as data and results from publications III and IV. The fourth part (chapter 4) describes computational approaches to deal with medical questions using both molecular and clinical variables. As a case study, chapter 4 briefly introduces data, methodology and results published in the fifth study included in this thesis. The last part (chapter 5) summarizes the novel proposed methods and clinically relevant findings of this work. The original publications included in this thesis are presented in the Appendix section.

2 Computational and Machine Learning Algorithms for Discovering Disease Markers from High-Throughput Gene Expression Data

In a broad point of view, the field of molecular biology investigates the role of biomolecules in regulation of cellular activities that defines phenotypes in living organisms. A central concept of the field is to understand gene expression process and transcriptional mechanisms to control this process. High-throughput sequencing techniques have revolutionized the scope of our knowledge in various biological aspects including gene expression quantification and profiling. At molecular level, this thesis focuses on computational methods for the analysis of transcriptomics data. Starting with biological background, this chapter provides an overview of data producing technologies and computational pipelines for the analysis of high-throughput transcriptomics data. The extended focus will be on statistical methods for measuring differential gene expression with RNA-seq data. Conducting a comparative study, the methodology behind the state-of-the-art tools for differential gene expression testing is elaborated. As a result, this study provides a practical guideline to assist RNA-seq users to choose the optimal method fitting to their experimental design. Additionally, a novel tool is proposed to perform RNA-seq differential expression analysis which aims to eliminate weaknesses of the field. The efficiency of our proposed method is confirmed in a complex biomedical application by identifying novel prognostic markers for ccRCC.

2.1 General overview of gene expression

The genome of living organisms carries the substantial heredity information encoded with primary biomolecules called deoxyribonucleic acid (DNA). Inside the cell nucleus, the massively long DNA strand is densely organized into a condensed structure called chromosome. Although almost all cells contain copy of the same DNA, different cell types vary in appearance and function. This difference is due to the fact that each cell turns on or expresses specific parts of DNA called genes whose overall activity determines the cell fate. Genes pass their flow of information to program a cell in two steps. The first step includes transcription and the gene coding information is used to form ribonucleic acid molecules (RNA). The RNAs can be either the initial template for the second step of gene expression called translation; or, can be the final products called non-coding RNA. The RNA which is used for translation step is called messenger RNA (mRNA) and encodes the final product of gene expression and primary element of cell functions the so-called proteins [12]. In most studies involving bioinformatics analysis, the aim to address the question of interest through identification, quantification and exploration of DNA, RNA and proteins. More specifically, cellular activity and behaviour can be studied either directly by investigating cell's proteins, or indirectly by exploring DNA or mRNA molecules as initiative and medium cell functioning products. In laboratory, high-throughput technologies are used to translate the genetic information from macromolecules of interest into quantified raw data [13]. The term omics is added to the obtained quantified data referring to the comprehensive investigative strategies to study biological entity of interest [14]. For instance, the term genomics refers to the comprehensive study of genome, while genome includes the whole set of genes obtained from sequencing of DNA molecules of a cell or tissue. The same approach is followed to explain other high- throughput omics datasets such as transcriptomics, proteomics, epigenomics and metabolomics. Transcriptomics is the data type used in publications I and II and so sections 2.2 and 2.3 give brief overview about the data and corresponding methodology.

2.2 Transcriptomics technologies

The heredity information encoded in genes need to be transcribed into RNA molecules or transcripts which serve as the mediator to control cell functions. Transcriptome is the full set of these transcripts made of RNA molecules in a cell, a tissue or an organism. The RNA molecules vary based on their functional roles. Examples include mRNA molecules as protein production templates, rRNAs as ribosomal assemblers, tRNAs as protein synthesis regulators or non-coding RNAs that have potential epigenomic impact on certain phenotypes such as disease development. Transcriptome has a dynamic nature and can vary under different conditions including intracellular stages or environmental circumstances [12]. The ultimate goal of transcriptomics is to identify and quantify all types of transcripts available in a cell of a desirable tissue or organism. The rapidly evolving high-throughput technologies have provided the possibility of transcriptome study via two primary approaches including DNA microarray and sequencing-based techniques.

With the history of more than two decades, microarray platforms serve as the initial tool in transcriptomics studies. This technique aims to detect intensity signals emitted from hybridization of predefined fluorescently labeled sequences to their complementary probes. The detected signals confirm the availability of transcripts and their intensities determine their corresponding expression levels. Microarray data highly rely on RNA molecules concentration and their binding affinities to complementary targets [15]. Due to their high-throughput capacity, microarray platforms allow cost effective analysis of thousands of probes simultaneously in a single experiment. However, the technology is superseding with sequence-based techniques due to several limitations. Microarray assays depend on prior knowledge about the genome of interest. Consequently, there is no possibility of investigating unknown genomes and detecting novel transcripts or fusion genes. In addition, they suffer from intrinsic properties which affect the reliability of differential expression analysis. These include high level of background noise due to cross-hybridization and inaccurate expression levels originated from sequence concentration and saturation biases [16]. Today, due to the significant decrease of costs and increase of precision in sequencing platform assays, number of newly produced microarray studies are declining dramatically [17] [18] [19]. Nevertheless, projects with the aim of data reusability still utilize previously produced microarray datasets either alone or in combination with sequencing data. Clinical diagnostic applications are also heavily relied on array-based platforms.

Next generation sequencing (NGS) techniques have become the standard approach for gene expression analysis in biology and medicine. The technology overcome microarray limitations while proving reproducibility and low technical variation in real-world applications [20]. RNA-seq is a highly sensitive NGS-based technique for detecting and measuring RNA molecules. In contrast to microarray techniques, RNA-seq does not depend on a priori known genome to perform the sequencing task. Instead, RNA-seq allows de novo assembly of transcriptomes in the absence of a well annotated reference genome [21]. RNA-seq data de novo assembly provides the ability to identify alternative novel splicing, fusion genes [22] and allelic variations within the transcribed regions [23]. The widely used platforms available for RNA-seq include Illumina, SOLiD, Roche 454, Ion Torrent and PacBio. Despite the variation in platform protocols, independent studies have confirmed relatively inter-platform and intra-platform concordance in identification and quantification of targeted RNA molecules [24]. RNA-seq platforms are free from predefined expression detection ranges, supporting their sensitivity in detection of very lowly or very highly expressed genes [19].

General approach in an RNA-seq workflow includes three main steps. The first step is called library preparation and starts by extracting and purifying RNA molecules from target cells or tissues. Next, the purified RNA strands will be randomly sheared into smaller fragments and then reverse transcribed into their original complementary DNA (cDNA) libraries. At this stage, some unique sequence strands called adapters will be ligated to cDNA fragments to help starting the RNA strand elongation during sequencing process. Finally, PCR (polymerase chain reaction) amplification techniques are used to produce sufficient copy of adapter ligated fragments required for sequencing platforms. The second step in RNA-seq workflow is to sequence the amplified cDNA strands using the desirable sequencing platform. Although there are notable differences in the chemistry of available sequencing platforms, the overall trend and the output of the platforms remain very similar. Here, the common output contains millions of short reads and their associated sequencing scores usually in regular simple formats like fastq files. The generated files from sequencing machines are the basic material for the third step in RNA-seq workflow which is computational data analysis. The overview of a typical RNA-seq data analysis pipeline is provided in the next section.

2.3 RNA-seq data analysis pipeline

The enormous amount of data produced by RNA-seq technologies require specialized computational tools to be analyzed and interpreted. A typical RNA-seq data analysis pipeline involves the following parts: preprocessing and quality control, reads alignment, summarization and normalization; and, finally differential expression testing followed by complementary data mining and data enrichment analysis (Figure 1). The experimental design and study subject are the main factors to determine the workflow details. Several statistical tools and computational methods have been proposed for RNA-seq data analysis. These bioinformatics tools are commonly built using Python, C and R programming languages. Majority of R-based software tools are freely available under Bioconductor project which is an open-source environment for computational biology. For instance, most of software tools utilized for data preprocessing and differential expression testing in publication I, are available as open source in Bioconductor.

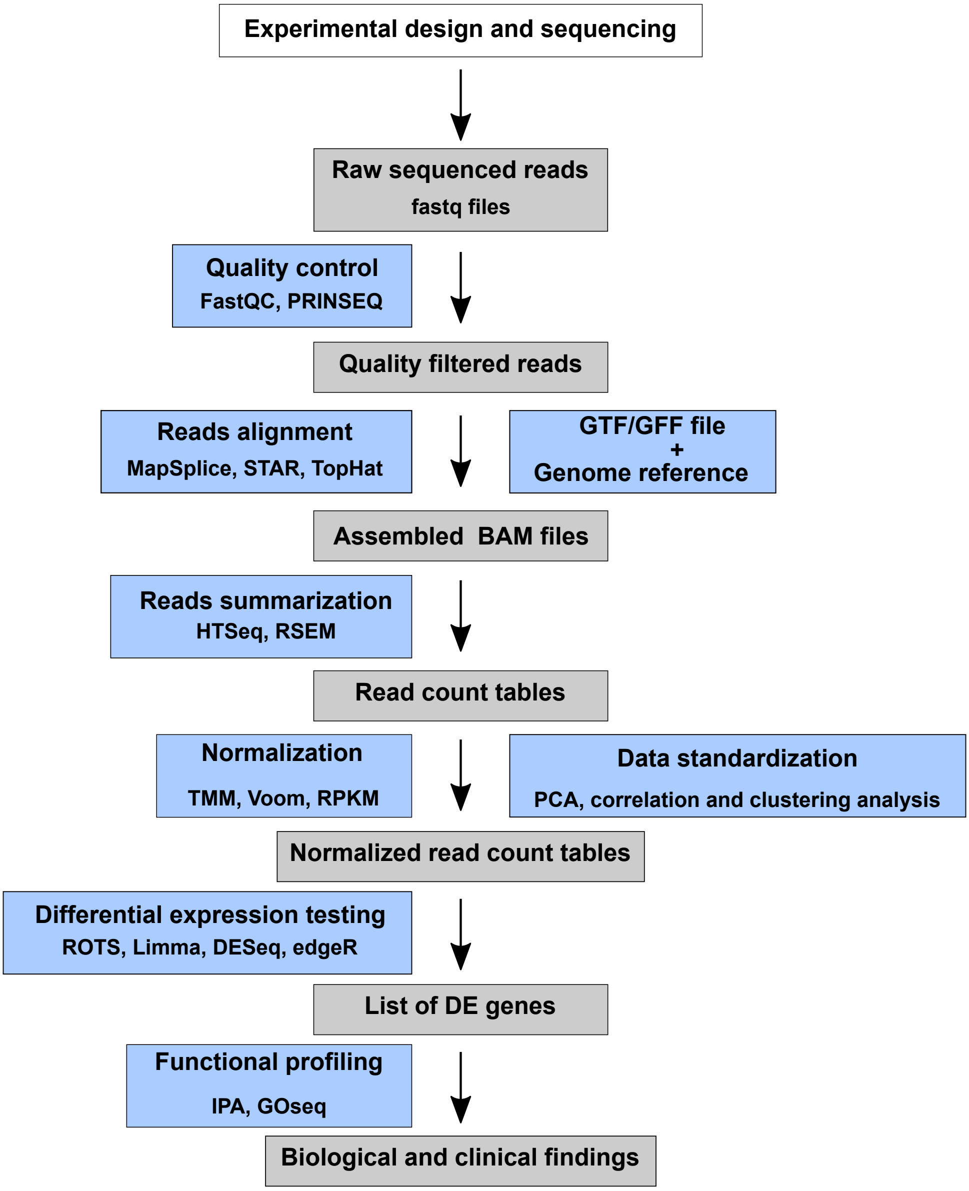

Figure 1. RNA-seq pipeline: the general workflow for the analysis of RNA-seq data. Blue boxes represent analytical methods and gray boxes represent data materials.

2.3.1 Sequenced reads quality control

Despite the considerable improvements, RNA-seq technology still suffers from systematic and random technical biases that affect the quality of sequenced reads. In order to obtain reliable downstream results, it is necessary to estimate the accuracy of data and correct for problematic detections. The quality control process checks quality scores per base calls, the PCR amplification problems, guanine-cytosine (GC) content distribution, presence of adapters and high rate of short length motifs (k-mers) and duplicated reads detection. Several bioinformatics tools have been developed to address each of the mentioned potential error sources [25][26]. FastQC[1] is a common and user-friendly tool to generate quality control report from sequenced raw reads supporting by visualization graphs. MultiQC [27] is also a modular reporting tool which aggregates results from different analytic softwares into a single output. There are other application tools such as NGS QC Toolkit [28] that can be used for both quality check and quality improvement in NGS data. In addition to possibility to discard low quality reads and trimming adapters or poorly sequenced bases, NGS QC Toolkit contains format conversion and statistical tools to facilitate the general data analysis workflow. In publication I, the FastQC tool is utilized for controlling the quality of raw fastq files.

2.3.2 Read Alignment

Following a typical RNA-seq pipeline, read mapping or alignment is the next step after quality control. Ideally a mapping algorithm intends to map all short-sequenced reads into a unique and identical location of a genome or transcriptome reference. However, this goal is not fulfilled in practice due to several intrinsic biological facts including alternative splicing, genome repetitive or mutated regions and existence of pseudogenes. In addition, RNA extraction method and PCR procedure have been proved to affect the performance of alignment algorithms [29]. Various alignment tools have been developed to address the available challenges, yet none of them are able to map all the input sequenced reads. MapSplice [30], ReadsMap , STAR [31] and TopHat2 [32] are among the most popular developed alignment softwares. With most of the available methods, the percentage of identical mapped reads is a determinant factor in evaluating the accuracy of results. TopHat2 and ReadsMap were used in publications I and II respectively. For publication I, the results from alignment step were acceptable if more than 70% of the reads were mapped to non-redundant regions [33]. Recently, TopHat2 has been replaced by a new tool called HISAT2 [34] which was not available at the time of our studies.

2.3.3 Expression summarization and quantification

Once the alignment phase is accomplished, the next step is to estimate the expression abundances by quantifying the number of mapped reads associated to genes or

[1] Andrews S. (2010). FastQC: a quality control tool for high-throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

transcripts. Here, the most challenging issue is to deal with multireads that have been mapped to multiple transcripts. The simplest approach when encountering with multireads is to discard them. This approach has been used even in some widely used methods like HTSeq [35] which discards reads mapped to a specific number of regions determined by user. This way, a noticeable proportion of information will be lost which results in potential systematic biases in downstream analysis. Alternatively, some recently developed methods have tried to accommodate the challenge by replacing the alignment step with mapping step while accounting for multi-mapped reads (e.g., Salmon [36] and Kallisto [37, p.]). The current recommended state-of-the-art methods include Cufflinks [38], RSEM [39], Salmon [36], and Sailfish [40] which have proved rather similarly acceptable performance in independent comparative studies [41]. In this thesis, Cufflinks, HTSeq and RSEM have been extensively used to obtain the expression levels.

2.3.4 Preprocessing and normalization methods

Normalization is an essential part of RNA-seq analysis workflow ensuring the accuracy of downstream analysis. Normalizing techniques aim to remove (or highly reduce) the data systematic technical biases and provide a uniform comparative circumstance within and between sequenced samples [42]. Most considerable within sample biases include gene length, RNA composition and GC content. With gene length bias, it is simply expected that longer genes comprise more sequenced reads compared to shorter ones which could provide more statistical power in false detection of long genes as differentially expressed genes. In particular, this can lead to misleading biomedical conclusions when performing enrichment analysis without a proper normalization strategy. RNA-seq technology is also biased towards detection of fragments with low or rich GC content [43]. More specifically, sequencing results from Illumina platforms are prone to under-represent the reads with GC content [44]. Several normalization methods exploiting quantile regression and linear modeling approaches have been proposed to account for RNA-seq GC content [45][46]. However, some model developers state that within sample effects can be ignored due to their consistency within all experimental samples.

In case of differential expression analysis, state-of-the-art methods aim to standardize data by removing between samples biases. Here, sequencing depth, the determinant of total counts per sample (library size) is the most critical bias origin. When extracting RNA strands for sequencing, samples with more RNA molecules will end up with more reads compared to other samples; introducing rise to false positive differential expression detections. RPKM (Reads Per Kilobase per Million mapped reads) is a classic but still widely used method which provides an integrated solution for both gene length and library size biases. For gene g the RPKM value R_g is calculated as

$$R_g = \frac{10^9 n_g}{N L_g}, \tag{1}$$

where n_g is the total number of reads mapped to gene g, N is the total library size of the sample and L_g is the length of gene g. Currently, RPKM values are most often used for expression level summarization rather than normalization similar as in the approach in publication I.

In addition to library size, a practical normalization method has to tackle with skewness of read counts distribution in the presence of extremely varied expressed features before running any statistical tests. For instance, in experiments with few highly expressed genes, the proportional distribution of the reads for the remaining genes can be imbalanced or underestimated. In this thesis, TMM (Trimmed Mean of M-values) [42] method has been frequently used to reduce biases originated from RNA composition. TMM method assumes that the majority of genes are equally expressed among sample groups and aims to compute TMM scaling factors for each sample after trimming the genes with extreme absolute expression levels. To do so, one sample will be chosen as reference and the log ratios between other samples and the reference for all genes will be calculated. More specifically, the log ratio of gene g from sample i to reference sample r is calculated as

$$M_{g,i}^{r} = \log_2 \frac{q_{g,i}/N_i}{q_{g,r}/N_r}, \tag{2}$$

where $q_{g,i}$ is the observed expression level of gene g from sample i; and N_i and N_r are the total library size in sample i and reference sample respectively. At this stage, 30% of the most extreme M_g values will be trimmed, and the weighted mean of the remaining log ratios will be used to calculate TMM scaling factors. This whole procedure is implemented in calcNormFactors() function in edgeR [47] Bioconductor package.

In publication II, we employed log-counts per million (log-cpm) transformation implemented in voom (variance modeling at the observational level) [48] trend from Limma [49] package. The log-cpm method aims to transform the discrete distribution of count data to better fit the normal distribution; so that the normal-based statistical methods become applicable in RNA-seq data analysis. For each experimental sample the expression levels of genes are transformed to log-cpm values

$$y_{g,i} = \log_2 \left(\frac{n_{g,i} + 0.5}{N_i + 1} \times 10^6 \right), \tag{3}$$

where $n_{g,i}$ is the total reads mapped to gene g from sample i and N_i is the total library size for sample i. Voom is more a non-parametric analysis approach rather than a normalization technique. It provides an analytic platform to model the data variance and facilitates downstream statistical analysis. Independent studies suggested that statistical methods which use voom transformation and TMM in their internal procedures produce comparably reliable results [50] [51]. Also, in publication II, comparably robust identifications were obtained when performing differential expression analysis over kidney cancer data.

2.3.5 Differential expression testing

In RNA-seq experiments, the most common task is to identify transcribed genomic regions that are differentially expressed across two or more distinct biological conditions. The transcribed genomic regions can be exons, transcripts or genes. In this thesis, different genomic regions will be referred as genes for convenience. Bioinformatics community has proposed a considerable number of tools to perform differential expression analysis. A well performing tool has to optimally assess the technical and biological variations available in RNA-seq data. In general, differential expression testing tools can be divided into two categories of parametric and non-parametric methods. As shown in publication I, non-parametric methods present limited statistical power in experiments with few replicates. A frequently used example of this class of methods is NOISeq [52] which produced very conservative results with lowest precision among tested methods in our comparative study. It is a data-adaptive method which performs differential expression testing in two steps. First, it empirically models the so-called *noise* distribution of expression levels out of the actual data by contrasting fold change differences and absolute expression differences of all available genes among samples within the same condition. In the second step, the modeled noise distribution is used to justify the significance of observed differences between the sample groups and distinguish the true positives.

The early parametric differential expression testing methods, proposed Poisson models to account for RNA-seq discrete data. This class of methods fitted well to studies with technical replicates and low level of between-samples variability [20]. When taking a random read from an experiment, the probability that the taken read comes from gene g from sample i can be calculated as

$$\pi_{gi} = \frac{n_{gi}}{N_i}, \tag{4}$$

where n_{gi} is the expression level of gene g (proportion of counts mapped to gene g) from sample i and N_i is the total mapped reads obtained for sample i (library size). Assuming that the sequenced reads have been sampled independently, the sequencing process can be modeled as a simple random sampling from the binomial distribution (success: if the sequenced read is from gene g, failure: otherwise). Accordingly, the number of reads mapped to gene g from sample i, Y_{gi}, can be modeled as follow:

$$Y_{gi} \sim B(N_i, \pi_{gi}). \tag{5}$$

In the absence of biological replicates, considering that N_i is always very large and $\pi_{gi} \ll 1$, the above distribution could be very well approximated by a Poisson distribution with mean and variance both equal to $\pi_{gi}.N_i$. Previous studies, have successfully confirmed the practicality of Poisson distribution in RNA-seq experiments with technical replicates [20] [53] . However, soon it was discovered that the Poisson models underestimate the elevated variability, referred as overdispersion [54], observed in RNA-seq data with biological replicates. In this case, the estimated values for π_{gi} vary between samples within same experimental group and an additional parameter is required to account for over-dispersed biological data (mean $>$ variance) and better controlling of type I error. Hence, several commonly used tools employ negative binomial models (e.g. edgeR, DESeq [55]) or beta binomial models (e.g. BBSeq [56]) to estimate the distribution of expression levels.

To model the expression levels as negative binomial distribution, the variance of Y_{gi} is computed as a function of mean μ

$$var(Y_{gi}) = \mu_{gi}(1 + \phi_g \mu_{gi}), \tag{6}$$

where $\mu_{gi} = \pi_{gi}.N_i$ and ϕ_g is the dispersion for gene g. Accordingly, one can estimate the distribution of Y_{gi} is follows:

$$Y_{gi} \sim NB(\mu_{gi}, \phi_g), \tag{7}$$

where in case of $\phi_g = 0$, then the negative binomial distribution is converted to Poisson model.

Once the parameters are estimated, the likelihood ratio test is performed to compare the likelihood of the read counts with no differential expression under condition of interest (e.g., disease) against the likelihood of the data representing differential expression due to the same condition. In this case, thousands of hypotheses are tested simultaneously and so correction for multiple testing is a necessity to avoid encountering high rate of false positives. Therefore, most tools including ROTS, DESeq2 and Limma correct the obtained p values for multiple testing before reporting the final results.

DESeq is a widely used parametric method in RNA-seq data analysis. Independent comparative studies have confirmed the strength of DESeq in precise estimation of variance especially in relatively highly or lowly expressed genes. Also, DESeq is capable of estimating data parameters in datasets with no replicates using a so-called *blind* method. However, end users are warned about the potentially overly restrictive results that may be drawn from experiments with no replicates.

In publication I, we concluded that Limma method outperforms other state-of-the-art packages under a wide range of experimental designs tested in our study. Limma was first used to identify differentially expressed genes in microarray experiments but over the time, it has been modified so that it is compatible with other data types such as RNA-seq data. In contrast to parametric methods such as DESeq, Limma aims to model the mean-variance relationship rather than introducing the additional dispersion variable under negative binomial distribution assumption. This package provides two approaches to detect differentially expressed genes: the Limma-trend and the voom-trend [48]. The main difference between the trends is that the mean-variance modeling is performed at gene level for all samples in Limma-trend; while, the voom-trend models the variance at sample level individually.

To control the library size bias, voom transformation is used to convert the raw expression levels to log-cpm values (refer to 2.3.4 for details). RNA-seq counts have shown to be heteroscedastic meaning that the variance fluctuation increases with increasing expression levels. Using log-cpm transformation, larger variances would be moderated with higher rate, stabilizing variance for relatively long or highly expressed genes. When the data is preprocessed, the mean-variance relationship is modelled using an empirical Bayes analysis pipeline. The voom-trend includes an additional step in which, it calculates the precision weights for each individual sample and incorporates these weights into the Limma analysis pipeline [57]. The expression levels are fitted into a linear model implemented in very similar way for both Limma-trend and voom-trend:

$$Y_{gi} = \beta_1 x_i + \beta_0 + \epsilon_{gi}, \tag{8}$$

where β_0 and β_1 are the model slopes, ϵ_{gi} is the model error and x_i is the differentiating covariates for experimental groups. Limma is freely available as a Bioconductor software package and the developers actively answer technical questions via Bioconductor community mailing list. The reader is referred to publication I for detailed information about the software packages available in our comparative study.

2.4 ROTS

In publication I, we focused on differential expression analysis and demonstrated that the performance of the state-of-the-art methods strongly depends on the data under analysis, and the field lacks a powerful method with capacity of fitting to real data. With parametric methods, opting a suitable distribution which can fully capture the data heterogeneity seems a challenging task. To address the challenge, in publication II we propose our data-adaptive method called ROTS [58] [59] [60]. ROTS method avoids making a priori assumption about the data distribution by employing a test statistic learning directly from the data. More specifically, we utilize a dynamic t-type statistic as a core to optimize the reproducibility of k top-ranked candidate genes from group-preserving bootstrap datasets. Here, each resampling round and the corresponding list of top-ranked genes provide information to optimize the statistic parameters.

Suppose that x_{gi}^{j} represents the normalized expression level of gene g in sample i from condition j, ROTS estimates the mean and variance of gene g within each condition respectively as follow:

$$\bar{x}_g^j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{gi}^j \text{ and } \left(s_g^j\right)^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(x_{gi}^j - \bar{x}_g^j\right)^2, \tag{9}$$

where n_j is the number of samples in condition j. Next, the rank of gene g is calculated using the following modified t-statistic:

$$d_\alpha(g) = \frac{|\bar{x}_g^1 - \bar{x}_g^2|}{\alpha_1 + \alpha_2 s_g}, \tag{10}$$

where $\alpha_1 \in [0, \infty)$ and $\alpha_2 \in [0, 1]$ are the common parameters to be optimized, and s_g represents the estimated pooled standard error:

$$s_g = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)\left(s_g^1\right)^2 + (n_2-1)\left(s_g^2\right)^2}{n_1+n_2-2}}. \tag{11}$$

ROTS statistic is characterized by optimization and specification of α parameters. The predefined cases of $d_\alpha(g)$ include signal log-ratio ($\alpha_1 = 1, \alpha_2 = 0$), ordinary t-statistic ($\alpha_1 = 0, \alpha_2 = 1$) and SAM-statistic ($\alpha_2 = 1, \alpha_1$ a percentile of the standard deviations) [61]. Unlike these predefined forms, the ROTS statistic is determined directly from the input data through a reproducibility optimization procedure ensuring the accurate estimation of biological and technical variability. Accordingly, we optimize $d_\alpha(g)$ by maximizing its reproducibility ($\max\{R_k(d_\alpha(g))\}$) through pairs of bootstrap datasets estimated as follow:

$$R_k\,(d_\alpha) = \frac{1}{B}\sum_{b=1}^{B} R_k^b\,(d_\alpha), \tag{12}$$

where B is the predefined number of bootstrap rounds. Assuming that D_1^b and D_2^b are the bootstrapped pairs resampled from the original dataset D (with replacement while preserving the groups), one can calculate R_k^b as follow:

$$R_k\ (d_\alpha) = \frac{1}{B}\sum_{b=1}^{B}\frac{\#\left\{g \mid (r_g(\alpha, D_1^b) < k, r_g(\alpha, D_2^b) < k)\ \right\}}{k}, \tag{13}$$

where $r_g(\alpha, D_{\mathrm{i}}^b)$ is the rank of gene g in D_{i}^b dataset and # denotes the number of genes that fulfill the above condition. Finally, we transform the reproducibility into a z-type statistic and maximize it to determine the optimal ROTS statistic:

$$Z_k(d_\alpha) = \frac{R_k(d_\alpha) - R_k^0(d_\alpha)}{s_k(d_\alpha)}. \tag{14}$$

Here, we define $R_k(d_\alpha)$ and $R_k^0(d_\alpha)$ as the reproducibility of the observed and random datasets, and the dominator,$s_k(d_\alpha)$, is the estimated standard deviation of the bootstrap distribution of the observed reproducibility. $Z_k(d_\alpha)$ is maximized over a dense lattice of α parameters $(\alpha_1,\ \alpha_2)$ and k as the optimal number of top ranked genes. For publication II, we defined the parameters so that $\alpha_1 \in \{0, 0.05, \dots, 5\}$, $\alpha_2 \in \{0,1\}$ and $k \in \{5, \dots, G\}$ where G denotes the total number of genes available in the experiment.

To perform analysis, ROTS takes matrix of normalized data for input (columns: experimental samples, rows: genomic features) along with vector of sample labels and groups. For each gene, the package reports ROTS statistic, fold change, p value and FDR estimate for assessing differential expression. Additionally, the general validity of the analysis depends on the optimized *Z*-score and reproducibility values. In practice, the reliability of detections cannot be confirmed in experiments with *Z*-score below 2. The ROTS package and detailed manual is freely available through Bioconductor: (http://bioconductor.org/packages/ROTS).

2.5 Functional enrichment analysis and biological interpretation

Once the list of differentially expressed genes is obtained, it is essential to biologically characterize the concluding list and infer their associative role with the study subject. For example, we are interested to see how and to what extent the differentially expressed genes or their interactions would affect the biological condition under study. In RNA-seq data analysis pipelines, investigating the functional significance of differentially expressed genes can be performed using two different approaches. The first standard approach is gene category over-representation analysis which classifies the original list of differentially expressed genes into smaller *gene sets* based on their association with a priori annotated pathways; and, determines whether a significant over-representation of differentially expressed genes in certain pathways is observed. Here, appropriate statistical tests such as hypergeometric test are applied to compare the significance of difference between distribution of detected differentially expressed genes and distribution of all sequenced genes within a desirable gene set. The most commonly used a priori annotated pathway databases include Gene Ontology (GO) [62], Database for Annotation, Visualization an Integrated Discovery (DAVID) [63] and Kyoto Encyclopedia of genes and genomes (KEGG) [64]. In publication II, the Ingenuity pathway analysis (IPA), a commercial web-based tool, was used for functional enrichment analysis. IPA includes both automated and manually annotated ingenuity ontology to identify the list of over-represented gene sets and can be applied to all types of high-throughput omics data. In publication II, our objective was to identify significantly changed pathways with potential impact on developing severe form of ccRCC with poor prognosis. To this aim, the list of differentially expressed genes detected by ROTS was uploaded to IPA software; and the Core Analysis function was applied to identify pathways and gene networks relevant to poor prognosis. The significance of detections was tested by the Fisher's exact test p value and the final pathway candidates were scored on the basis of number of genes over-represented in each pathway (see publication II, supplementary table S3).

Gene set enrichment analysis (GSEA) is the alternative popular approach to detect the enriched functional gene sets. Similar as in previous enrichment profiling approach, a gene set is a module of genes that share common functions with underlying impact on certain pathways or diseases. GSEA-based methods rank the differentially expressed genes or transcripts according to their differential expression measurements (e.g. FDR and fold change), and then run Kolmogorov family tests to assess their enrichment significance level. Several GSEA-based software tools like SeqGSEA [65] combine differential expression information and splicing strength for each gene before adopting the statistical tests to improve the analysis outcome [66].

2.6 RNA-seq databases and public repositories

Today, most of high-profile journals require the study materials to be publicly available by authors before publishing their research results. This policy ensures reproducibility of the results and provide a valuable source to the whole research community. In bioinformatics, data serves as the fundamental necessity in method development and due to their cost effectiveness, many bioinformatics publications depend on data sources stored in public data portals. Currently, many public data repositories have been developed to fulfill the growing demands of data sharing and reusability. While some databases are specialized in one type of data or a specific organism (e.g. Integrated Microbial Genomes (IMG) for microbes), many of them include multiple omics data levels focusing on several topics like various cancer types or mental diseases. In practice, it is very common for computational projects to demand combining data from different databases which can cause technical challenges due to the databases' inconsistencies. In this case, the technical challenges include different naming conventions used by different databases, incomparable data preprocessing approaches (in case of absence of raw data) or significant variation in data collection time which can yield to utilization of different data format and production technologies. Accordingly, most data repositories use an embedded or stand-alone tool to automatically homogenize data from different resources. BioMart application provided by Ensembl is one of the widely used tools to address some of the challenges originated from databases inconsistency. The studies in this thesis utilized data from public data portals including ReCount [67], TCGA, European Genome-phenome Archive (EGA) [68] and Project data Sphere [69]. Among the utilized data portals by this dissertation, TCGA is perhaps the most studied data repository in computational biology projects. It is a joint program between the National Cancer Institute and the National Human Genome Research Institute which was launched in 2006. TCGA utilizes high-throughput sequencing techniques to produce multi-level genomic data for more than 30 types of most prevalent cancers with poor prognosis. The program ultimate goal is to develop a multi-dimensional comprehensive atlas of genomic profile of selected cancers. In addition to biomedical sciences, TCGA is an invaluable resource in bioinformatics as a source of real-world data with huge sample size, freely available for research purposes. The large volume and diversity of the TCGA datasets can significantly assist in developing accurate and well-tolerating tools when coping with heterogenous data [70]. For transcriptome studies, the NCBI Gene Expression Omnibus (GEO)[71] and the EBI ArrayExpress (AE) [72] also provide public repositories for gene expression data.

2.7 Differential expression testing comparative study: Publication I

As described in section 2.3.5, in most RNA-seq experiments the primary goal is to detect the differentially expressed transcript products. Validation and reproducibility of this type of analysis ensure the implementation of utilized statistical approaches in real-world conditions. Bioinformatics community has proposed a considerable number of tools to perform differential expression analysis. However, in practice, there is no clear guideline about how to select the best fitted methods to organize an optimal experimental design and obtain reliable results. Of note, several comparative studies including publication I have demonstrated the determinative impact of choice of method in final reporting results [73][74][75]. This publication is very important as it shows a significant inconsistency between results from different pipeline analysis. It also provides a practical guideline for selecting the right method for organizing an optimal RNA-seq experiment. The main advantage of our study over the preceding similar works is the utilization of two relatively large real datasets with biological replicates rather than either small or simulated data. Under optimal experimental design conditions, it is expected to obtain a close set of differentially expressed features using different computational methods. However, our analysis demonstrated a notable level of variation and inconsistency among the significant detections of methods under study.

2.7.1 Methods and material

Data. Two datasets from mouse and human experiments (with relatively large sample sizes at the time of this study) from publicly available repositories are employed in publication I comparative analysis. The mouse RNA-seq data consists of the striatum samples of 21 mice, 10 of the C57BL/6J strain and 11 of the DBA/2J strain [76]. Gene expression profiling was performed to identify differences between the two strains. The human data includes lymphoblastoid cell lines of 56 unrelated Nigerian individuals, 28 males and 28 females [77]. We identified differentially expressed genes between males and females. Both datasets were sequenced using Illumina Genome Analyzer II and TopHat aligner with default parameters was used to assemble the sequenced reads. For all comparisons except CuffDiff2, the genes abundance levels were inferred using HTSeq software. In mouse data, samples are divided into more homogenous groups with significantly higher within-strain correlation values (Wilcoxon test, $P < 0.01$) compared to human data.

Methods. This study presents a systematic practical comparison of eight state-of-the-art software packages including edgeR [47], DESeq [55], baySeq [78], NOISeq [52], SAMseq [79], Limma [49], Cuffdiff2 [80] and EBSeq [81]. The performance of the methods was assessed using 1) the number of significant detections for different sample sizes, 2) detections consistency within and between methods, 3) false discovery detection rates and 4) methods running times. Data preprocessing and normalization was performed following the approaches recommended by the software manuals. In order to investigate the relation between gene expression level and their appearance in differential expression results, genes were divided into four categories on the basis of their RPKM values: very lowly or not expressed, lowly, medium or highly expressed genes. The Bioconductor package easyRNASeq [82] was used to calculate the RPKM values and the categories were organized as follow: genes with an average RPKM value across samples below 0.125 were considered as very lowly or not expressed, genes with an average RPKM value between 0.125 and 1 were considered lowly expressed, between 1 and 10 medium expressed, and above 10 highly expressed [83].

2.7.2 Results and conclusions

As expected, with most of the software tools, we observed a positive correlation between the number of detected differentially expressed genes and the number of replicate samples (exceptions included NOISeq and Cuffdiff2) which implies the auxiliary effect of large sample size when estimating dataset statistics. Despite of this similar increasing trend, a remarkable variation was observed when comparing the number of detections at each sample size from different packages. For instance, when considering the whole sample size, number of significant detections varied from zero to few hundreds or couple of thousands in mouse and human datasets respectively (Figure 2A). Regarding the expression levels of detected genes, in both human and mouse datasets a relatively high proportion of detected genes belonged to very lowly and lowly expressed categories (Figure 2B). This observation can originate from technical limitations of RNA-seq technology in accurate quantification of expression levels especially for lowly expressed genes. Currently, filtering out the very lowly expressed genes is a common approach in practical RNA-seq studies to avoid expression level-based biases.

Finally, we examined the within methods consistency of the detections for both human and mouse datasets. Accuracy of variance estimation is a key factor ensuring the reliability the of statistical testing. However, with many of RNA-seq studies, variance estimation has remained a challenging issue due to few or no replicates condition. In this publication, we assumed that detections from complete datasets comparisons should have highest precision and so were used as validating reference to check the consistency of the results from smaller sample size comparisons. Accordingly, the precision of the methods was calculated as the overlap between detected genes in the subset of different sizes and significant detections in the complete data for both human and mouse datasets (Figure 2C). Robust methods are expected to detect relatively consistent features for different subsample sizes with lower level of variation in number of significant detections. However, with most of software packages (except Limma and DESeq) our analysis revealed the necessity of algorithm improvements especially with the more heterogenous human dataset.

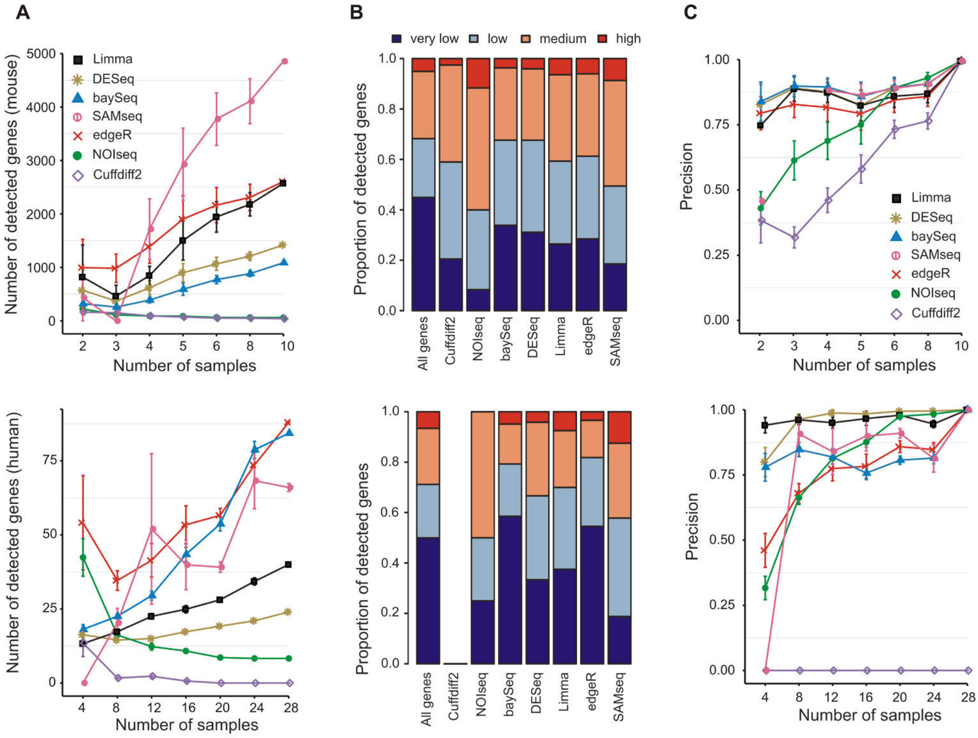

Figure 2. The effect of sample size and expression level on differential expression analysis using eight software packages in mouse and human data (upper and lower panel respectively). (A) Number of differentially expressed genes (y axis) with different sample size (x axis) for each software tool. The points correspond to average number of significant detections and the error bars represent the standard deviation. (B) The proportion of differentially expressed genes with respect to their expression abundance categories: very lowly expressed, lowly, medium and highly expressed genes. (C) Precision of calling differentially expressed genes (y axis) with regard to sample size (x axis). Number of detections in the complete data (N = 10 and N = 28 for mouse and human data respectively) were used as the reference to calculate the precision of methods. The points correspond to averages over 10 randomly sampled subsets; the error bars show the standard error of the mean. The figure is adapted with permission from publication I.

Methods' power in controlling type I error is another key metric to assess the validity of differential expression analysis findings. Calculating the false positive discovery rate is not straight forward for real datasets when the truth of the results is unknown. For this publication, we suggested performing artificial differential expression analysis within the same sample groups (i.e. same mouse strain or same human sex). The comparative arms included 10 times random sampling without replacement for different numbers of replicates for each sample group. A powerful method is expected not to detect any significantly differentially expressed features in such mock experimental design. In order to make comparable results across different methods, for each sample size, number of mock detections were divided by the average number of detections in the actual comparisons. In general, false discovery detections decreased when the sample size increased and the methods' performances were more promising with mouse dataset (Figure 3). We concluded that in the absence of data distribution knowledge, more restrictive false discovery rate control methods and cautious when reporting the results are needed especially with more heterogeneous datasets (e.g. human dataset in this study).

Following the significant variation observed between the methods' detections, additional analysis was performed to investigate the methods' similarities and differences. More specifically, for each method, the genes were ranked based on their level of significance in differential analysis and Spearman correlation values were calculated to reflect the methods' relation. This type of analyses can provide a useful guideline for studies which use more than one method to validate their findings. As could be expected, methods that shared similar data distribution assumptions were highly correlated. After a carful comparative investigation, we suggested Limma method as a robust tool to perform differential expression analysis.

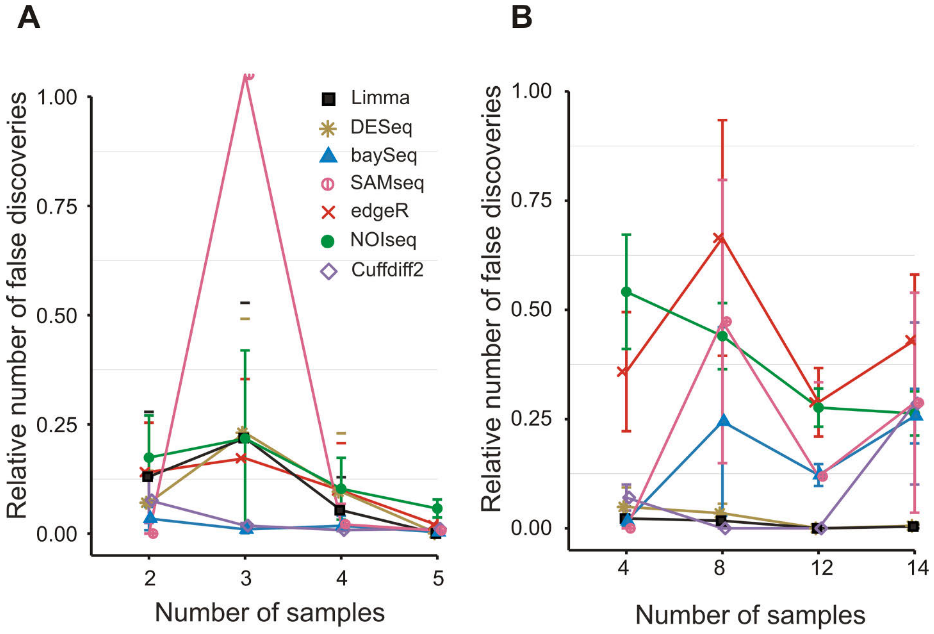

Figure 3. False discovery rates for the eight software tools on the basis of mock comparisons in (A) mouse and (B) human data. The points correspond to averages over 10 randomly sampled subsets; the error bars show the standard error of the mean. The figure is adapted with permission from publication I.

2.8 Practical issues and potential pitfalls of RNA-seq differential expression analysis

In publication I, the aim was to guide the selection of a suitable package to ensure the validity of differential expression analysis in terms of accuracy and robustness at various sample sizes. It was shown that different methods can produce widely contrasting results under similar study design. Of note, most of the investigated methods were sensitive to number of replicates meaning represented inconsistent results when the study sample size varied. In this comparative study, Limma and DESeq presented persistent results for different sample sizes and more importantly when the sample size was below 5. In addition to sample size, methods presented different levels of sensitivity when the input data was relatively heterogeneous (e.g., human dataset in current analysis). In fact, with RNA-seq data, the main difficulty that method developers are facing is to accurately estimate the variance. In this case, we are dealing with three main sources of variability including technical variation, biological variation and the systematic NGS inherent biases. The technical and biological variations originate from experimental replicate types. As we are dealing with count data, we can safely use Poisson distribution to model technical replicates and simply assume that the variance of samples is equal to their mean. However, this assumption does not hold when the biological replicates are introduced to the analysis. In experiments with biological replicates, the main challenge is to accurately estimate the variance of over-dispersed data. A parametric solution to account for biological variation is to use negative binomial distribution which poses additional parameter for variance. Although parametric methods with negative binomial distribution outperform Poisson bases methods, their accuracy in estimation of biological variation in experiments with relatively large sample size (e.g., sample size ≥ 10) is under question. For instance, edgeR is a popular method which utilizes negative binomial distribution to model count data. However, several studies have confirmed that the ability of edgeR in controlling type I error decreases when the sample size increases [73] [51] [84]. Here, one solution is to apply non-parametric approaches with no a priori assumption about the variance.

Majority of differential expression methods employ normalization and preprocessing techniques to account for NGS inherent biases [85]. Filtration of very lowly expressed genes with the aim of improving the statistical power is a common preprocessing practice [86]. This is mainly because most methods assume a large proportion of data noise is available among the lowly expressed genes as a result of RNA composition bias. Although several studies confirm the necessity of filtration operation, careful considerations should be taken into account to minimize the loss of potentially interesting expression changes. In our comparative study, we concluded that choice of normalization method does not have a strong impact on the final results (refer to Supplementary Figure S4 and Table S1, publication I for details). However, it should be mentioned that evaluating the effect of normalization techniques was not the main aim of our study and so further complementary analysis are required to rule it out.

Under special circumstances that the choice of suitable method is ambiguous, a typical approach is to perform differential expression analysis using couple of selected methods and take the overlapping results for enrichment analysis. Although this can be an efficient solution for some scenarios, it is very probable to obtain none or very few overlapping detections which is a big barrier for further analysis steps (refer to Figure 4 publication I for details). In publication I, it was concluded that Limma method is the safest choice among other tested methods based on our study design. Nevertheless, it was emphasized that the field lacks a standard method capable of creating reproducible and robust results under wide range of experimental design. To address this essential demand, ROTS method is proposed in publication II as a data-adaptive and powerful approach to perform differential expression analysis.

2.9 Robust identification of prognostic markers for ccRCC: Publication II

In publication II, a comprehensive study was launched to assess the efficiency of ROTS in real-world medical applications. In particular, I aimed to extract a robust and reproducible gene signature for risk stratification of patients with ccRCC. In cancer studies, prognostic markers are used to estimate the tumor progression and patient survival outcome. They are the principle factors in clinical decision making and clinical trials eligibility criteria. Accordingly, identification of robust biomarkers supported by biological validity play key role in development of efficient and personalized therapeutic agents.

Renal Cell Carcinoma (RCC) is the most common type of kidney cancer accounting for more than 90% of the diagnosed cases. With a steady rise in annual incidence, RCC is the seventh and ninth leading cancer in men and women respectively. Locally it originates from renal epithelium with high metastatic potential especially with possibility to spread into lung, liver, bone and lymph nodes system. Most of the non-metastatic diagnosed cases can benefit from surgical treatments with five years overall survival. However, the prognosis is poor for later stages. In respect of biology and genetics, RCC tumors are proved to be highly heterogenous with distinct histological subtypes. This genetic diversity is the underlying factor for necessity of subtypes-specific therapeutic strategies and consequently the disease outcomes. The most frequent histological subtypes of RCC include clear cell (75-80%), papillary (10-15%) and chromophobe (<5%). In clinical practice, it is still a challenge to predict patient progression and decide on proper treatment options. In this thesis, a novel data-driven approach was proposed to identify prognostic markers for ccRCC, the most common type of RCC with the highest morbidity rate.

2.9.1 Materials and methods

Spike-in data. The spike-in data were downloaded from GEO with accession number GSE49712. The RNA group samples A (pooled cell lines) and B (human brain) were from SEQC (MAQC- III) project spiked with 92 synthetic RNA molecules from the External RNA Control Consortium (ERCC) [1]. The ERCC spike-in controls were spiked to have 0.5, 0.67, 1 or 4 fold changes between the mixtures groups A and B. Samples were sequenced using Illumina HiSeq 2000 platform and TopHat (V 2.0.3) and HTSeq (V 0.5.p3) were applied for read alignment and expression level estimation respectively.

ccRCC datasets. The ccRCC datasets were from TCGA [70] [88] and a Japanese cohort published by Sato et al. [89]. In particular, we were interested in datasets with large heterogeneity and sample size to examine ROTS performance. Of note, the selected datasets were among the largest datasets publicly available at the time of our study (2015). The TCGA data (N = 442) was used to detect candidate prognostic markers and the dataset by Sato et al (N = 93; referred to as validation data in the following) was used as a completely independent data to validate the results. The TCGA dataset was directly downloaded from TCGA portal and the Japanese ccRCC cohort was downloaded from EGA under the accession number EGAS00001000509. The reader is referred to the original papers for the detailed information regarding the ccRCC datasets.

2.9.2 Results and conclusions

Spike-in dataset analysis. ROC (Receiver operating characteristic) curves and AUC (Area under the ROC curve) values are widely used to evaluate methods' performance in detection of differentially expressed features. In this study, calculating ROC analysis parameters was straight forward as a spike-in dataset with known differentially and non-differentially expressed features was utilized to assess ROTS performance over the selected state-of-the-art methods of the field. Accordingly, true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) were calculated by comparing significant detections of each method and a priori known ERCC spike-in controls. Next, ROC analysis was performed, and AUC values were calculated to assess the methods' sensitivity and specificity in detection of differentially expressed features. The results showed the outstanding performance of ROTS comparing to other benchmarking methods (Figure 4A; AUC

= 0.941; DeLong's test P < 0.01 compared to all other methods except for baySeq for which P = 0.077). Non-differentially expressed ERCC controls (fold change = 1) were further used to test methods power in controlling type I error rate. It is expected that methods report high FDR values (i.e., FDR > 0.05) for non-differentially expressed features. Among 23 non-differentially expressed spike-in controls, ROTS identified only one false positive at FDR < 0.05, whereas with other methods this number varied from 3 to 18 at the same FDR threshold (Figure 4B). Altogether, the AUC values and false discovery detection rates in spike-in dataset demonstrated the significant performance advantage of ROTS over the tested methods.

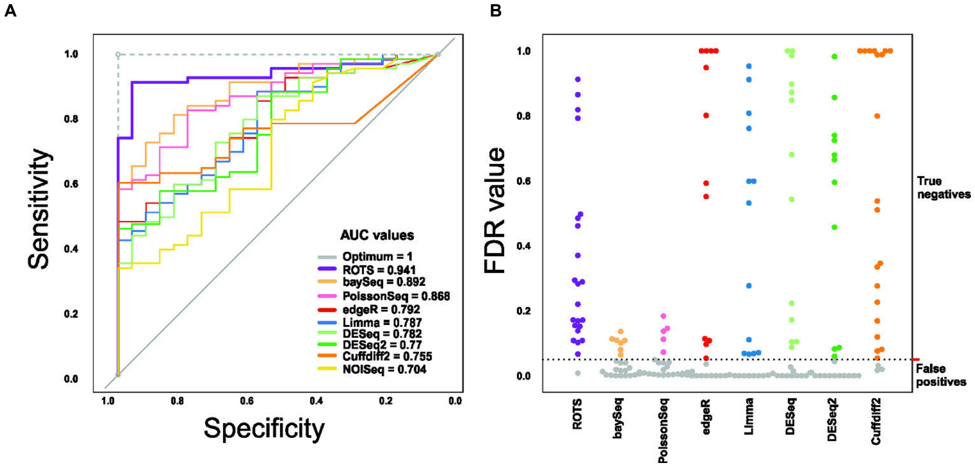

Figure 4. Differential expression analysis using spike-in data. (A) ROC curves and AUC values for detection of DE ERCC spike-in controls. Statistical methods were expected to detect ERCC controls with 0.5, 0.67 or 4 fold changes between the experimental mixtures groups. (B) False discovery rate values of non-differentially expressed ERCC controls. Statistical methods are expected to report FDR values greater than 0.05 for ERCC controls with fold change = 1. The figure is adapted with permission from publication II.

ccRCC risk stratification. Inspired by promising results from spike-in dataset, a complementary study was launched to assess the efficiency of ROTS in real world medical applications. This study aimed to identify ccRCC prognostic biomarkers using gene expression data. Among 442 patients of TCGA data, 40 patients with poor prognosis (survival time < 12 months) and 40 patients with better prognosis (survival time > 60 months) were selected for differential expression analysis. The expression level of 2208 genes differed significantly between the comparative groups at FDR < 0.05 (ROTS reproducibility parameters: R = 0.57, Z = 5.27). To minimize the intrinsic technical noise and improve stringency, additional filtration criteria were taken into account. Hence, significant genes with absolute log2 fold-change below 1.6 (~3-fold change) and average expression level below the lowest 30% of data were filter out remaining 152 genes for downstream analysis.

Unsupervised hierarchical clustering of the differentially expressed genes between pooled TCGA samples identified four ccRCC prognostic clusters (Figure 5A). Although the clustering analysis proved promising results, I was looking for a more stabilized risk stratification approach to ensure the reproducibility of findings. Accordingly, a supervised risk scoring system was developed to assess patient survival outcome. For each patient, prognostic score was defined as the difference between the average log2-transformed expression levels of up and down regulated genes in the prognostic signature of 152 ROTS detections. The risk scores were then clustered into four risk groups, named C1 to C4, using K-means clustering analysis (Figure 5B). Kaplan-Meier (KM) analysis (refer to section 3.3.1 for details) revealed a significant difference in the survival of patients; presenting the best and worst prognosis for C1 and C4 groups respectively while remaining the two other groups in between (Figure 5C; log rank test $P < 10^{-15}$). For low risk patients (C1 group) the five-year survival rate was around 80%, whereas the percentage dropped to below 20% for patients categorized in high risk group (C4 group).

Finally, a completely independent dataset of 100 ccRCC patients was used to evaluate the generalization ability of the proposed model. Figure 5D shows the corresponding scatter plot for signal log ratios of 152 ROTS detections between the best and poorest survived patients from TCGA and validation data. Notably, over 90% of the tested genes showed a significant pattern of regulation association which ensures practicality of validating analysis (Pearson correlation 0.796, $P < 10^{-15}$). Next, the validation data samples were divided into 4 risk groups (high risk, intermediate risk and low risk) and the association between risk groups and survival was estimated using KM analysis. Of note, a similar highly significant association between the classified risk groups and overall survival time (Figure 5E; log rank test $P < 10^{-4}$) was observed as in TCGA data. The spike-in and ccRCC data analyses confirmed the efficiency of ROTS in RNA-seq differential expression analysis. All in all, these analyses confirmed the reliable performance of ROTS on large and heterogenous real world data.

Clinical findings. The differentially expressed genes detected by ROTS were categorized into nine biochemical functioning groups by manual annotation (Figure 6A). The largest category comprised the genes from the cellular transporter and solute carrier (SLC) groups (~26%). Dysregulation of SLC genes and their association with ccRCC prognosis have been reported in several studies [90] [91] [88]. SLC family members regulate nutrient transporter proteins which are essential in both promotion and suppression of tumors growth. Among 26 differentially expressed SLC genes identified in this study, 19 genes were exclusively detected by ROTS. For instance, SLC38A5 is a previously undetected gene which fuels cancer cell by transporting glutamine required to build amino acids and consequently rapid proliferation of ccRCC tumors (Figure 6B). Accordingly, this study supports the theory of using SLC genes as potential therapeutic targets in ccRCC.

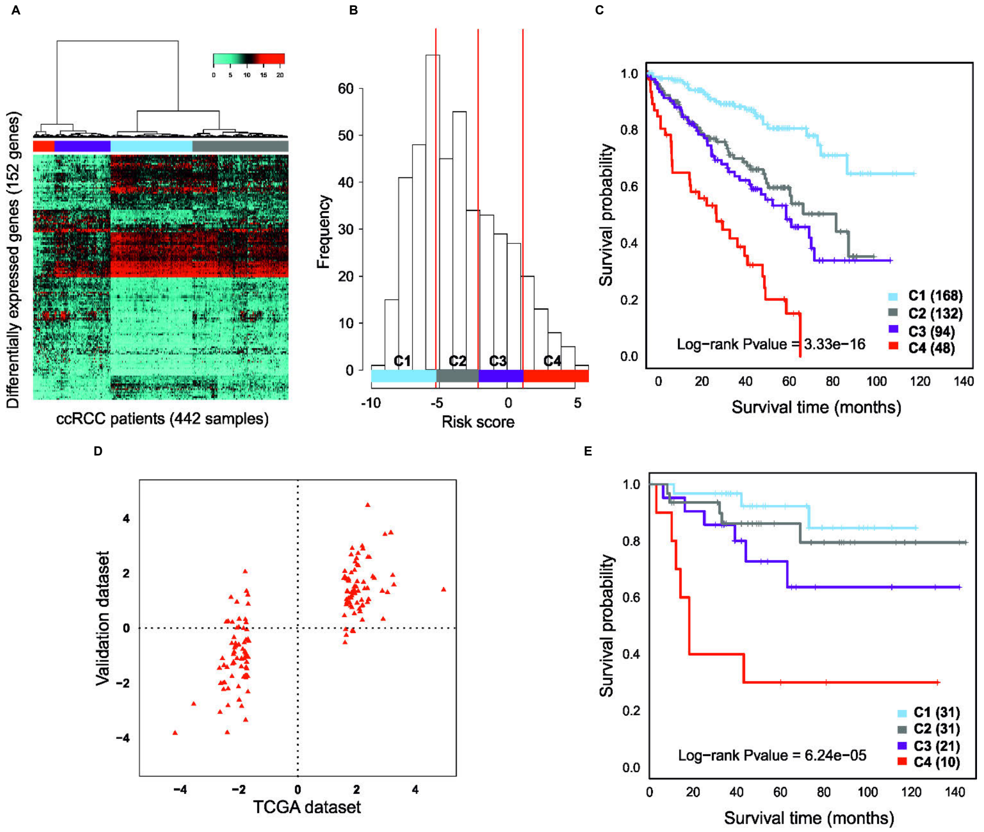

Figure 5. Identification of novel biomarkers to predict survival of patients with ccRCC. (A) Patients were clustered into four risk groups using unsupervised clustering analysis on the basis of differentially expressed genes detected by ROTS. (B) Frequencies of patient-specific risk scores for TCGA data. The red vertical lines show the cutoffs determined by K-means clustering to define ccRCC risk categories. The risk categories significantly overlapped with the similarly colored clusters in panel A (69, 92, 78 and 92 percent of overlap with C1, C2, C3 and C4 survival groups respectively; Fisher's exact test P-value < 0.01). (C) Kaplan-Meier plots of overall survival in TCGA data for different risk groups defined by risk scoring system presented in panel B. The risk groups indicating colors correspond to the colors in panel A and B. (D) Scatterplot of signal log-ratios of differentially expressed genes in TCGA (x axis) and validation (y axis) data for the patients with best (overall survival > 5 years) and poorest (overall survival < 1 year) survival. A highly significant Pearson correlation of 0.796 was observed (p value < 10^{-15}). (E) Kaplan-Meier plots overall survival in validation data. The four illustrated risk groups (C1, C2, C3 and C4) were identified on the basis of prognostic scoring system developed by TCGA data analysis. The figure is adapted with permission from publication II.

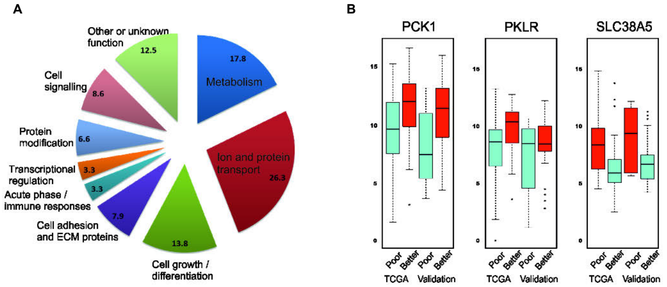

Figure 6. Biological insights from ccRCC prognostic markers. (a) Manually annotated functional groups of differentially expressed genes detected by ROTS. (b) Selected examples from list of 152 differentially expressed genes detected by ROTS. Boxplots show the median and the interquartile range (IQR) of normalized expression levels of selected genes for poor and better prognosis patients in training (TCGA) and validation data. The figure is adapted with permission from publication II.

The next major fraction of detected genes involved in regulating cellular metabolism confirming their well-known role in ccRCC progression [88] [92] [93]. Several previously reported markers such as the key glucose metabolism regulators ALDOB, G6PC and PKLR (Figure 6B) were among the 152 ROTS detections. Additionally, new interesting metabolism regulating markers such as the glucose transporter/sensor SLC2A2, and the central gluconeogenesis regulator PCK1 were identified (Figure 6B). Unlike several other cancer types (e.g. colon, lung and skin), glucose metabolism mediator genes can act as tumor suppressor in kidney cancer and have been shown to be correlated with better prognosis [94] [95]. It was also the case with ROTS detections as up-regulation of six glycolytic genes were observed in better prognosis group (ALDH1L1, ALDOB, G6PC, PCK1, PKLR, SLC2A2). For example, PCK1 contributes in regulation of gluconeogenesis pathway by catalyzing transforming of non-carbohydrate molecules into glucose. In kidney cancers, increased expression of PCK1 mediates death of glucose-starved cells and inhibits cancer proliferation [96]. Finally, the Ingenuity Pathway Analysis (IPA) was performed to find canonical pathways and relating genes interactions. The top affected pathways included molecular transport, small molecule biochemistry, and amino acid and lipid metabolism ($P < 0.05$). In line with earlier studies [88], network analysis revealed several interactions between better prognosis genes linked to cellular metabolism (Supplementary Figure S5, publication II). The poor prognosis genes were enriched in a variety of cell growth signaling molecules (e.g., phosphatases), extracellular matrix and remodeling proteins (collagens, metalloproteins) and acute phase/immune response genes (CRP, SAA family) (Supplementary Figure S4, publication II). For further information regarding the clinical insights the reader is referred to publication II, TCGA [88] and EGA [89] papers.

3 Computational and Machine Learning Algorithms for Predicting Prognosis and Treatment Outcome Using Clinical Data

Conventional clinical markers constitute a main resource in clinical decision making. This class of markers span a wide range of clinical aspects including demographic, pathology, laboratory and diagnosis and treatment history information. In addition to their reliability, clinical factors are mostly cheaper and easier to collect compared to molecular and omics data. Centralized data collection and storage systems provide possibility to distribute and integrate different levels of data for data mining applications. In biomedical sciences, computational techniques are applied to translate raw data into actionable knowledge that can benefit clinical practice. Conventional data analytical methods have proved to be inefficient when tackling with large volume, variation and complexity of new biomedical data. Accordingly, it is essential to develop more sophisticated computational algorithms to overcome data-oriented challenges of the field. This chapter reviews publications III and IV that propose novel computational approaches to model clinical data and predict treatment outcome in advanced metastatic prostate cancer. Furthermore, I evaluated the reliability and generalizability of trial-tailored prognostic models in real-world patients.

Prostate cancer is estimated to be the most common type of cancer in men and among the three top leading causes of cancer death in 2020 [97]. Treatment options vary based on the stage of the disease, risk of recurrence and patient overall characteristics including age and ECOG (Eastern Cooperative Oncology Group) performance status. The treatment strategy may include surgery, radiation therapy, hormonal therapy, chemotherapy and immunotherapy, or most possibly a combination of these options [98]. Surgery or radiation therapy combined with androgen depletion therapy (ADT) is the standard care for locally advanced diagnosed patients. However, majority of patients develop resistance to hormonal therapy and eventually enter into the inevitably fatal metastatic castration-resistant state with estimated median survival of three years [99]. During this state, the prostate-specific antigen (PSA) serum levels continue to increase and new lesion sites especially bone lesions emerge. Currently, the treatment strategy for metastatic castration-resistance prostate cancer (mCRPC) is very complicated with unclear clinical outcome. Cytotoxic medicines, anti-androgens, immunotherapeutic and radiopharmaceutical agents are the available options with proved survival prolongation effect [100]. Among the cytotoxic drugs, docetaxel is one of the first-line approaches in treating mCRPC; but the drug efficacy is still controversial as many patients die while experiencing a lower quality of life due to developing adverse events without any survival benefit. Accordingly, at clinic, oncologists struggle to find target patients who could benefit from docetaxel without putting those who may suffer, at risk of treatment failure. Computational biology can assist in clinical decision making by risk stratifying patients and providing reliable estimates about their clinical outcome following different treatment strategies. This thesis comprises two papers focusing on predicting the clinical outcome of mCRPC patients treated with docetaxel. Publication III presents the prostate cancer DREAM challenge top performing method and the postchallenge ensemble-based model for predicting early docetaxel treatment discontinuation due to adverse events. Publication IV systematically evaluates the reliability of trial-based prognostic models in mCRPC real-world patients. This is a crucial question to answer specially since there is evidence that real-world patients have shorter overall survival with more severe adverse events compared to clinical trial patient cohorts [101]. The ultimate goal in both papers is to provide practical advances in treatment decision making for everyday clinical routine.

3.1 Clinical data sources and types

Clinical data is mainly collected under two different categories: randomized clinical trials (RCTs) and electronic medical records (EMRs). RCTs are research-based studies that recruit a well-defined target group of patients to evaluate new drugs or other newly developed medical interventions. Clinical trials are launched just after the drug efficacy and safety has been proved using in vitro and in vivo models. The ultimate goal of preclinical research is to test the new drug behaviour in a comparable experimental design as in human subjects. Once the preclinical studies suggest promising results in animal models, the proposed drug becomes a potential candidate for clinical trial investigations. According to FDA requirements, a new therapeutic compound should successfully pass phases 0 to 3 of clinical trials to be approved for safety and efficacy. Phase 1 trials usually involve a small number of volunteers (e.g., 20-80 healthy and/or diseased cases) and focus on testing the safety and toxicity of the new drug. The aim is to find the maximum tolerated dose that can be safely prescribed and investigates the possibilities of developing adverse events. Once a new compound has been proved safe in phase 1, it will be subjected for exploratory investigations in phase II trials. Recruiting volunteer patients with the target disease, this phase aims to find out the drug efficacy in fulfilling its therapeutic goals. In addition, phase II trials provide preliminary suggestions on effective dosage and treatment strategy. Finally, a new compound requires to pass phase III trial as final stage to confirm its safety and efficacy in a large patient cohort to ensure adequate statistical power. The final results elaborate the drug efficacy compared to placebo and/or standard available therapies by randomizing patients to different comparative arms. In publication III, clinical data from four phase III prostate cancer RCTs were utilized to develop the proposed model for early prediction of adverse event induced by chemotherapy in mCRPC patients.

The other primary source of clinical data, EMRs, are collected during the routine patients care at healthcare centers or hospitals and significantly improve healthcare quality in comparison with paper-based documents. In particular, the electronically stored data is utilized to assist in real-time treatment decision making and consequently advances patients final outcome. In addition to their positive impact on patients’ care strategies, EMRs provide a valuable source in clinical research. Typically, the collected records are noticeably large, incomplete and heterogeneous and demands heavy preprocessing strategies to make the data suitable for research purposes. Currently, there is no established pipeline to systematically remove undesirable heterogeneity, address the data incompleteness and ultimately cure and summarize the data for further analytical steps. One general preprocessing approach includes versatile standardizing, transforming and imputing techniques to address the data complexities [102] [103].

The reliability of RCTs final outcomes are evaluated using internal and external validity metrics. High internal validity is obtained by appropriate experimental design and ensures that the medical intervention under study is the source of observed differences across comparative arms. While the internal validity is critical for proving the efficacy of a new therapy; however, it is not clear how well the results may be extrapolated to real-world clinical settings. Hence, the metric external validity determines to what extent the RCTs results can be generalized. Cancer clinical trials are the final stage to evaluate the efficacy of new therapies on patient survival and quality of life. However, it is estimated that less than 5% of cancer patients get the chance of participating in a clinical trial and others are either unaware of this opportunity or if they are aware, they fail to be recruited as their overall clinical condition do not meet clinical trials tight eligibility criteria [104]. In general, clinical trials exclude or underrepresent patients with advanced age, poor prognosis or multiple comorbidities. Here, the highly restricted patient recruitment criteria may violate the study external validity and recruited patients are no longer true representors of real-world patients in term of medication benefit and overall survival. This issue has been addressed thoroughly in publication IV.

3.2 Statistical data exploratory methods

Routinely, predictive analyses for medical applications start with a comprehensive data exploration and curation phase using the so called descriptive statistics and statistical learning techniques. The term exploratory data analysis (EDA) was first introduced by Tukey [105] which involved examining the data and producing new insights to it. He compared EDA procedure with detective work where the data analyst has specific question to answer but has to take completely unbiased and data-driven steps to solve the problem. These steps are integrated as a computational toolkit to summarize and visualize data before developing any modeling hypothesis [106]. The EDA toolkit covers wide range of approaches from data quality control to descriptive and analytical statistics. Ultimately, EDA aims to reveal the overall data description as the preparation step for model formation [107].

Descriptive statistics provide extensive data summarization and visualization focusing on data variability and central tendency. In addition, a powerful data screening approach is needed to determine data distribution and quality, data missingness and potential outliers. The data exploration step demands close collaboration with domain experts like biologists or clinicians. Graphical figures and representing tables are utilized to detect potentially primary interesting patterns and variables. In practice, EDA might be the last chance to diagnose data collection errors and assist in data cleaning or data refinement. In biological and medical studies, missing values are almost inevitable and are considered as potential threat to statistical validity and precision. In publication III, variables with missing values in at least 2/3 of the patients or absent in test data were removed from the analysis (refer to Table 1 and Supplementary Table S3 publication III). In publications III and IV missing at random variables were imputed using median values (refer to Supplementary Table S3 publication III and Supplementary Table 1 publication IV).

Identification and correction of potential outliers is another fundamental task during data exploratory phase. With clinical variables, outlier management procedure is unique and context specific. For instance, with laboratory values, the outlier observations may origin from patient lifestyle, medications in use or simply be the result of measuring equipment error. Accordingly, data specific approaches are required to specify the normal range and coping with outlier values while avoiding losing potentially interesting cases. The task of detection and handling of outlier values can be performed using conventional graphical displays (e.g. boxplots or histograms), appropriate statistical tests (e.g. Grubb's test [108] for normally distributed data and chi-square test both implemented in R package "outliers" and

Hample test and Quantile method for both clinical and molecular variables [109]) or multivariate analysis such as PCA (see section 3.2.4 for details). In publication III and IV, graphical displays and PCA method were utilized to detect highly skewed variables. When necessary, outliers were log transformed or discarded to obtain robust modeling results.

3.2.1 Clustering analysis

Cluster analysis is a data exploratory technique that aims to classify data into homogeneous subsets (clusters) in such a way that subjects within the same cluster have maximum similarity with each other and minimum similarity with other clusters. Clustering methods assess each data point based on its measurements and relation to the other data points and eventually discover hidden patterns between subjects without any predefined rules or information. In most cases, detecting the relation between the discovered clusters also provides valuable insight to the question of interest. This relation can be shown by grouping more similar clusters into the same level of hierarchy such that the distance between each hierarchy class resemble the difference between them. The primary goal of all clustering techniques is to calculate a similarity metric between individual subjects to provide a reduced notation of the data. There are extensive range of clustering algorithms which are very close in theory but may vary in final grouping results. In the context of computational biology, hierarchical clustering algorithms and partitioning clustering algorithms are the most popular approaches in use.

3.2.2 Hierarchical clustering

This family of algorithms output a hierarchical cluster tree such that each level of hierarchy is formed by combining its lower level clusters. Agglomerative algorithms are the most popular hierarchical clustering technique in bioinformatics. They follow a bottom-up approach to develop the data hierarchy. The algorithm starts with assigning individual data points as a singleton cluster at the lowest level. Next, at each step successively the most similar clusters are merged (agglomerated) into a new cluster and build the next upper hierarchy level. The merging procedure continues until all the clusters are combined into one cluster which comprises the whole dataset at the highest level of hierarchy. In order to find and fuse the closets (least dissimilar) clusters at each step, a dissimilarity metric should be predefined [110]. There are versatile approaches to calculate the clusters similarity. For instance, R language supports the following schemes in measuring the distance between the clusters: ward, single [111], complete, average (UPGMA), mcquitty (WPGMA), median (WPGMC) and centroid (UPGMC).

In the single-link method (also called nearest neighbor method), the similarity between two clusters is defined as the similarity between their closest observation pairs and clusters with minimum similarity measure are combined together. Complete-linkage method estimates the clusters distance by calculating similarity between the most dissimilar observation pairs and clusters with distance are merged together. As the name suggests, the average-linkage uses the average distance between all members of the two clusters and combines clusters with smallest average distance. For Ward's method, the distance between two clusters are defined as cost of losing information when combining them. The clusters with smallest cost will merge together. The merging cost is calculated as sum of squared errors.

In contrast with different linkage methods that focus on similarity/dissimilarity measures, the Ward algorithm aims to minimize the loss of information when merging two clusters. To do so, this algorithm computes the sum of squared errors (SSE) between the elements of two clusters and merges the clusters with minimum increase in their SSE value. This way the algorithm is able to minimize heterogeneity or variation within the merged clusters and preserves maximum unity among the combined observations. In publication III, hierarchical clustering was performed using Ward's method and Manhattan distance for risk stratification of mCRPC patients suffering from docetaxel adverse events.

3.2.3 Partitioning clustering

Partitioning clustering algorithms divide the data into a flat set of distinct clusters. The algorithm initially decomposes elements of data into predefined number of clusters and then iteratively moves the data points between the clusters until the algorithm criterion function is optimized. K-means algorithm is the most common and simplest example of partitional clustering techniques. The method starts with defining one centroid for each cluster and then assign all other data elements to their closest centroid (cluster).

3.2.4 Data dimensionality reduction

Current biological and clinical datasets are commonly high-dimensional with significant number of redundant variables or irrelevant to the study purpose. Dimensionality reduction is a vital step to ensure the feasibility of computational analysis while improving the accuracy and power of the developed models.

Feature selection and feature extraction are the most common techniques to reduce data dimension. Feature selection techniques aim to detect a subset of the most relevant features with potential high power in predicting the outcome variable. Depending on data type and model development strategy, feature selection techniques are divided into three main categories: filter based, wrapper based and embedded feature selection methods. Filter based methods are independent from model development and are usually performed during the preprocessing phase. These methods utilize statistical tests or metrics to assess the association between input variables and model outcome and filter features with low-ranked association. Wrapper based methods follow an iterative searching approach which perform feature selection and model learning simultaneously. In each iteration step, a model is trained using a specific subset of features and the model performance determines the relevance of selected features to the outcome variable. When all the designed iterations with intended features subsets are performed, the most accurate model with its associated features are selected. Embedded feature selection methods are integrated within the model development procedure and the candidate features are selected during the model training phase. Decision trees employed in boosting algorithms Lasso method are the widely used learning algorithms that perform feature selection as embedded task.

Feature extraction is a data transformation technique which aims to reduce the data dimension by aggregating the potentially informative variables into a new reduced format. Here, the main focus is to decrease the features space while preserving the meaningful data variation and maximizing the accuracy of the learning algorithm [112]. In publication III, we extracted some additional features from the core table (refer to 3.6.1 for detailed information about core table) to predict early discontinuation of chemotherapy in mCRPC patients. These included LESIONS, DRUGS, DISEASES and PROCEDURES, which were defined as arithmetic sums of the numbers of lesions, medications, diseases and medical interventions respectively. These newly proposed features were representors of 54 separate variables which their solo usage could have smaller effect on our proposed learning algorithm.

Principal component analysis (PCA) is a non-parametric method widely used for dimensionality reduction with molecular and clinical data. PCA performs the dimensionality reduction task by transforming a class of correlated variables into a smaller set of uncorrelated components (also called principal components) representing the linear combination of the original variables. The computed principal components are ordered on the basis of amount of variance they explained about the data. For instance, the first principal component (PC1) accounts for the largest possible proportion of variance in the input data. Additionally, PCA is a powerful technique for exploring data distribution and detecting outlier units or observations [113]. In publications III and IV, PCA was widely used for both feature extraction and data distribution exploration purposes.

3.3 Survival analysis

In Survival analysis, the focus is to assess the time duration until an event of interest occurs. For medical studies, the event of interest is mostly time to death or time to development or progression of a disease. One practical example are the clinical trials that aim to benchmark the efficacy of a new medication on patients' survival prolongation. Since the data is mostly collected prospectively, it is important to have consistent starting and ending points for all the data subjects to have a robust experimental design. In survival analysis, if for an observation the information regarding the time to event is not available, the situation is called censoring and usually specific tools and graphical methods are demanded to deal with the situation. The more common censoring type in medical studies is right censoring which includes observations that do not experience the study outcome or experience it after the study ends. For instance, if the project aims to assess the cancer patients' survival in five years, the patients who abandoned the study before its completion or died after the study period; are counted as the right censored observations. Censored data is one type of data incompleteness which is a barrier to estimate data distribution parameters and so typical regression algorithms are not applicable in this context. One common approach to deal with the situation is to utilize semi-parametric and non-parametric algorithms.

3.3.1 Non-parametric survival methods

Non-parametric survival methods maximize the utilization of dataset information by using censored observations till their censoring time point. Here, the survival probability is calculated as a time dependent function. Let's assume that $T \geq 0$ is a random variable denoting the survival time, then $S(t)$ represents the probability that an observation survives beyond time t

$$S(t) = P(T > t) = \frac{No\ of\ alived\ observations_{at\ (t=\ 0)} - No\ of\ dead\ observations_{at\ (t=t)}}{No\ of\ alived\ observations_{at\ (t=\ 0)}}. \qquad (\ 15)$$

If the aim of analysis is to calculate the survival probability during a certain follow-up period rather than an individual time point, then $S(t)$ can be calculated using Kaplan-Meier (KM) method:

$$S(t_i) = S(t_{i-1})\left(1 - \frac{d_i}{n_i}\right)\ when\ t_0 = 0, S(0) = 1, \qquad (16)$$

where $S(t_{i-1})$ is the probability of surviving at t_{i-1} time interval, d_i denotes number of dead observations and n_i corresponds to number of alive observations, both at

t_{i-1}. By plotting the survival probabilities against follow-up time, a step-like KM curve will be obtained. In biomedical applications, KM curves are widely used to illustrate censored data, especially when comparing survival probability of two or more experimental conditions [114] [115]. Log-rank test is used to test whether the differences between certain survival distributions are statistically significant or not. In this thesis, I have frequently used KM analysis to evaluate the association between risk categories and patients' prognosis.

3.3.2 Semiparametric proportional hazard models

Precise assessment of patient prognosis is a key factor in oncological decision making. Cox proportional hazard model is the most widely used survival analysis method to predict cancer prognosis and recurrence. The model is used to examine the potential relationship between the study covariates (e.g. clinical features and tumor characteristics) and time-dependent clinical outcome of interest. For example, it evaluates whether a new medication is associated with survival prolongation. Cox proportional hazard models are classified in semiparametric proportional hazard models as the method does not make any assumption about the distribution of the baseline hazard function. Suppose we are interested in studying time to outcome of interest in a cancer patient cohort with sample size = n and $x = \left(x_1, x_2, \ldots, x_p\right)$ as vector of potential covariates. One can denote each patient's information as a vector (t, δ, x) where the study outcome at time t is complete if $\delta = 1$ or is right censored if $\delta = 0$. Following this setting the hazard function for Cox proportional hazards function at time t is defined as:

$$h(t; x) = h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i x_i\right), \tag{17}$$

which can be simplified as

$$h(t; x) = h_0(t) \exp(x^T \beta), \tag{18}$$

where $\beta = \left(\beta_1, \beta_2, \ldots, \beta_p,\right)$ is the vector of regression coefficient for p covariates and $h_0(t)$ is the baseline hazard function of the underlying survival distribution when all the covariates are assumed to be zero. Since $h_0(t)$ is unspecified, the regression coefficient β can be estimated using partial likelihood approach rather than other likelihood methods. Accordingly, β is estimated by maximizing $l(\beta)$ as:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ x_i^T \beta - \log\left(\sum_{k \in R(t_i)} exp\left(x_k^T \beta\right) \right) \right\}, \tag{19}$$

where $R(t_i) = \{i : t_i \geq t\}$ denotes a risk set at time t_i [116].

Traditional maximum likelihood estimators aim to determine the vector of coefficients such that they maximize the probability of the observed data. With high dimensional data, developing a model with highest likelihood may require all covariates incorporation leading to a low-level generalization ability. Accordingly, methods that produce non-zero estimation for all covariates are inapplicable as the final model is difficult to interpret and suffers from overfiring. Here, the main challenge is to detect an efficient subset of covariates with substantial contribution on hazard function. Penalized likelihood estimation is a solution to reduce model complexity by shrinking some covariates to zero. A family of penalizing methods are proposed by Fan and Li [117] that performs variable selection and coefficient estimation simultaneously. Although the proposed method was designed for parametric models, it is applicable to semi-parametric Cox model using appropriate transformations.

Suppose $X = (x_1, x_2, \dots, x_p)$ denotes a vector of covariates, $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ represents the vector of corresponding regression coefficient and let y denotes the study response. The penalized regression problems are commonly formulated as

$$\hat{\beta} = arg \min_{\beta} \left(\|y - X^T \beta\|^2 + \lambda \sum_{j=1}^{p} P(\beta_j) \right), \tag{20}$$

where λ is a tuning parameter and $P(\beta_j)$ is the penalty function. The choice of the penalty function is the underlying difference among regularization methods. The Lasso (Least absolute shrinkage and selection operator) method [118] [119] and Elastic net [120] are the well-known members of this family penalizing approach. The Lasso utilizes the l_1 penalty to shrink the negligible coefficients to zero and produce a sparse solution for high dimensional data. The simplified sparse models are easier to interpret with reduced risk of overfitting. The Elastic net method combines l_1 penalty and l_2 penalty to solve the model regularization problem. Both Lasso and Elastic net and their variations (e.g., adaptive Lasso) are widely used in predicting cancer prognosis. In publication IV, all the benchmark models utilize regularization approaches including Lasso, adaptive Lasso and Elastic net to predict overall survival in patients with advanced prostate cancer.

3.4 Boosting algorithms

Boosting is one of the most powerful machine learning techniques which is extensively used in medical applications. The boosting original idea was to combine many weak learners to produce a powerful model with optimally enhanced performance. Besides enabling flexible non-linear solutions, the benefits of the boosting strategy include its capability of feature selection, capacity of handling missing values, as well as relatively low run times. The first practically successful algorithm was Adaboost (Adaptive boosting) proposed by Freund and Schapire [121] for binary classification problems with the possibility to expand into multiclass problems. Adaboost is an iterative algorithm which predicts a sample class using the weighted combination of weak learners produced at each iteration step. Here, a weak learner is expected to achieve accuracy slightly better than random chance (50%). Suppose a training set of size N as $(x_1, y_1), \dots, (x_N, y_N)$ where x_i represents the vector of predictor variables and $y_i \in \{-1, 1\}$ represents the binary outcome, the Adaboost algorithm is described as follows:

- Initialize the weight for each observation as:

$$w_i = \frac{1}{N}, i = 1, 2, \dots, N$$

- For $m = 1\ to\ M$ do:

  a. Fit weak learners to the training data and select the learner $G_m(x)$ with lowest error rate computed as:

$$err_m = \frac{\sum_{i=1}^{N} w_i\, I(y_i\ \neq G(x))}{\sum_{i=1}^{N} w_i}$$

  b. Calculate the weight for the selected weak learner:

$$\alpha_m = \frac{1}{2} \ln\left(\frac{1 - err_m}{err_m}\right)$$

  c. Update the weights for each member of training set for the next iteration:

$$w_{m+1} = w_m \,.\exp\left[\alpha_m \,.I\big(y_i\ \neq\ G_m(x_i)\big)\right], i = 1, 2, \dots, N$$

- Output the final prediction function:

$$G(x) = sign\left[\sum\nolimits_{m=1}^{M} \alpha_m G_m(x)\right]$$

Here, $G(x)$, the final learning function combines the weak learners ($G_m(x)$) while incorporating their respective weights (α_m) to make decision about the class label of the test data. Initially, the weights of all data points are set to $w_i = \frac{1}{N}$ but the weights get updated on the basis of previous iteration learner accuracy. More specifically, at each iteration m, a weak learner is trained using the samples receiving their weights from the iteration $m - 1$. This trained weak learner $G_m(x_i)$ should have the lowest error rate as computed (line a. Adaboost algorithm). Next, the weak learner weight

α_m is calculated in such a way that a learner with lower error rate receives higher weight and consequently has greater impact on the final prediction function (line b. Adaboost algorithm). Finally, the observations weights are updated as a result of applying the developed weak learner to all data points (line c. Adaboost algorithm). At this step, the aim is to increase the weights of misclassified observations while decreasing the weights of correctly classified samples. Accordingly, the misclassified samples from iteration m receive more exploration considerations when training the G_{m+1}. After M iteration, the prediction of an unseen sample is obtained by summing up the weighted votes of all weak learners. The original Adaboost algorithm returns a discrete value as the predicted class label. Modifications to algorithm return value yields to generalization of the method to wider range of problems including regression problems [110].

3.4.1 Forward stagewise optimization

For both classification and regression problems, boosting algorithms fit an additive model with linear combination of weak learners as

$$f(x) = \sum_{m=1}^{M} \beta_m b_m\,(x; \gamma_m), \tag{21}$$

where x denotes the input vector, β_m and γ_m are the model parameters to be optimized and b_m are the arbitrary modeling functions of input x. Here, the task is to estimate the optimized values of β_m and γ_m by minimizing a general loss function L averaged over the input data x for individuals $i = 1, 2, \dots, N$ as follow

$$\langle \beta_m, \gamma_m \rangle_1^M \leftarrow \arg\min \sum_{i=1}^{N} L\left(y_i, \sum_{m=1}^{M} \beta_m b_m(x; \gamma_m) \right). \tag{22}$$

With most common loss functions, direct optimization of above phrase could be a complicated task. One alternative solution to minimize the loss function in () is using an iterative approach called forward stagewise additive modeling (FSAM) [122]. Following this approach, it is possible to approximately solve equation 23 starting by single base function

$$min_{\beta,\gamma} \sum_{i=1}^{N} L\left(y_i, \beta b_m(x; \gamma)\right) \tag{23}$$

and then sequentially adding new basic functions ($m = 1\ to\ M$) and their corresponding coefficients without updating parameters of previous models. More specifically, the FSAM algorithm finds the optimal loss function via solving a series of subproblems in a greedy approach while preserving the parameters of previously added sub-functions. In practice, many of boosting methods with adjusted Adaboost algorithm can be simply fit into a forward stagewise approach using an exponential loss function. The reader is referred to [110] for more details on the topic.

3.4.2 Gradient boosting machines

The boosting algorithms building on Adaboost can be extremely powerful and highly robust especially when decision trees serve as base learner [123]. As mentioned earlier, efficient estimation of loss function is the main difficulty for this class of learners. In 2001, Friedman introduced a gradient decent based function estimation method, called *gradient boosting machine (gbm)*, which utilizes the stagewise additive manner similar as in FSAM algorithm [124]. The proposed GBM algorithm, is a greedy function estimation approach which is capable of minimizing a rich set of arbitrary differentiable loss functions. GBM performs the learning task by consecutively adding a new *step* or *boost* to the sequence of successive expansions while minimizing the prediction error rate of the previous steps. At each iteration, the new boost or base learner is developed so that it is fitted to the gradient of residuals calculated on the basis of previous predictions. This way, gradient boosting algorithms convert a model fitting problem into a parameter optimization task. Hastie et al. [110] modified the gradient boosting algorithm proposed by Friedman for regression problems with decision trees as choice of base learner as follows:

1. Initialize $f_0(x) = \arg min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$ with a constant.
2. for $m = 1$ to M do:
3. Compute the negative gradient for all the N training samples
$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \qquad \forall i \in \{1, \dots, N\}.$$
4. Fit a regression tree to the targets r_{im} giving terminal regions
$$R_{jm}, j = 1,2, \dots, J_m.$$
5. Find the optimal step length gradient decent:
$$\gamma_{jm} = \arg\min \textstyle\sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \qquad \forall j \in \{1, \dots, J_m\}.$$
6. Update the function estimate:
$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}).$$
7. end for.
8. return $f_M(x)$.

Here M is the number of iterations and $L(y, f)$ is the selected loss function. The optimal function is initialized with a single node tree $f_0(x)$ as a constant starting guess (line 1). Next, the algorithm estimates the model parameters following an additive approach to obtain the final function $f_M(x)$ implemented in lines 2-7. Line 3 calculates (negative) gradients r_{im} for every sample in the training set using their fitted values from the successive predictions. The r values, also known as pseudo residuals, are then used to determine the direction and corresponding parameters of the next base learner (line 4). At this step, the steepest decent which gives least error will be chosen to compute the new base learner parameter (line 5) and this new estimate will be added to the ensemble solution (line 6). Finally, the optimized function is produced using the sum of all the estimated base learners from the M iteration step (line 8). Following this approach, the gradient boosting algorithm can be compatible with considerable selection of both loss functions and base learner models required for regression and classification problems. This thesis utilizes adapted version of gradient boosting algorithms for survival analysis. In this case, the negative gradients of partial likelihood for Cox model are used to fit the base learners. The algorithm explained in this section is implemented in R package *gbm* which is used for analysis in publications III and V.

3.4.3 Gradient boosting regularization

For all model development procedures, regularization is an essential part ensuring that the model is not overfitted to the training data. Gradient boosting is highly prone to overfitting due to its greedy manner while expanding the weak base learners. The number of boosting iterations (M) is a primary parameter that can affect generalization properties of a GBM-based predictive model. Controlling or early stopping of the M value plays as a tradeoff between minimizing the error rate and model regularization ability. Friedman suggested to estimate the optimal M value by monitoring the prediction error curve versus a wide range of M values using an independent validation data or cross-validation techniques.

Shrinkage is the next widely considered technique for controlling the model complexity. It aims to scale the impact of additive expansions by penalizing their weights using an additional *learning rate* $0 < v \leq 1$ parameter. This can be simply implemented in the gradient boosting algorithm by modifying line 6 as follow

$$f_m(x) = f_{m-1}(x) + v \,. \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}). \tag{24}$$

This way, the algorithm mitigates the negative effect of error rate corresponding to each base learner at their subsequent iteration step. It should be noted that, M and v are correlated parameters and should be optimized with respect to each other. In practice, smaller values of v parameter may give smaller error rate, but this would simultaneously increase the value of M and computation costs. For an optimal composition, Friedman suggested a $v - M$ tradeoff strategy which aims to define small value for v while optimizing the M value early stopping approach. With real-world data, it is common to have very small learning rate, often in the range between 0.05 and 0.1.

Additional procedure with direct impact on model regularization and computational efficiency is subsampling. Applying this method, a random subset of training data rather than complete dataset is utilized for development of the base learners. Here, the main goal is to prevent base learners fitting to data noise and consequently improving the accuracy by introducing randomness via data sampling. However, it is important to establish a balance when setting the sampling ratios as the insufficient sample size may reduce the power of analysis. In practical experiments, subsampling coupled with appropriate shrinkage approach may lead to better performing models [5]. The ratio of selected subsets is specified by an extra parameter called bag fraction which is dependent to the problem sample size. For example, Hastie et al. [110] suggested the bag fraction equal to $\frac{1}{2}$ can be a reasonable ratio for typical datasets with moderate sample size. Of note, subsampling can significantly reduce the burden of computational efforts by decreasing the sample size into $\frac{1}{2}$ or even $\frac{1}{3}$ for very large datasets.

3.4.4 Relative importance of selected variables

With tree ensemble models, interpretation of the fitted model in term of specifying the selected predictive variables and their influence is different from typical regression analysis. More specifically, this class of models describe the relation between the selected variables and the response through relative importance of selected variables rather than reporting variable coefficients [125] [124]. For selected variables, the relative importance measures the average influence of that variable in decreasing the prediction error rate over all the decision trees. More specifically, this measure of influence is proportional to the number of times a variable has been chosen to form a split; and also, to the average impact this split has had in minimizing the loss function. Breiman et al. [125] proposed a heuristic approach to calculate the relative influence of a variable l for a single decision tree T with J splits as follow:

$$Influence_l\,(T) = \sum_{t=1}^{J-1} i_t^2\, I(v_t = l), \tag{25}$$

where all the $J - 1$ non-terminal nodes of the tree are evaluated to calculate the variable effect. v_t represents the current splitting variable under evaluation and so expression $I(v_t = l)$ will determine the number of times the selected variable l is utilized as splitting point; and, i_t^2 is an empirical coefficient representing the squared prediction improvement corresponding to the examined split. This approach can be simply expanded to calculate the relative influence of a variable over all the trees by averaging its influence on single trees

$$Influence_l = \frac{1}{M}\sum_{m=1}^{M} Influence_l\ (T_m). \tag{26}$$

When the relative influences of all variables are obtained, the variable with largest influence serves as reference value and the influence of the remaining variables are represented relative to the reference variable.

3.5 Power analysis

The reliability of statistical hypothesis testing is strongly dependent to controlling type I and type II errors. Typically, type I error can be avoided by fixing the significance level during the experimental design (e.g., P-value = 0.05). Unlike type I error, controlling type II error is not straight forward in biomedical applications and requires more considerations. Sufficient sample size is a key factor in controlling type II error without threatening type I error justification criteria. Power analysis are performed to obtain the optimal sample size and consequently avoid false negative results. The statistical power is defined as the probability of detecting a true difference between comparative groups using an appropriate statistical test. Similar as in justifying the significance level, power is an arbitrary value which is most often set to 80% to 90%.

The secondary objective of publication III was to establish a quantitative benchmark to plan more efficient clinical trial design by optimizing patient selection and recruiting criteria. For clinical trials used in this study, high number of patients discontinued docetaxel treatment due to adverse events could lead to clinical trial failure, while detection of at-risk patients during trial design could enhance success rate and cost effectiveness. Accordingly, we conducted a simulation study aimed to assess justified number of patients which is adequate to obtain statistically significant results while possessing the ability to refuse enrolment of patients with high risk of treatment discontinuation. More specifically, we aimed to demonstrate the determinant role of prespecification of statistical power and sample size in final clinical trial results. In this simulation analysis, we assumed a balanced two-arm trial with patients randomly assigned to each arm (e.g. treatment and control groups) and the primary end point was overall survival. In order to establish a realistic setting, the ENTHUSE33 trial (the challenge validation data), was used to inform simulation parameters. The survival times were generated from an exponential distribution and the comparator arm had consistent hazard ratio between patients that did and did not discontinue docetaxel as in ENTHUSE 33 trial. For the control arm we assumed 0% of patients discontinued treatment early while this rate was set to 10.4% for the treatment group as reported in ENTHUSE 33 trial. Using the estimated parameters, 100 randomized trials were simulated, each with 10,000 total patients. Within each of 100 simulated datasets, randomly sampled subsets (with replacement) of patients at a ratio of 1:1 from control and treatment groups were selected to estimate the required sample size to detect a survival difference between the comparator arms at 80% statistical power and a false positive rate of 0.05. The survival difference was examined on the basis of hazard ratios (Hazard ratio = {1.30, 1.40, 1.50, 1.60, 1.70, 1.80, 1.90, 2.00}) representing a significant decrease in the risk of death in the treated group. The baseline prediction models were defined as the models with accuracy of 0%, 25%, 50%, 75% and 100% in detecting patients at risk of early discontinuation. For each baseline model, the at-risk patients were excluded from the randomization and the model with 0% accuracy in identifying true discontinuation cases was used as a reference benchmark for comparison with baseline prediction models with 25%, 50%, 75% and 100% accuracy at detecting true at-risk patients.

3.6 An ensemble-based prediction model for short-term discontinuation of docetaxel in mCRPC patients: Publication III

Publication III represents a collaborative study between 34 international teams to build a prediction model for early treatment discontinuation in patients with advanced metastatic prostate cancer. Docetaxel is a chemotherapy agent which has been proved to have survival benefit for mCRPC patients. However, about 20% of patients would eventually discontinue docetaxel because of sever adverse events before completion of the treatment plan. In this paper, an ensemble-based model is proposed for prediction of at-risk patients using routinely collected clinical features.

3.6.1 Materials and Methods

mCRPC data. The aim of this study was to identify risk factors associated with development of adverse event in mCRPC patients treated with docetaxel. The study was designed as an international crowdsourced DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge and my developed model was selected as top performer among 34 international teams. In total, the clinical record of 2070 patients were selected from four phase III prostate cancer randomized clinical trials with identical experimental design (ASCENT2: n = 476; VENICE: n = 598; MAILSAIL: n = 526; and ENTHUSE 33: n = 470). Following the predefined study design ruled by the challenge organizing team, three of the trials (ASCENT2, MAINSAIL and VENICE) were utilized as training dataset for model development and the remaining ENTHUSE 33 trial was used as test dataset for independent model validation. All the patients were chemotherapy-naïve with progressive disease and treated with docetaxel as their first-line chemotherapy treatment. The trials' records were carefully curated and transformed into a homogenous standardized format for further analysis by the DREAM challenge organizers. More specifically, the standardized data were divided into five tables of laboratory values, lesion measurements, prior medications, medical history, and vital signs with trials follow-up details. In addition, a so-called core table from integration of the above tables with baseline values (day 0) were generated to be utilized as the analysis main reference. In total the core table included 129 baseline variables plus study outcomes and patients' follow-up time or outcome incidence. Figure 7 illustrates the detailed study design. The primary steps of data collection and curation strategy and process was designed and performed by the challenge organizers as an independent work which is beyond the scope of this thesis. The data and detailed information about collection and curation is provided at data supplement of publication III.

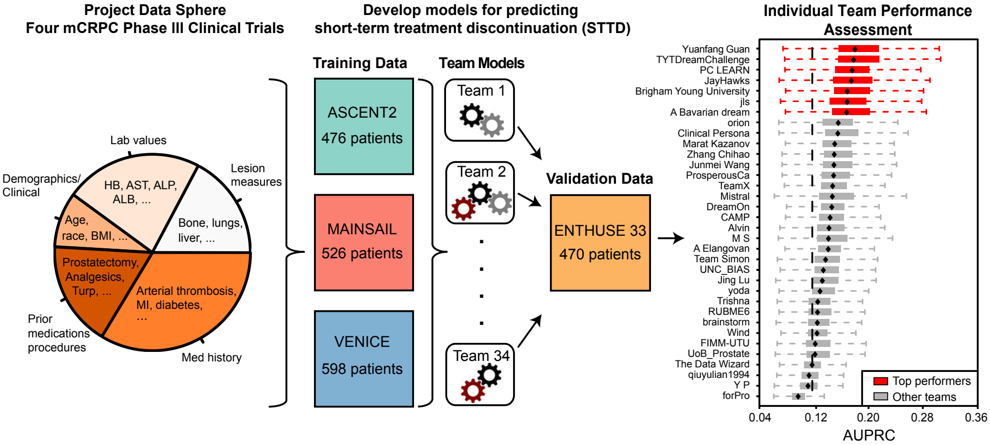

Figure 7. Prostate Cancer DREAM Challenge overview. Three of the trials including ASCENT2, MAINSAIL and VENICE served as training data, while ENTHUSE 33 trial was used for model validation. The figure is adapted with permission from publication III.

3.6.2 TYTDream Challenge model

The model proposed by my team, the TYTDreamChallenge, was selected as the top-performing model to predict early chemotherapy treatment discontinuation due to adverse event in mCRPC patients. The procedure of model development was implemented in three main steps: data preprocessing, model training and model evaluation. Initially, the data were preprocessed to filter out variables with 1) had missing values in at least 2/3 of the patients (including testosterone), 2) were absent in the test dataset (including blood urea nitrogen, glucose) or 3) were highly collinear (including alanine transferase). The remaining laboratory variables (except PSA values that were only log_2 transformed) were scaled to their normal range references as

$$x_s = 2 \times \frac{x - \alpha}{\beta - \alpha} - 1 \tag{27}$$

where x is the observed value of the laboratory test and α and β are the corresponding lower limit and upper limit of the reference range. Next, the remaining highly skewed variables were truncated and/or logarithm transformed to mitigate the data providers batch effects (including neutrophils and alkaline phosphatase). Furthermore, we considered extracting additional variables from binary variables including lesions, prior medications, prior diseases and prior surgical procedures. The extracted variables are called LESIONS, DRUGS, DISEASES and PROCEDURES risk scores and represent the arithmetic sum of the presence of the individual variable in their corresponding class. In addition to reduce the data dimensionality, the proposed risk scores were proposed to minimize the effect of study dependent missingness that was observed in the four selected binary variable classes.

For model training, the gradient boosting algorithm implemented in R package *gbm* was utilized. The study outcome consists of two variables: 1) a binary variable indicating the status of treatment discontinuation which was equal to “1” if the treatment was discontinued within three months due to development of adverse event and “0” if the treatment was tolerated for more than three months, and 2) a discrete variable that measures treatment discontinuation period in days. Accordingly, the outcome was modelled as a survival problem, where time (in days) to treatment discontinuation is employed for right censoring. The model performance and its regularization ability were evaluated and improved following a two-step strategy. First, a preliminary set of models was built using ten-fold cross-validation over the MAILSAIL study and the models were validated on the VENICE study. Variables with higher relative influence among all the preliminary models were selected for final modeling. Second, the final predictive model was built using all of the three input datasets (ASCENT2, MAILSAIL, VENICE) and selected variables.

3.6.3 Results

TYT markers in order of relative influence. The initial aim of the prostate cancer DREAM challenge was to identify patients who are expected to develop adverse events and discontinue docetaxel regimen therapy on the basis of baseline clinical variables. The novel model developed by my team (TYTDream Challenge) significantly outperformed other DREAM challenge proposed models (BF > 3). The final trained model had a slow learning rate of 0.07 and the optimized maximum depth and number of trees were 3 and 232 respectively. Figure 8 represents the relative importance of the identified predictive factors associated with development of adverse event in mCRPC patients. More specifically, my top performing model identified five laboratory values including PSA (prostate-specific antigen), neutrophils (NEU), hemoglobin (HB), alkaline phosphatase (ALP) and aspartate aminotransferase (AST) as critical factors in predicting docetaxel adverse event. Furthermore, the ECOG performance status, analgesic use and risk scores derived from prior medications and therapeutic procedures, medical histories and lesion sites proved to have deterministic role in pre-identification of at-risk patients.

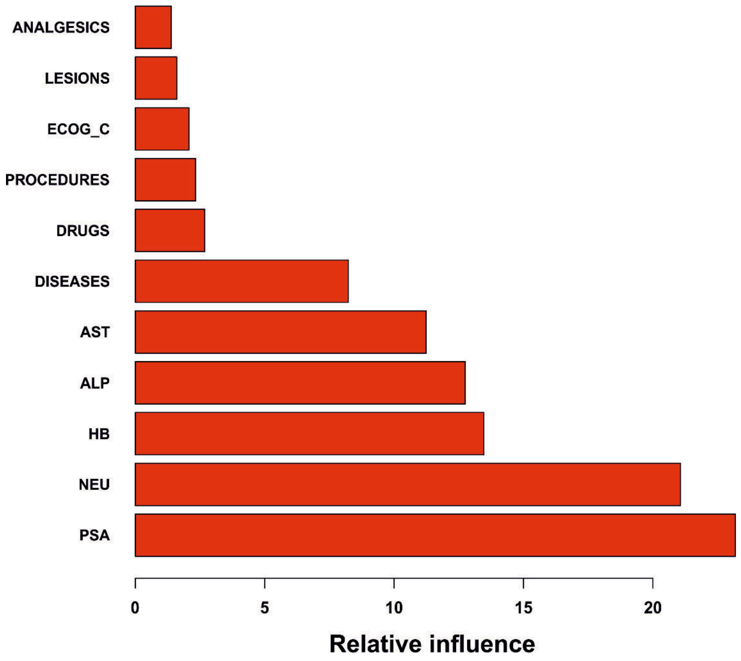

Figure 8. Bar graph illustrating the important clinical features ranking on the basis of their influence to predict early treatment discontinuation.

Post challenge Analysis results. After the completion of the challenge, postchallenge collaboration between top performing teams and challenge organizers was formed and novel results were obtained from the meta-analysis. In order to explore correlations between top performing models' predictions, hierarchical clustering analysis was performed on the rank-normalized risk scores using Ward's method and Manhattan distance. Three subgroups emerged from the cluster analysis (Figure 9A): patients consistently at high risk of early discontinuation (concordant high risk; n= 50), patients consistently at low risk of early discontinuation (concordant low risk; n = 170) and patients with discordant risk scores (discordant risk; n = 234 patients). There was approximately two-fold increase in risk of developing adverse events for concordant high-risk group compared to concordant low risk and discordant risk groups. As expected, a similar trend was observed when evaluating death as competing risk event (Figure 9B). Further analysis investigating association of baseline characteristics and identified risk groups showed a significant difference in the distribution of several key variables (adjusted p-value <0.05). In case of laboratory values, albumin (ALB), HB, lactate dehydrogenase, PSA, sodium, red blood cells (RBC), ALP, calcium, AST, creatinine clearance, and total protein had significantly different distribution across risk groups (Figure 9C). Compared to concordant low risk and discordant groups, patients clustered in concordant high-risk group had significantly poorer ECOG performance status and most often required opioid analgesics and ace inhibitors. Additionally, the incidence of liver metastases and medical history of genetic disorders were evident approximately four times higher in the high-risk group (Figure 9D).

Furthermore, the postchallenge collaborative analysis contributed to development of an ensemble-based model with improved performance in comparison with individual top performing teams. The ensemble-based model was generated as the weighted sum of the individually predicted risk scores from the seven top performing models (Figure 10A). To calculate teams' weight, the training data, D, was split into two random subsets of D^{70} (contained 70% of patients from training data; n= 1120) and D^{30} (contained 30% of patients from training data; n = 480). Next, the top performing teams were asked to train their learning algorithms $L_i(.)$ on D^{70} and report their new predictors ($C_i^{70}(.), i = 1, \dots, 7$). Finally, the newly produced models C_i^{70} , were used to estimate the early docetaxel discontinuation risk for patients in D^{30}. The accuracy of C_i^{70} models in estimating discontinuation risk of D^{30} patients was set as the teams' weights (w_i^{70}) in developing the ensemble-based model. In particular, the ensemble-based learner was proposed as the linear combination of weighted classifiers trained on D data as follow:

$$C_w(r) = \sum\nolimits_i^7 w_i^{70} \, C_i(r)$$

Finally, the ensemble-based classifier $C_w(r)$ was applied to the validation trial (ENTHUSE 33) and resulted in an AUPRC of 0.23 (Figure 10B) which significantly outperformed the top individual classifiers (BF > 2.75).

Although this study primarily aimed to develop a risk prediction model for identifying treatment discontinuation, the results are very likely to be useful in the area of clinical trial design by assisting in efficient selection of eligible patients. In general, patients with baseline chronic conditions are at risk of developing treatment adverse events and their condition can introduce confusions when assessing the new treatment efficacy. Furthermore, when testing the efficacy of a drug, patients' quality of life and potential treatment toxicity effects should be carefully considered. Accordingly, making the reasonable recruitment criteria is a challenging part for most clinical trials determining their success or fail in demonstrating the treatment efficacy. Here, we conducted a comprehensive simulation study to estimate the ability of ENTHUSE 33 trial design in demonstrating the efficacy of docetaxel while incorporating the information regarding the patients with high risk of developing adverse event during the trial recruitment phase. The simulation analysis revealed that when the risk of early discontinuation was considered in trial eligibility criteria, fewer patients were required for the trial without loss of statistical power. For example, without information about discontinued patients, around 1,548 patients were required to detect an HR of 1.30 at 80% statistical power and false discovery rate of 0.05. However, when selection into the trial was based on the postchallenge ensemble-based model, 1,306 patients were sufficient to detect an HR of 1.30 with similar power and significance level.

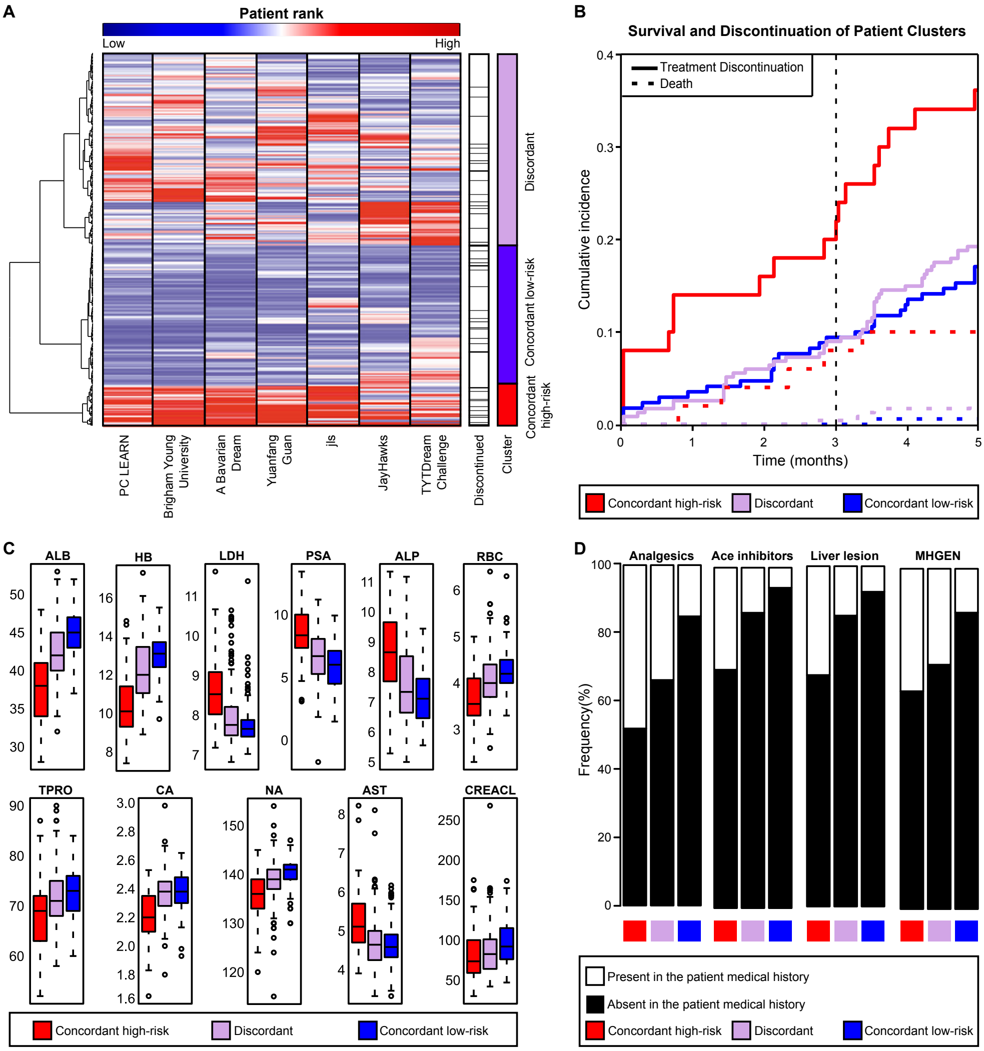

Figure 9. Meta-analysis of risk scores reported by the seven top-performing teams. (A) Unsupervised clustering of the patients' risk scores (columns) from the validation data (ENTHUSE 33, n=470) computed across the seven top-performing teams. (B) KM curves representing the association between risk groups and study outcome (death or treatment discontinuation). (C) Boxplots representing the distribution of laboratory test values found to be significantly different among the risk groups. (D) Distribution of binary predictors including prior medical and medications found to be significantly different among the risk groups. ACE, angiotensin-converting enzyme; ALB, albumin; ALP, alkaline phosphatase; Ca, calcium; CREACL, creatinine clearance; HB, hemoglobin; LDH, lactate dehydrogenase; MHGEN, medical history: general disorders and administration site conditions; Na, sodium; PSA, prostate-specific antigen; TPRO, total protein. The figure is adapted with permission from publication III.

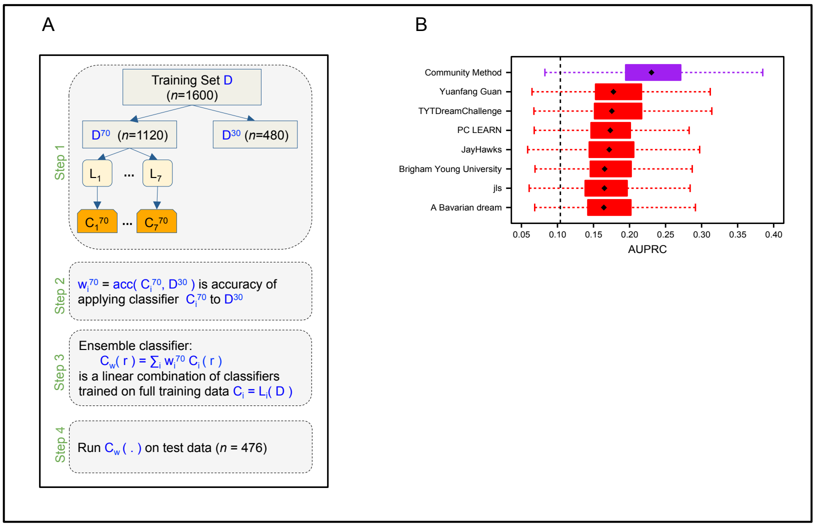

Figure 10. Postchallenge ensemble model development workflow and its performance compared to top-performing teams. (A) Schematic workflow illustrating the different required steps to develop the post-challenge ensemble-based prediction model. (B) AUPRC values from applying the top-performing and postchallenge ensemble models to the validation data (ENTHUSE 33, n=470). Diamonds represent the observed AUPRC and the horizontal boxplots and corresponding whiskers represent the distribution of AUPRC values from applying the models to 5000 bootstrap samples from the validation data. The vertical dotted line represents the performance a totally random model. The figure is adapted with permission from publication III.

3.7 Assessing the generalizability of trial-based prediction models in real-world data: Publication IV

The utility of trial-tailored prognostic markers in everyday practice is controversial. The DREAM Challenges community has proposed robust prognostic models to predict the survival of patients with mCRPC. Using a real-world cohort from Turku University Hospital, I evaluated the reliability of proposed prognostic models in clinical practice. Precise prediction of survival is a key factor in treatment plan and patient quality of life in mCRPC.

3.7.1 Materials and methods

Data. In this manuscript, we evaluated the reliability of trial-based prognostic models in real-world mCRPC patients. The prognostic models were trained using three clinical trials data (n = 1600 patients) and validated in an independent validation trial (n = 470 patients). The performances of the developed models were evaluated using the integrated time dependent area under the curve (iAUC) by the challenge organizers (Figure 11A). Next, we validated the DREAM challenge top three models and a previously developed model by Halabi and colleagues (referred as the challenge reference model) using an in-house real-world data (Figure 11B). Turku University hospital granted us the unique opportunity of accessing to a real-world (RW) dataset corresponding to the trial-based DREAM dataset. The data was classified into separate eight tables of ICD10 diagnosis values, laboratory values, hospital admissions, pathology records, demographics, chemotherapy and radiotherapy courses and clinical procedures. Patient cohort inclusion criteria included clinical diagnosis of prostate carcinoma (ICD10:C61), antiandrogen therapy (ACT code G03HA) as prior medication and docetaxel-based regimens as first-line chemotherapy treatment. After exclusion of patients with several malignancies, 289 eligible patients were identified for further analysis. In line with the trial-based cohorts, the baseline laboratory values were defined as the last test result within four weeks before chemotherapy administration. Missing laboratory values including lactate dehydrogenase and aspartate transferase were imputed using median values. The missing data from EMR documents was manually collected and confirmed by a physician.

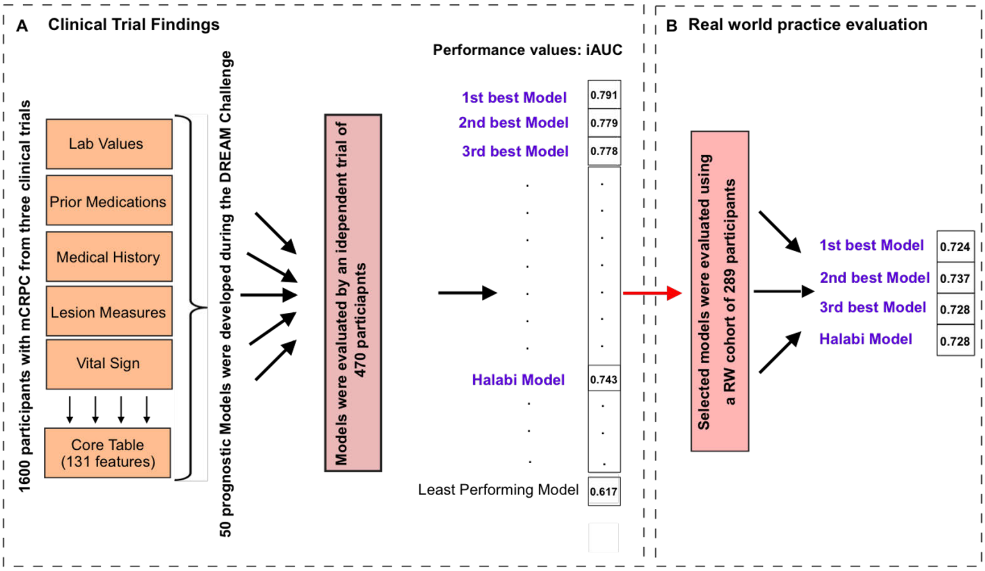

Figure 11. Study design. (A) Baseline clinical variables from three phase Ⅲ RCTs (ASCENT2, VENICE, and MAINSAIL, n = 1600) were used to develop prognostic models for patients with mCRPC. The developed models in addition to the study reference model (Halabi model) were evaluated using an independent RCT (ENTHUSE 33, n = 470) and their performances were reported as iAUC values. (B) Three top-performing teams and study reference model were selected as the state-of-the-art mCRPC prognostic models and their performance was evaluated using a real-world cohort (n = 289) from Turku University Hospital.

3.7.2 Results

Patient baseline characteristics and overall survival. The main goal of this study was to assess the utility of clinical trials tailored prognostic models in real-world patients with mCRPC. As expected, patients in real-world cohort were significantly older and had worse ECOG performance status compared to clinical trial patients. PCA was performed to examine similarities and differences between the real-world and clinical trial cohorts. More specifically, PCA was conducted for the variables from reference model by Halabi and colleagues including analgesic use, metastasis site (defined as lymph node only, bone metastases with no visceral involvement, or any visceral metastases), ECOG performance status, LDH, ALB, HB, ALP, and PSA were used to assess the patients composition in PCA analysis (Figure 12A). Of note, the distribution of key variables was very similar across both real-world and clinical trial cohorts, confirming the reliability Halabi model variables in various cohort groups. Kaplan-Meier analysis revealed that no significant difference in the overall survival rates was observed between the real-world and trial cohorts (log rank P = 0.11, Figure 12B).

Validation and calibration of prognostic models. The top performing prognostic models from the prostate cancer DREAM challenge were externally validated using the Turku University Hospital cohort. For the clinical trial datasets, all the selected prognostic models outperformed the Halabi reference model as confirmed by integrated AUC (iAUC) values (Figure 13A). In real-world data validation analysis, there was not a significant outperforming model (BF < 3 for all comparisons) but all the models including the Halabi reference model performed similarly well (Figure 13B). In general, the performances of models were slightly lower in real-world data with iAUC values ranging from 0.743 to 0.792 but the performance remained steady over time which is a positive sign in term of model reliability. Furthermore, the models' calibration was performed by comparing the predicted with the observed survival at 18, 24, 30 and 36 months follow-up times. The calibration plots demonstrated a relatively high agreement between predicted and observed survival proportions especially at 24 and 30 months (Supplementary Figure 2, publication IV). Despite the significant differences in some critical predictors including age and ECOG performance status, the discriminative potency of models was still acceptable ensuring that the recent trial-based prognostic models can be safely used for survival stratification in real world mCRPC patients.

Next, I assessed the generalizability of the prognostic models in terms of the availability of utilizes features and the way they were collected in real-world cohort. Notably, the number of utilized features markedly varied from eight features to 91 features in Team 3 and Team 1 models respectively (Table 1). Clinical trials follow a restrict conventional setting for data collection which will be implemented for all the recruited patients. However, with real-world data, clinical features are collected or recorded based on the patient individual condition and healthcare provider routine. For example, while LDH was utilized by all the tested prognostic models, it was fully missing from the real-world data as LDH test has not been among the routine tests for prostate cancer in Turku University Hospital. Another example is the significantly lower incidence of lymph node metastases in real-world cohort compared to clinical trial data (26% vs. 48%, $P < 0.001$). However, it should be noted that this comparison is meaningless as lymph lesion evaluation is a routine practice in clinical trials but for the real-world patients it was measured only through pathological examination. Missing data and incompatible data collection strategies are potential source of bias and could adversely affect the statistical power and model performance.

Finally, I reassessed the performance of the prognostic models in trial eligible real-world patients. Following the recruitment criteria derived from clinical trial design protocol, 245 (85%) real-world patients were identified to be eligible for trial inclusion. Most ineligible patients were older and had additional comorbidities. With this RCT eligible subset data, Team 2 model achieved an iAUC of 0.739 and significantly outperformed the other three models (BF > 3, Table 1). Interestingly, during the first follow-up year, there was a significant drop in performance of models (except for Team 2 model) utilizing the new RCT eligible sub cohort compared to the whole real-world data. It was speculated that the performance drop was linked to exclusion of patients with worse overall condition at higher risk of early death consequently easier to predict. In concordance with this hypothesis, a significant drop in survival rate during the first follow-up year was observed 72 % versus 62 % deaths in RCT non-eligible and eligible patients, respectively ($P < 0.01$).

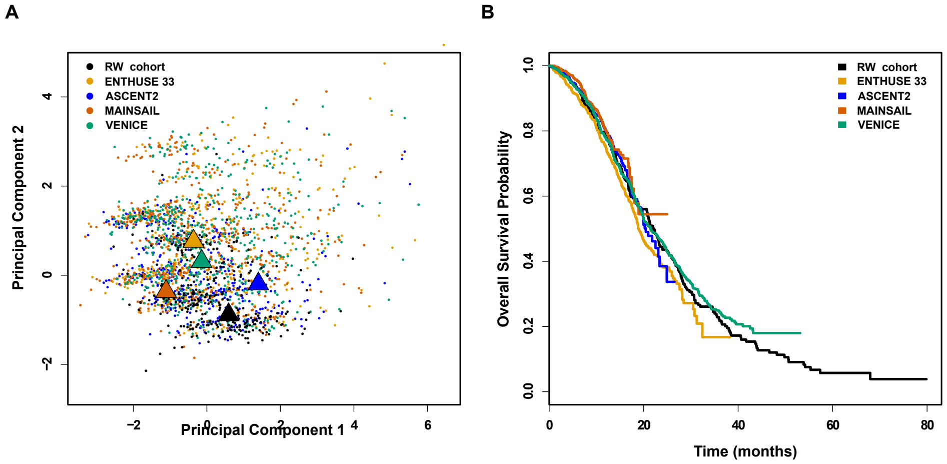

Figure 12. Comparison of distribution of RCT and real-world data in terms of predictive variables and overall survival. (A) PCA of the four RCT datasets and real-world cohort using predictive variables from the Halabi reference model including ECOG performance status, albumin, alkaline phosphatase, hemoglobin, lactate dehydrogenase, prostate specific antigen, analgesics use, metastasis site (defined as lymph node only, bone metastases with no visceral involvement, or any visceral metastases). (B) KM curves representing the overall survival of in RCTs and real-world cohorts. No significant differences were observed between the overall survival of RCTs and real-world patients (log-rank test $P = 0.70$). The figure is adapted with permission from publication IV.

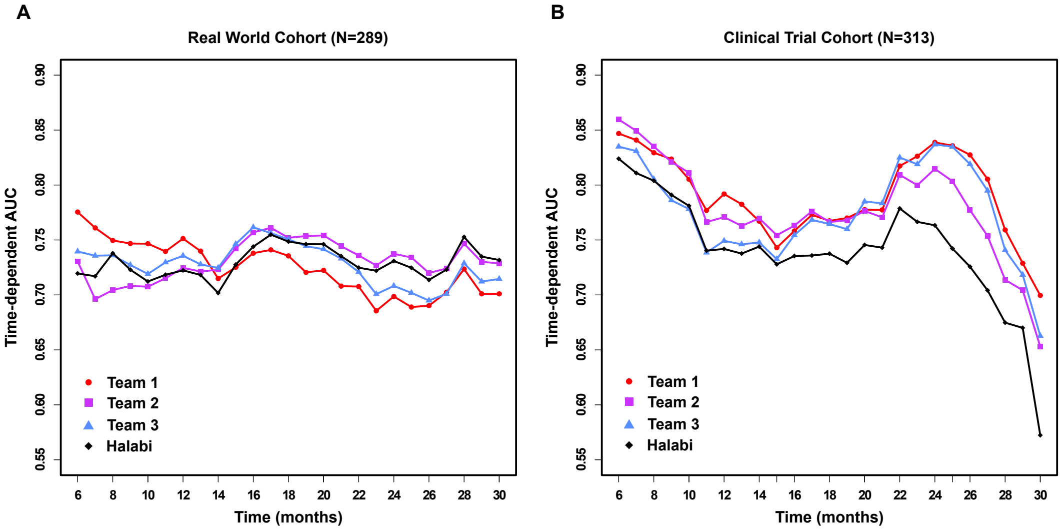

Figure 13. (A) Performance of the RCT based state-of-the-art prognostic models (including three top-performing models from the prostate cancer DREAM challenge in addition to the halabi reference model evaluated by real-world cohort from Turku University Hospital (n = 289). (B) Performance of the RCT based state-of-the-art prognostic models (including three top-performing models from the prostate cancer DREAM challenge in addition to the halabi reference model evaluated by the prostate cancer DREAM challenge validation data (ENTHUSE 33, n = 313). The figure is adapted with permission from publication IV.

Table 1. Number of predictive variables and performance of selected models including the prostate cancer DREAM challenge top-performing teams (Teams 1-3) and the study reference model (Halabi model).

Team	**Model**	**Number of employed features**	**iAUC (6-30 months) in ENTHUSE 33 trial**	**iAUC (6-30 months) in RW cohort**	**iAUC (6-30 months) in RCT eligible RW cohort**
Team 1	Penalized Cox regression with Elastic net	91 independent features plus all their interactions	0·792	0·724	0·721
Team 2	Weighted Penalized Cox regression	22	0·779	0·731	0·739
Team 3	Penalized Cox regression	8	0·778	0·728	0·710
Halabi	Penalized Cox regression with Lasso	10	0·743	0·729	0·712

4 Predicting Disease Risks Using Genetic and Clinical Data

Chapters 2 and 3 exploit molecular (genomic) and clinical data as separate sources for development of computational tools in biomedical research. The availability of versatile types of data motivates development of integrative strategies which may potentially improve the performance of proposed mathematical models. This can include integration of same type of data at different levels (e.g. integration of different levels of omics data) or combination of data from different sources such as clinical and genomic data. Generally, with genomic data, preprocessing is more complex and computationally intensive compared to clinical data. Hence, in case of combining genomic and clinical data, effective data curation and standardization is required prior to initiating model development. Differences in volume and dimension in addition to inconsistency across data representations are the further factors which require bioinformatic solutions when combining different data types. When the data is efficiently integrated, appropriate statistical and machine learning methods are applied to extract information from the well-structured multi-level data. In order to better explaining the methodology, in this chapter we propose an ensemble based predictive model which utilizes early life clinical and genomic measurements to predict adulthood obesity. The methods and results from this study are presented in publication V of this thesis.

During the last two decades, the prevalence of obesity has markedly increased worldwide [126]. Obesity is an underlying risk factor for several chronic comorbidities including type 2 diabetes, cardiovascular diseases and cancer [127] [128] [129]. Unhealthy dietary habits and lack of physical exercise are the well-known risk factors in developing obesity. For a large proportion of cases with excessive weight, the problem originates from childhood time. Of note, the risk of becoming an overweight adult could be 3-fold higher among obese children (body mass index (BMI) > 80th percentile) compared with nonobese children [130]. Statistical models have been used for risk stratification of children on the basis of their genetics and clinical and environmental factors. Currently, genome-wide association studies are performed to identify obesity susceptibility genetic variants in different population cohorts [131]. Early detection of at-risk children and precautionary health actions play crucial role in treatment and prevention of adolescent obesity. This aim is followed in publication V where we examine the role of genetic risk factors coupled by childhood clinical and environmental factors in developing adulthood obesity in the Cardiovascular Risk in Young Finns Study (YFS). The YFS is a multi-center, on-going follow-up study initiated in 1980 which aims to study the cardiovascular risk factors and their consequences from childhood to adulthood. The participants were randomly selected from five different cities in Finland (Helsinki, Kuopio, Oulu, Tampere, and Turku) and aged 3 to 18 years old. During the study follow-up time the conventional risk factors have been systematically collected at certain time points at 1983, 1986, 2001, 2007 and 2011. This included health surveys and self-reported questionnaires (e.g. assessing general wellness and health status, life-style factors, diet, alcohol and smoking habits, physical activity and socioeconomic status), physical measurements (e.g. height, weight, blood pressure) and blood tests (e.g. C-reactive protein (CRP), insulin and serum lipoproteins). Additionally, genome-wide association study was performed for individuals participated in 2007 follow-up. The YFS data has been widely used to assess whether chronic health issues such as type 2 diabetes, hypertension and obesity originate from early life stage [132].

In the following sections, first I will give a brief introduction about genome-wide association studies (see section 4.1) as the genomic data source in current predictive analysis. Next, I will give brief overviews about data integration methods (see section 4.2) and obesity associated risk factors as well as materials and methods utilized in publication V (see section 4.3). In particular, gradient boosting algorithm was utilized over the most updated list of obesity-related genetic variants combined with clinical and environmental factors (confirmed by Juonala et al. [133]) to predict adulthood obesity. The findings of this study were validated with an independent dataset. Finally, the last section presents the novel clinical findings of this study (see section 4.4). Compared to previous similar studies [133], this study includes larger sample size and longer follow-up time contributing to a higher confidence level and more statistical power. Additionally, we used the most updated list of genetic risk factors from the largest GWAS experiment at the time of our study [134]. This gives us better possibility of detecting interactions between genetic risk factors responsible for development of obesity. Finally, we utilized boosting algorithms as one the most powerful machine learning tools for model training and the study novel results confirms the importance of our methodology.

4.1 Genome-wide association studies

Whole-genome sequencing (WGS) allows to study the complete set of an organisms' genome at the DNA base level in a robust and cost-effective way. Genome-wide association studies (GWAS) aim to investigate the population-specific genetic variants associated with complex traits and diseases. In this case, frequently occurred single base differences in DNA sequence of an organism, commonly referred as single nucleotide polymorphisms (SNPs), are supposed to be the variant causative for developing a phenotypic difference when compared to control population. In human, SNPs have two alleles or variant forms (one from each parent) and so in respect of each gene, a person is either homozygous with identical alleles or heterozygous with two different alleles. Once the DNA is sequenced, statistical methods are required to confirm the difference in the frequency of each allele between the case and control groups. In practice, GWAS experiments include relatively modest effect size; therefore, the study sample size can be a determinant factor to obtain the desired statistical power. Moreover, the power of a GWAS analysis is directly affected by the frequency of candidate alleles associated with a trait of interest. For common diseases (e.g. type 2 diabetes) or common disorders (e.g. obesity), GWAS assume that the causative allele should be also very commonly observed across the case sample group [135]. Accordingly, if GWAS discovers a genetic variant which is not as common as the associated trait, it may be concluded that a group of (relatively few) SNPs rather than a single one with overall higher frequency has caused the trait. However, it should be considered that the discovered trait-associated SNPs might be false positives in linkage disequilibrium (LD) with true trait-associated SNPs that are not detected by GWAS analysis. LD is a genetic phenomenon stating that two SNPs from same chromosome with short distance from each other remain physically in the same location within a population. Therefore, the LD between the detected genetic variant and the possibly undiscovered variant can reveal an indirect association which requires further considerations.

When the experimental and computational prerequisites are settled, association analysis are performed to assess the relationship between a trait or a disease and genetic variants. With GWAS design, often logistic regression algorithms (due to binary outcome e.g. disease or healthy) are applied to estimate the significance of variant-trait associations. In this case the variant-trait association is evaluated using the regression coefficients (β: log (odds ratios)) and their corresponding level of significance. Some publications report odds ratios (coupled with chi-squared test) rather than β values as the main measure of association and SNPs effect size. For example, for patients with type 2 diabetes, the odds ratios for hypothesized SNPs associated with this disease are estimated as ratios between patients with and without the particular SNPs. Of note, in most GWAS experiments (except for studies focusing on rare traits or diseases) the general odds ratios reported for final list of associated SNPs are close to 1 (although they may include highly significant p values). To avoid confusing issues influencing the reliability of acquired results, three factors should be taken into account. First, complex diseases or disorders (e.g. obesity studied in publication V) are supposed to be polygenic meaning that the trait is subjected to cause by a network of interacting genes (SNPs) rather than a single variant. Hence, the cumulative effects from detected SNPs are evaluated to make conclusions about the study findings. The second important factor when interpreting the results is the experimental sample size. Studies with large sample size represent improved statistical power by detecting larger number of associating SNPs. For example, when investigating association of genetic variants and body mass index, Locke et al. [134] were able to increase number of detections from 32 SNPs in their previous study [136] to 97 SNPs associated with obesity when they increased the sample size from 249,769 to 339,224 in their latest study. Finally, always it should be reminded that the detected SNPs with very small odds ratios could simply represent the population heterogeneity which have no effect on the trait under study [137] [138].

4.2 Methods to combine molecular and clinical features

The predictability role of non-genetic factors such as clinical, environmental and socioeconomic variables has been extensively proved in various medical applications. For example, laboratory measurements, diagnostic imaging and pathological staging are the main traditional factors in cancer prognosis [139]. Although highly informative, these factors are not sensitive enough for a personalized and accurate adjuvant therapy plan. One solution to achieve a more accurate estimation of patient outcome, is to integrate genetic data into their clinical profile. Several studies suggest combination of clinical and genetic factors play complementary role in better understanding of the disease state and characteristics [140] [141] [142]. With majority of integrative studies, the aim is to combine the often high-dimensional data into EMRs collected for clinical trials or from hospitals and healthcare systems. The information combination task can be done following two approaches. During the first modeling approach, separate models for clinical and genetic data are developed and the patient final risk score is obtained by fusion of the individual models. A simplest way to fuse the individual risk scores is to calculate the average risk from the separate prediction models [143]. The second common integrative strategy is to treat clinical and molecular variables uniformly to develop a single predictive model. A well-known example are studies that combine data dimension reduction step to their modeling workflow. This way, the often high dimensional genetic data can be transformed into a genetic risk score or signature and then will be incorporated to the clinical variables set to develop the prediction model [144]. In publication V, I proposed a weighted genetic risk score to be added to the clinical features before building the predictive model.

4.3 Predictive value of obesity associated risk factors: Publication V

Previous studies have categorized obesity as a multifactorial trait with several causing factors correlating or interacting with each other. These factors can broadly be divided into two main categories: genetic risk factors and clinical and environmental risk factors. From genetic point of view, obesity is a polygenic heritable trait meaning that several alleles with small individual and considerable cumulative effect are involved in becoming obese. The most updated human GWAS study by Locke et al. [134] has identified 97 BMI associated loci of which 41loci have been previously shown to be associated with obesity. The main childhood clinical risk factors include parental obesity, increased birth or childhood BMI and early puberty. Imbalanced total calorie intake, low physical activity, maternal smoking and low socioeconomic status are the well-known environmental risk factors. In publication V, the aim was to answer the question of whether the integration of genetic risk factors into predictive analysis can improve our capacity to detect onset of adulthood obesity. To answer this question, the SNP values and their corresponding β coefficients were transformed to weighted genetic risk scores suitable to integrate with clinical and environmental risk factors. In this study, I utilized gradient boosting algorithms to create our predictive model and validate it with an independent validation set. It was based on the hypothesis that the updated genetic variants list together with longer follow-up time will increase the statistical power and performance of the proposed predictive model.

4.3.1 Materials and methods

Data. A total of 2262 participants with complete set of genotype data and baseline clinical and environmental variables were selected from the YFS. The study outcome, adulthood obesity was defined as BMI $\geq$ 30 (kg/m^2) and the baseline explanatory variables included age, gender, baseline BMI (calculated as weight (kg) divided by height in meters squared and adjusted for age and gender), parental BMI, family income and childhood CRP measurements. It should be noted that, although previous studies have shown significant effect of CRP in predicting adulthood obesity, it was not among our final list of predictive variables as we did not observe any performance improvement while utilizing CRP measurements. The reader is referred to Table 1 publication V for detailed information about baseline clinical variables. The follow-up time for this manuscript data was 31 or 32 years. The SNP genotyping was performed with the Illumina BeadChip 670K and the Illumina clustering algorithm was used for genotype calling [145]. Imputation of genotypes was performed using IMPUTE2 software package [146] and the 1000 Genome Phase 1, Version 3 as reference panel [147]. The genetic risk factors utilized in this study included the 97 significant genome-wide SNPs that have been reported by Locke et al. [134] in their latest study using data on 339,224 individuals. The participant cohort was randomly split into training (n=1625, 72%) and validation (n=637, 28%) subsets to avoid overfitting and ensure the generalization ability of the final model.

Methods. In publication V, I examined the contribution of clinical factors integrated with a genome-wide polygenic score with adulthood obesity. Differences in the distribution of obese and nonobese participant characteristics were determined using Wilcoxon rank-sum test and χ^2 test for continuous and categorical variables respectively. The explanatory effect of the genetic risk factors was integrated into a weighted genetic risk score which was defined as the arithmetic sum of SNP values (x) weighted by their corresponding β scores as follow

$$WGRS = \sum_{i=1}^{n} \beta_i x_i,$$

where β scores were derived from the original study by Locke et al. For this publication, we calculated two weighted genetic risk scores $WGRS97$ and $WGRS19$, where $WGRS97$ denoted the weighted risk score depending on the whole 97 SNPs reported in the original study and $WGRS19$ utilized a subset of 19 SNPs out the 97 SNPs we found more relevant using univariate regression analysis (Supplementary Table 1, publication V).

The gradient boosting algorithm implemented in the R package *gbm* was used to develop the prediction model. The reader is referred to section 3.4 for detailed method description. In order to reduce overfitting effects, regularizing parameters including learning rate and subsampling fraction were fine-tuned. To capture interaction effects between up to two variables, the decision stumps was set to three. Five-fold cross validation was used to penalize model overfitting. The AUC values were used to assess the performance of predictive models. The effect of $WGRS19$ on BMI trajectories was further explored for different age groups. More specifically, the participants were divided into four groups based on their $WGRS19$ quartiles and comparative analysis using Wilcoxon rank-sum test was performed to examine the age-related effect of genetic variants. Here, the BMI trajectory plots represented the mean BMI of all study participants with measured BMI values at each given age in each defined quartile.

4.3.2 Predictive analysis results

When examining the association of childhood clinical factors and adulthood obesity, it was observed that obese participants had significantly higher childhood baseline BMI, maternal BMI and WGRS19 compared to nonobese participants (*P*<0.0001). Out of 97 SNPs reported by Locke et al., 19 SNPs had P values < 0.1 in the univariate regression analysis and were selected as an independent covariate in the form of weighted genetic risk score WGRS19 (Table 3). Gradient boosting algorithms were utilized to investigate the association of genetic and childhood clinical risk factors with adulthood obesity. Comparison of AUC values demonstrated that combining the genetic risk factors (WGRS19) and clinical characteristic improves prediction accuracy also in validation data (AUC=0.769 versus AUC=0.747, *P*=0.026) when compared with clinical factors alone (Table 2). Next, the study participants were divided into 3-year age groups for further age-wise model performance evaluation. When the genetic risk factors were incorporated to the model, the model performed significantly better in the youngest age group (3-6 years) also in the validation data (AUC=0.771 versus AUC=0.700, *P*=0.002). However, combination of clinical and genetic risk factors did not significantly improve the model performance in older age groups when the model was tested in the validation data.

Finally, the association of the genetic risk scores WGRS19 and WGRS97 with early onset of excessive BMI incidence was investigated (Figure 14). The participants were divided into quartiles based on their WGRS19 and WGRS97 scores and participants from the highest and lowest quartiles were selected as BMI trajectory groups for further analysis. The comparative analysis revealed that the genetic risk scores act as significant stratifying factors across BMI trajectory groups. More specifically, BMI trajectory groups were significantly different starting from age of 9 and 6 for WGRS19 and GWRS97 respectively (*P*<0.05, Kruskal–Wallis test). These results of this study support our hypothesis about association of genetics and childhood baseline characteristics with adulthood obesity.

Table 2. Performances of the developed models

Training data (n=1625)					
	WGRS19 and clinical factors	WGRS97 and clinical factors	Clinical factors	WGRS19 vs. clinical	WGRS97 vs. clinical
Age	AUC	AUC	AUC	p	p
All	0.787	0.782	0.744	<0.0001	<0.0001
3-6 y	0.754	0.736	0.692	<0.0001	<0.0001
9-12 y	0.809	0.805	0.777	0.002	<0.0001
15-18 y	0.793	0.795	0.752	0.001	<0.0001

Validation data (n=637)					
	WGRS19 and clinical factors	WGRS97 and clinical factors	Clinical factors	WGRS19 vs. clinical	WGRS97 vs. clinical
Age	AUC	AUC	AUC	p	p
All	0.769	0.749	0.747	0.026	0.785
3-6 y	0.771	0.742	0.700	0.002	0.020
9-12 y	0.799	0.767	0.786	0.293	0.049
15-18 y	0.734	0.738	0.740	0.743	0.922

Table 3. Univariate logistic regression analysis for SNPs with P < 0.1.

SNP	Chr	Position	Gene	OR	P	Ranking in Locke et al.
rs2207139	6	50,953,449	TFAP2B	1.326	0.001	6
rs16951275	15	65,864,222	MAP2K5	0.727	0.004	15
rs2112347	5	75,050,998	POC5	0.809	0.006	17
rs1558902	16	52,361,075	FTO	1.203	0.014	1
rs492400	2	219,057,996	USP37	0.835	0.015	56
rs11688816	2	62,906,552	EHBP1	0.889	0.019	73
rs7239883	18	38,401,669	LOC284260	1.321	0.035	69
rs9400239	6	109,084,356	FOXO3	1.171	0.042	71
rs16851483	3	142,758,126	RASA2	1.082	0.0491	44
rs9641123	7	93,035,668	CALCR	1.183	0.0762	42
rs17724992	19	18,315,825	PGPEP1	0.847	0.078	89
rs2245368	7	76,446,079	PMS2L11	0.855	0.079	87
rs1460676	2	164,275,935	FIGN	1.134	0.081	96
rs11727676	4	145,878,514	HHIP	0.845	0.0849	79
rs7164727	15	70,881,044	LOC100287559	1.197	0.088	53
rs13078960	3	85,890,280	CADM2	1.134	0.09	21
rs9374842	6	120,227,364	LOC285762	1.145	0.0906	81
rs11057405	12	121,347,850	CLIP1	1.05	0.0922	74
rs12885454	14	28,806,589	PRKD1	0.881	0.096	41

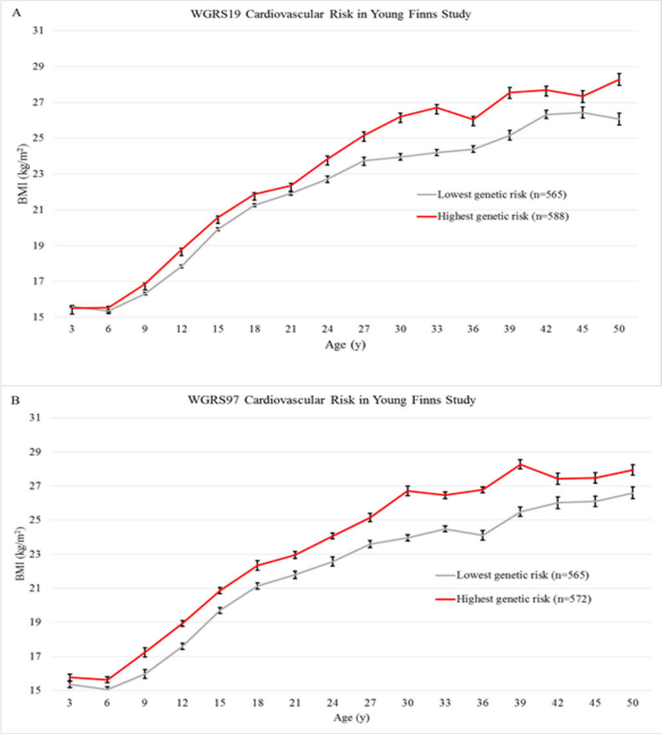

Figure 14. The BMI trajectories in YFS participants for low-risk and high-risk groups according during the whole study period. A) The BMI trajectories of high-risk and low-risk groups based on WGRS19 score and B) the BMI trajectories of high-risk and low-risk groups based on WGRS97 score. The figure is adapted with permission from publication V.

4.4 Clinical findings

Early life properties and conditions may affect individuals’ health and disease risk in older ages. In this study, we observed that genetic factors are better representors of obesity in younger children for longer follow-up periods. Also, this analysis revealed the significant contribution of childhood BMI, maternal BMI and family income in development of adulthood obesity. When using weighted GRS, higher prediction accuracy was achieved using the reduced $WGRS19$ form rather than the complete $WGRS97$ risk scores. It was concluded that $WGRS97$ score may include variants whose effects have been revealed already during childhood time; therefore, its predictive ability decreases compared to clinical factors. Considering the refined $WGRS19$ score, its influence on BMI was manifested between 9 to 12 years of age; whereas, for $WGRS97$ score the effect was manifested at 6 years of age. This is in line with previous findings that the influence of genetics on adulthood obesity is already manifested during childhood [148] [149]. For instance, Hakanen et al. [150] showed that the influence of a very well-known obesity related gene, FTO, is already manifested at age of 7 years which supports findings of current study. In contrast, the effect of genetic factors in older children is mitigated; and clinical factors play stronger role in predicting adulthood obesity for individuals between 12 to 18 years of age. Now that we know genetic data may assist in early identification of individuals with high risk of obesity, this information can be used to propose personalized precautionary plans. In fact, the decreasing trend in genome sequecning experiments costs may result in utilization of genetic variants in clinical practice in near future. Accordingly, studies similar to ours, can be implemented in everyday routine for identification of susceptible individuals and providing them clinical preventive services.

5 Discussion

Early prediction of diseases developmental stages has direct effect on treatment decision-making and outcomes. In this thesis, I developed several statistical and machine learning methods for patient risk stratification with case studies in cancer and obesity. Despite the significant advancements in management and therapy, cancer has remained a leading cause of death worldwide. The field of computational biology has invested remarkable efforts to address the challenges of the field. Bioinformatics software tools provide crucial insights into genetic and epigenetic mechanisms of tumorigenesis and cancer progression. Various types of mathematical models are developed to predict metastasis, patient prognosis as well as drug resistance and response with significant clinical impact. The gene expression profiles provided by high throughput technologies are a central source of information in cancer research. The dynamic nature of transcriptome can explain major differences with causal effect on disease development and state. More specifically, genetic mutations and dysregulation of gene expression profiles may disrupt the cellular function and increase the chance of developing malignancy. Gene expression profiles are mainly used to classify patients into distinct phenotypic subtypes with specific treatment plans. Publications I and II in this thesis, focused on task of transcriptome profiling and biomarker discovery using RNA-seq data.

Today, RNA-seq technology is the standard approach for transcriptome characterization. This technology is extensively used for global quantification of gene expression in all bioscience areas including cancer research. From bioinformatics point of view, a large set of software packages are developed for detection of differential expression in RNA-seq studies. Here, the expression levels are represented in the form of integer read counts. Accordingly, most of the state-of-the-art methods (e.g. DESeq2, edgeR and baySeq) use negative binomial distribution to model the expression levels. Another practically appropriate approach is to apply suitable transformation strategies so that the transformed data would fit in a normal distribution system (e.g. Limma); and then model the gene counts on the basis of normal distribution assumptions. When the model is fit, appropriate statistical tests (e.g. modified t-test in ROTS and Wald test in DESeq2) are utilized to assess the differences in expression levels across sample groups. Finally, the significance of detections is corrected for multiple testing to control the false discovery rate (e.g., using Benjamini–Hochberg method). The processed list of significant detections from differential expression testing may further serve as disease biomarkers or predictive signatures in diseases studies. Unfortunately, it is very likely to obtain inconsistent set of biomarkers using different statistical testing methods as observed in publication I. This particularly can be a challenging issue if the study sample size is small or high level of heterogeneity exists between replicate samples. Additionally, careful considerations are required when including very lowly expressed genes in a predictive signature as they are at higher risk of being false positives induced by RNA-seq technical biases and limitations.

The significant inconsistency observed among the available statistical methods demonstrated the need for a data-adaptive method to ensure the reproducibility of findings in sensitive medical applications. The problem of selecting a suitable statistic was addressed by proposing ROTS in publication II. In contrast to the parametric methods tested in publication I, ROTS statistic is optimized directly based on the input data by maximizing the reproducibility of the selected genes (or transcripts) across bootstrap samples. In order to prove the efficacy of ROTS in real biomedical applications, the method is used to identify prognostic markers in ccRCC. ccRCC tumors are highly heterogenous and always there is a risk that the expression profile of an identified marker may vary from patient to patient. In treatment plan of ccRCC patients, disease stage and patient prognosis are the main deterministic elements. With the aim of assisting clinicians to make advance treatment plans, I proposed a prognostic signature for survival stratification of ccRCC patients using ROTS method and complementary statistical procedures. In addition to known markers such as key genes in glucose metabolism (e.g. *ALDOB*, *G6PC* and *PKLR*), the proposed signature includes novel markers with potential prognostic and therapeutic effects. These include, for instance the genes regulated by the pVHL-HIF pathway (EPO, REN, IGFPB1 and FABP1) known to be dysregulated in ccRCC and several solute carrier family members (e.g. SLC38A5) which are known to supply crucial glutamine to cancer cells and contribute in malignant growth. Utilizing this gene expression signature, I was able to classify patients into poor (< 12 months) and better (> 60 months) prognosis groups. To account for tumor heterogeneity and assess the reproducibility of the findings, the proposed signature was validated using an independent cohort of ccRCC patients.

Despite the increasing trend of demanding genetic-based markers in cancer management, traditional clinical parameters are still appeared to be the main index to rule out treatment strategy. In mathematical modeling of cancer-specific outcomes, machine learning algorithms are becoming an increasingly popular approach to capture linear and non-linear relations between clinical variables. For these studies, limited sample size is a big barrier in development of accurate and generalizable predictive models. While the current systematic workflow of data management coupled with new data sharing policies could assist in obtaining a suitable sample size, many models still lack sufficient statistical power to support their claims. This is specially challenging if we are dealing with an imbalanced cohort where the target subjects (e.g. patients with clinical outcome of interest) belong to the rare class. Here, maximizing the global accuracy is not necessarily valuable as it may downweigh the model performance in detection of truly interesting subjects. A common approach to handle the problem of imbalanced data is utilizing boosting techniques to develop an ensemble learner rather than a single model with high global accuracy. In publication III, an ensemble-based model is proposed to predict the clinical event of interest for an imbalanced cohort.

Participating in Prostate Cancer DREAM Challenge, provided me the opportunity of accessing and analyzing clinical data from a relatively large cohort of mCRPC patients. The utilized data were from the comparator arms of four phase III clinical trials in first-line mCRPC with a total of 2,070 patients. For these patients, the challenge participants had access to laboratory values, lesion measurements, prior medications, medical history, and vital signs with trials follow-up details. With the aim of improving the survival rate and quality of life, these patients were treated with a docetaxel-based chemotherapy regimen. However, as discussed in chapter 3, around 20% of mCRPC patients experience toxicity-induced adverse events and have to prematurely discontinue their treatment plan. The Prostate Cancer DREAM Challenge aimed to develop predictive models to estimate the survival and risk of early docetaxel discontinuation in mCRPC. The publication III of this thesis provided a community-based approach to predict patients with higher risk of developing adverse events. Early identification of high-risk patients for short-term treatment discontinuation is crucial for enhancing their quality of life. As only 10% of patients were at risk of developing early adverse events, the problem was approached as an imbalanced situation and an ensemble learner was proposed to predict high risk patients. The proposed model utilizes the baseline clinical variables to assess treatment discontinuation. The main predictive features included ECOG status, liver lesions, use of analgesics and angiotensin-converting enzyme inhibitors and laboratory values of HB, ALB, LDH, PSA, RBC, calcium, AST, creatinine clearance, and total protein. Although the performance of the proposed model was modest, our results can serve as an initial step in development of similar tools in clinical practice.

In addition to early treatment discontinuation, the Prostate Cancer DREAM Challenge aimed to improve the prediction of survival in patients with mCRPC. Precise estimation of survival is an important factor that influences the choice of treatment and patient's quality of life. Since the survival models were developed using trial-based data, a critical question was how reliable are the proposed models in real-world patients. This is a serious oncology concern as a large proportion of real-world mCRPC patients are of advanced age with severe extra comorbidities and so very likely different from RCT eligible patients. Accordingly, it is not clear if the RCT participants can represent the entire mCRPC patients or not. On the other hand, validation of trial-based models in real-world setting is a challenging issue due to various limitations associated with data properties. In order to fairly validate the prediction models, medical records of real-world patients should be standardized to have a uniform format as RCT data. With data standardization, imputation of missing data is a challenging task where either entire or fraction of values for a predictor are not measured or accessible. Publication IV of this thesis provides an example for validation of trial-based prediction models in read-world data. Interestingly, our results suggested the strength and generalizability of examined methods [10] in real-world patients. In an optimal setting, real-world and trial-based data could be used as complementary sources for development of prediction models.

The prediction models in publications II-IV were developed for better management of cancer as a complex disease. Typically, complex diseases are caused or influenced by a combination of genetic and non-genetic risk factors such as clinical and environmental factors. Considering this multifactorial nature, it is expected to obtain better performance when integrating genetic and non-genetic variables in model development. In practice, however, combination of genetic and non-genetic factors is not straight forward [151]. The first challenging issue in integrative frameworks is data heterogeneity. In non-genetic features, the heterogeneity may arise from data collection and registration protocols. A very well-known example in medical applications is the inconsistency observed in self-report surveys [152]. Similar problematic issue may occur when defining specific clinical conditions or outcomes. For instance, different studies may use various time intervals to define early treatment discontinuation. Additionally, missing data, outlier values and measurements with non-uniform scales are common barriers to data integration. Therefore, precise standardization procedures are required to have a uniform data for model development. For genetic data, the main source of heterogeneity is biological variation which should be estimated precisely before data modeling. Moreover, suitable dimensionality reduction methods are required to handle sparse or large-scale genetic data. Publication V of this thesis provides a successful example of combining genetic and environmental factors to improve the prediction ability. In this paper, I proposed a novel ensemble-based model to predict adulthood obesity using both genetic and social-environmental factors. Using gradient boosting machines, the significant contribution of genetic risk factors, childhood BMI, maternal BMI and family income in development of adulthood obesity was confirmed.

Overall, the main goal of biomedical modeling projects is to translate the often high dimensional and noisy molecular and clinical data into medically enriched findings. Despite significant advancements in model development, clinical utility of identified markers and predictive signatures is still unclear. Obviously, traditional data analysis methods are not sufficient to process today's large and complex data and advanced statistical and machine learning methods are required to address the challenges of the field.

Acknowledgements

This dissertation was carried out at the Medical Bioinformatics Centre in Turku Bioscience Centre, University of Turku during 2014 to 2018.

For most, I would like to express my deepest gratitude to my supervisor Professor Laura Elo for her essential scientific inputs and guidance without those I was not able to accomplish this work. Laura, thank you for accepting me as a member of your unique research group and for your continuous support throughout the years. I am thankful to my research director Professor Marko Mäkelä for his guidance and prompt responses to my inquiries and requests.

I sincerely thank Professor Merja Heinäniemi, from University of Eastern Finland, and Associate Professor Hedi Peterson, from University of Tartu, for their detailed reviews and constructive comments on my dissertation. I thank Assistant Professor Ewa Szczurek, from University of Warsaw, for accepting the invitation to be the opponent of this dissertation.

The University of Turku Graduate School (UTUGS) is gratefully acknowledged for funding and supporting my dissertation. I wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

I want to thank all my co-authors and colleagues at Turku Bioscience Centre for their valuable efforts and fruitful collaborations, as well as the joyful time we spent together. In particular, I want to acknowledge Asta Laiho, Anna Pursiheimo, Tomi Suomi, Arfa Mehmood, Ye Hong, Deepankar Chakroborty, Veronika Suni, and Krista Rantanen. I also would like to thank Dr. Jukka Lehtonen for thorough technical advices on data management and high-performance computing.

I am immensely grateful for the privilege of working with many clinical experts whose contribution is an essential asset to this dissertation. I want to specially thank Prof. Panu Jaakkola, Prof. Sirkku Jyrkkiö, Prof. Tarja Laitinen, Prof. Outi Hirvonen and Prof. Olli Raitakari for providing the suitable data to test methods proposed in this work and for their valuable bio-clinical insight to the findings of this dissertation. I extend my thanks to Prof. James C. Costello and organizers of Prostate Cancer DREAM Challenge and my co-authors in publication III for the great impact on expanding my machine learning skills.

I would like to express an especially heartful thanks to Eija Nordlund, the former coordinator of master programme in Bioinformatics at University of Turku for her constant encouragement and help during my studies. Along with this, I wish to thank Dr. Martti Tolvanen our great lecturer and instructor for inspirational courses and discussions during master studies.

I owe a great deal of thanks to my groupmates and friends in Turku: Maria Jaakkola, An Le thi Thanh, Jarkko Peltomäki, Daniel Laajala and Mehrad Mahmoudian. I appreciate our joyful social-togethers as well as our scientific and non-scientific discussions. Maria, since the first time we tried that Iranian New year dish in my flat till today, there have been moments I am not sure how could I have survived without you. Just as simple and deep our friendship is: thank you! An, I admire your courage and calmness and the way you care about other people around you, including me. I have always valued our friendship. Jarkko, thank you for being a real friend and great company during happy and also sad moments. Daniel, I have had the privilege of finding in you, the many facets which are so very rare in the aspects of one's character. I truly believe the world could have been a much better place with more scientists like you. Mehrad, thank you for being a joyful company and for always offering to help without me even having to ask. I would also like to thank my friends from outside academia, Heta Aali, Johanna Vikman-Toivio, Matti Toivio, Salla Nivala and Zahra Abbaszadeh for their support and all the great times we spent together.

I would like to warmly thank Tiina Ahlsten and Heikki Jaakkola for extending your hand in friendship to my family. You have opened up your home and created a sort of connection which has forged a family atmosphere in a foreign land for us. Thank you for your kindness and care.

I sincerely thank my parents and sisters for their love and support. My beloved mom, during the last 10 years there have been dozens of moments which I wish I could have shared with you, including this one. I admire you for always having been a source of inspiration and how you always maintained a sense of hope; even in the face of all difficulties along the way. I admire your beautiful lust for life, your open-mindedness and trustworthy you have been always. I wish I could tell you how grateful I am to be your daughter. I love you and always will. Marzieh and Razieh, my wonderful sisters, thanks for knowing when I am not fine and not letting me down when I have worries. I love you and your support means a lot to me. I am thankful to my grandmother, Batool, for her unconditional love and care.

Last but not least, I am grateful for my amazing family more than anything in my life. Soren, my little happy boy, words are not enough to express how thankful I am for being your mom. Loving you is the only and only thing that would never change in me. I am grateful to my best friend and the love of my life, Mohammad, for sharing your journey with me. Thanks for being there for me with your encouragement and unconditional support no matter what. Thanks for knowing me better than I know myself. Thank you for all the brilliant moments and memories we have made together and will make. I hope you would always remember me by this:

دیده از دیدار خوبان برگرفتن مشکل است
هر که ما را این نصیحت می‌کند بی‌حاصل است
گر به صد منزل فراق افتد میان ما و دوست
همچنانش در میان جان شیرین منزل است

December 2021
Fatemeh Seyednasrollah

List of References

[1] T. Reichhardt, “It’s sink or swim as a tidal wave of data approaches,” *Nature*, vol. 399, no. 6736, pp. 517–520, Jun. 1999, doi: 10.1038/21044.

[2] H.-S. Kim, S. Lee, and J. H. Kim, “Real-world Evidence versus Randomized Controlled Trial: Clinical Research Based on Electronic Medical Records,” *J Korean Med Sci*, vol. 33, no. 34, Jun. 2018, doi: 10.3346/jkms.2018.33.e213.

[3] M. Munson, “A study on the importance of and time spent on different modeling steps,” *Sigkdd Explorations*, vol. 13, pp. 65–71, May 2012, doi: 10.1145/2207243.2207253.

[4] J. Burdack, F. Horst, S. Giesselbach, I. Hassan, S. Daffner, and W. I. Schöllhorn, “Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning,” *Front. Bioeng. Biotechnol.*, vol. 8, 2020, doi: 10.3389/fbioe.2020.00260.

[5] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front Neurorobot*, vol. 7, Dec. 2013, doi: 10.3389/fnbot.2013.00021.

[6] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, and O. Lyashevska, “Predictive analytics with gradient boosting in clinical medicine,” *Ann Transl Med*, vol. 7, no. 7, Apr. 2019, doi: 10.21037/atm.2019.03.29.

[7] J. Zhao *et al.*, “Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction,” *Sci Rep*, vol. 9, Jan. 2019, doi: 10.1038/s41598-018-36745-x.

[8] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors),” *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, Apr. 2000, doi: 10.1214/aos/1016218223.

[9] A. Bayat, “Bioinformatics,” *BMJ*, vol. 324, no. 7344, pp. 1018–1022, Apr. 2002.

[10] J. Guinney *et al.*, “Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data,” *The Lancet Oncology*, vol. 0, no. 0, Nov. 2016, doi: 10.1016/S1470-2045(16)30560-5.

[11] S. Halabi *et al.*, “Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer,” *J. Clin. Oncol.*, vol. 32, no. 7, pp. 671–677, Mar. 2014, doi: 10.1200/JCO.2013.52.3696.

[12] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4th ed. Garland Science, 2002.

[13] J. A. Reuter, D. Spacek, and M. P. Snyder, “High-Throughput Sequencing Technologies,” *Mol Cell*, vol. 58, no. 4, pp. 586–597, May 2015, doi: 10.1016/j.molcel.2015.05.004.

[14] M. V. Schneider and S. Orchard, “Omics Technologies, Data and Bioinformatics Principles,” in *Bioinformatics for Omics Data: Methods and Protocols*, B. Mayer, Ed. Totowa, NJ: Humana Press, 2011, pp. 3–30. doi: 10.1007/978-1-61779-027-0_1.

[15] A. E. Pozhitkov, D. Tautz, and P. A. Noble, “Oligonucleotide microarrays: widely applied--poorly understood,” *Brief Funct Genomic Proteomic*, vol. 6, no. 2, pp. 141–148, Jun. 2007, doi: 10.1093/bfgp/elm014.

[16] Z. Wu and R. A. Irizarry, “Stochastic models inspired by hybridization theory for short oligonucleotide arrays,” *J. Comput. Biol.*, vol. 12, no. 6, pp. 882–893, Aug. 2005, doi: 10.1089/cmb.2005.12.882.

[17] A. Lachmann *et al.*, “Massive mining of publicly available RNA-seq data from human and mouse,” *Nat Commun*, vol. 9, no. 1, p. 1366, Apr. 2018, doi: 10.1038/s41467-018-03751-6.

[18] M. S. Rao *et al.*, “Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies,” *Frontiers in Genetics*, vol. 9, p. 636, 2019, doi: 10.3389/fgene.2018.00636.

[19] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, “Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells,” *PLoS ONE*, vol. 9, no. 1, p. e78644, Jan. 2014, doi: 10.1371/journal.pone.0078644.

[20] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008, doi: 10.1101/gr.079558.108.

[21] F. Birzele *et al.*, “Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing,” *Nucleic Acids Res*, vol. 38, no. 12, pp. 3999–4010, Jul. 2010, doi: 10.1093/nar/gkq116.

[22] H. Edgren *et al.*, “Identification of fusion genes in breast cancer by paired-end RNA-sequencing,” *Genome Biology*, vol. 12, p. R6, Jan. 2011, doi: 10.1186/gb-2011-12-1-r6.

[23] A. Horvath *et al.*, “Novel insights into breast cancer genetic variance through RNA sequencing,” *Sci Rep*, vol. 3, p. 2256, 2013, doi: 10.1038/srep02256.

[24] S. Li *et al.*, “Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study,” *Nat. Biotechnol.*, vol. 32, no. 9, pp. 915–925, Sep. 2014, doi: 10.1038/nbt.2972.

[25] M. G. Ross *et al.*, “Characterizing and measuring bias in sequence data,” *Genome Biology*, vol. 14, p. R51, May 2013, doi: 10.1186/gb-2013-14-5-r51.

[26] R. Schmieder and R. Edwards, “Quality control and preprocessing of metagenomic datasets,” *Bioinformatics*, vol. 27, no. 6, pp. 863–864, Mar. 2011, doi: 10.1093/bioinformatics/btr026.

[27] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[28] R. K. Patel and M. Jain, “NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data,” *PLoS ONE*, vol. 7, no. 2, p. e30619, Feb. 2012, doi: 10.1371/journal.pone.0030619.

[29] M. Sultan *et al.*, “Influence of RNA extraction methods and library selection schemes on RNA-seq data,” *BMC Genomics*, vol. 15, no. 1, Aug. 2014, doi: 10.1186/1471-2164-15-675.

[30] K. Wang *et al.*, “MapSplice: accurate mapping of RNA-seq reads for splice junction discovery,” *Nucleic Acids Res.*, vol. 38, no. 18, p. e178, Oct. 2010, doi: 10.1093/nar/gkq622.

[31] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[32] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions,” *Genome Biology*, vol. 14, no. 4, p. R36, 2013, doi: 10.1186/gb-2013-14-4-r36.

[33] A. Conesa *et al.*, “A survey of best practices for RNA-seq data analysis,” *Genome Biology*, vol. 17, p. 13, Jan. 2016, doi: 10.1186/s13059-016-0881-8.

[34] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nature Biotechnology*, vol. 37, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41587-019-0201-4.

[35] S. Anders, P. T. Pyl, and W. Huber, “HTSeq—a Python framework to work with high-throughput sequencing data,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015, doi: 10.1093/bioinformatics/btu638.

[36] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference,” *Nat Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.

[37] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nat Biotechnol*, vol. 34, no. 5, Art. no. 5, May 2016, doi: 10.1038/nbt.3519.

[38] C. Trapnell *et al.*, “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation,” *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–515, May 2010, doi: 10.1038/nbt.1621.

[39] B. Li and C. N. Dewey, “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome,” *BMC Bioinformatics*, vol. 12, p. 323, Aug. 2011, doi: 10.1186/1471-2105-12-323.

[40] R. Patro, S. M. Mount, and C. Kingsford, “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms,” *Nature Biotechnology*, vol. 32, no. 5, pp. 462–464, May 2014, doi: 10.1038/nbt.2862.

[41] C. Zhang, B. Zhang, L.-L. Lin, and S. Zhao, “Evaluation and comparison of computational tools for RNA-seq isoform quantification,” *BMC Genomics*, vol. 18, no. 1, p. 583, Aug. 2017, doi: 10.1186/s12864-017-4002-1.

[42] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biology*, vol. 11, p. R25, 2010, doi: 10.1186/gb-2010-11-3-r25.

[43] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing,” *Nucleic Acids Res.*, vol. 36, no. 16, p. e105, Sep. 2008, doi: 10.1093/nar/gkn425.

[44] M.-K. Tilak, F. Botero-Castro, N. Galtier, and B. Nabholz, “Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA,” *Genome Biol Evol*, vol. 10, no. 2, pp. 616–622, Jan. 2018, doi: 10.1093/gbe/evy022.

[45] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, “GC-Content Normalization for RNA-Seq Data,” *BMC Bioinformatics*, vol. 12, no. 1, p. 480, Dec. 2011, doi: 10.1186/1471-2105-12-480.

[46] W. Zheng, L. M. Chung, and H. Zhao, “Bias detection and correction in RNA-Sequencing data,” *BMC Bioinformatics*, vol. 12, no. 1, p. 290, Jul. 2011, doi: 10.1186/1471-2105-12-290.

[47] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[48] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, “voom: precision weights unlock linear model analysis tools for RNA-seq read counts,” *Genome Biology*, vol. 15, no. 2, p. R29, Feb. 2014, doi: 10.1186/gb-2014-15-2-r29.

[49] G. K. Smyth, “Linear models and empirical bayes methods for assessing differential expression in microarray experiments,” *Stat Appl Genet Mol Biol*, vol. 3, p. Article3, 2004, doi: 10.2202/1544-6115.1027.

[50] H. Han and K. Men, “How does normalization impact RNA-seq disease diagnosis?,” *Journal of Biomedical Informatics*, vol. 85, pp. 80–92, Sep. 2018, doi: 10.1016/j.jbi.2018.07.016.

[51] C. Soneson and M. Delorenzi, “A comparison of methods for differential expression analysis of RNA-seq data,” *BMC Bioinformatics*, vol. 14, p. 91, Mar. 2013, doi: 10.1186/1471-2105-14-91.

[52] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, “Differential expression in RNA-seq: A matter of depth,” *Genome Res.*, vol. 21, no. 12, pp. 2213–2223, Dec. 2011, doi: 10.1101/gr.124321.111.

[53] H. Jiang and W. H. Wong, “Statistical inferences for isoform expression in RNA-Seq,” *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, Apr. 2009, doi: 10.1093/bioinformatics/btp113.

[54] A. L. Oberg, B. M. Bot, D. E. Grill, G. A. Poland, and T. M. Therneau, “Technical and biological variance structure in mRNA-Seq data: life in the real world,” *BMC Genomics*, vol. 13, no. 1, p. 304, Jul. 2012, doi: 10.1186/1471-2164-13-304.

[55] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. 10, p. R106, Oct. 2010, doi: 10.1186/gb-2010-11-10-r106.

[56] Y.-H. Zhou, K. Xia, and F. A. Wright, “A powerful and flexible approach to the analysis of RNA sequence count data,” *Bioinformatics*, vol. 27, no. 19, pp. 2672–2678, Oct. 2011, doi: 10.1093/bioinformatics/btr449.

[57] R. Liu *et al.*, “Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses,” *Nucleic Acids Res.*, vol. 43, no. 15, p. e97, Sep. 2015, doi: 10.1093/nar/gkv412.

[58] L. L. Elo, S. Filen, R. Lahesmaa, and T. Aittokallio, “Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 423–431, Jul. 2008, doi: 10.1109/tcbb.2007.1078.

[59] F. Seyednasrollah, K. Rantanen, P. Jaakkola, and L. L. Elo, “ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer,” *Nucl. Acids Res.*, p. gkv806, Aug. 2015, doi: 10.1093/nar/gkv806.

[60] T. Suomi, F. Seyednasrollah, M. K. Jaakkola, T. Faux, and L. L. Elo, “ROTS: An R package for reproducibility-optimized statistical testing,” *PLOS Computational Biology*, vol. 13, no. 5, p. e1005562, May 2017, doi: 10.1371/journal.pcbi.1005562.

[61] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001, doi: 10.1073/pnas.091062498.

[62] M. A. Harris *et al.*, “The Gene Ontology (GO) database and informatics resource,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D258-261, Jan. 2004, doi: 10.1093/nar/gkh036.

[63] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat Protoc*, vol. 4, no. 1, pp. 44–57, 2009, doi: 10.1038/nprot.2008.211.

[64] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, “The KEGG resource for deciphering the genome,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D277-280, Jan. 2004, doi: 10.1093/nar/gkh063.

[65] X. Wang and M. J. Cairns, “SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing,” *Bioinformatics*, vol. 30, no. 12, pp. 1777–1779, Jun. 2014, doi: 10.1093/bioinformatics/btu090.

[66] Y. Rahmatallah, F. Emmert-Streib, and G. Glazko, "Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline," *Brief Bioinform*, vol. 17, no. 3, pp. 393–407, May 2016, doi: 10.1093/bib/bbv069.

[67] A. C. Frazee, B. Langmead, and J. T. Leek, "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinformatics*, vol. 12, p. 449, Nov. 2011, doi: 10.1186/1471-2105-12-449.

[68] I. Lappalainen *et al.*, "The European Genome-phenome Archive of human data consented for biomedical research," *Nature Genetics*, Jun. 26, 2015. https://www.nature.com/articles/ng.3312 (accessed Apr. 05, 2018).

[69] A. K. Green *et al.*, "The project data sphere initiative: accelerating cancer research by sharing data," *Oncologist*, vol. 20, no. 5, pp. 464-e20, May 2015, doi: 10.1634/theoncologist.2014-0431.

[70] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemp Oncol (Pozn)*, vol. 19, no. 1A, pp. A68–A77, 2015, doi: 10.5114/wo.2014.47136.

[71] T. Barrett *et al.*, "NCBI GEO: archive for functional genomics data sets--update," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D991-995, Jan. 2013, doi: 10.1093/nar/gks1193.

[72] N. Kolesnikov *et al.*, "ArrayExpress update--simplifying data submissions," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D1113-1116, Jan. 2015, doi: 10.1093/nar/gku1057.

[73] F. Seyednasrollah, A. Laiho, and L. L. Elo, "Comparison of software packages for detecting differential expression in RNA-seq studies," *Brief. Bioinformatics*, vol. 16, no. 1, pp. 59–70, Jan. 2015, doi: 10.1093/bib/bbt086.

[74] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, no. 1, p. 91, Mar. 2013, doi: 10.1186/1471-2105-14-91.

[75] V. M. Kvam, P. Liu, and Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data," *Am. J. Bot.*, vol. 99, no. 2, pp. 248–256, Feb. 2012, doi: 10.3732/ajb.1100340.

[76] D. Bottomly *et al.*, "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays," *PLoS ONE*, vol. 6, no. 3, p. e17820, Mar. 2011, doi: 10.1371/journal.pone.0017820.

[77] J. K. Pickrell *et al.*, "Understanding mechanisms underlying human gene expression variation with RNA sequencing," *Nature*, vol. 464, no. 7289, pp. 768–772, Apr. 2010, doi: 10.1038/nature08872.

[78] T. J. Hardcastle and K. A. Kelly, "baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, no. 1, p. 422, Aug. 2010, doi: 10.1186/1471-2105-11-422.

[79] J. Li and R. Tibshirani, "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data," *Stat Methods Med Res*, vol. 22, no. 5, pp. 519–536, Oct. 2013, doi: 10.1177/0962280211428386.

[80] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nat Biotech*, vol. 31, no. 1, pp. 46–53, Jan. 2013, doi: 10.1038/nbt.2450.

[81] N. Leng *et al.*, "EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics*, p. btt087, Feb. 2013, doi: 10.1093/bioinformatics/btt087.

[82] N. Delhomme, I. Padioleau, E. E. Furlong, and L. M. Steinmetz, "easyRNASeq: a bioconductor package for processing RNA-Seq data," *Bioinformatics*, vol. 28, no. 19, pp. 2532–2533, Oct. 2012, doi: 10.1093/bioinformatics/bts477.

[83] N. R. Hackett *et al.*, "RNA-Seq quantification of the human small airway epithelium transcriptome," *BMC Genomics*, vol. 13, p. 82, Feb. 2012, doi: 10.1186/1471-2164-13-82.

[84] Z. H. Zhang *et al.*, "A comparative study of techniques for differential expression analysis on RNA-Seq data," *PLoS ONE*, vol. 9, no. 8, p. e103207, 2014, doi: 10.1371/journal.pone.0103207.

[85] J. Zyprych-Walczak *et al.*, "The Impact of Normalization Methods on RNA-Seq Data Analysis," *Biomed Res Int*, vol. 2015, p. 621690, 2015, doi: 10.1155/2015/621690.

[86] Y. Sha, J. H. Phan, and M. D. Wang, "Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2015, pp. 6461–6464, 2015, doi: 10.1109/EMBC.2015.7319872.

[87] L. Jiang *et al.*, "Synthetic spike-in standards for RNA-seq experiments," *Genome Res*, vol. 21, no. 9, pp. 1543–1551, Sep. 2011, doi: 10.1101/gr.121095.111.

[88] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, no. 7456, pp. 43–49, Jul. 2013, doi: 10.1038/nature12222.

[89] Y. Sato *et al.*, "Integrated molecular analysis of clear-cell renal cell carcinoma," *Nat Genet*, vol. 45, no. 8, pp. 860–867, Aug. 2013, doi: 10.1038/ng.2699.

[90] S. Schrödter *et al.*, "Identification of the dopamine transporter SLC6A3 as a biomarker for patients with renal cell carcinoma," *Molecular Cancer*, vol. 15, no. 1, p. 10, Feb. 2016, doi: 10.1186/s12943-016-0495-5.

[91] J. Hansson *et al.*, "Overexpression of Functional SLC6A3 in Clear Cell Renal Cell Carcinoma," *Clin. Cancer Res.*, vol. 23, no. 8, pp. 2105–2115, 15 2017, doi: 10.1158/1078-0432.CCR-16-0496.

[92] J. C. van der Mijn, D. J. Panka, A. K. Geissler, H. M. Verheul, and J. W. Mier, "Novel drugs that target the metabolic reprogramming in renal cell cancer," *Cancer Metab*, vol. 4, p. 14, 2016, doi: 10.1186/s40170-016-0154-8.

[93] H. I. Wettersten, O. Abuaboud, P. N. Lara, and R. H. Weiss, "Metabolic reprogramming in clear cell renal cell carcinoma," *Nature reviews. Nephrology*, vol. 13, no. 7, pp. 410–419, Jul. 2017, doi: 10.1038/nrneph.2017.59.

[94] R. R. Raval *et al.*, "Contrasting properties of hypoxia-inducible factor 1 (HIF-1) and HIF-2 in von Hippel-Lindau-associated renal cell carcinoma," *Mol. Cell. Biol.*, vol. 25, no. 13, pp. 5675–5686, Jul. 2005, doi: 10.1128/MCB.25.13.5675-5686.2005.

[95] C. Shen *et al.*, "Genetic and functional studies implicate HIF1α as a 14q kidney cancer suppressor gene," *Cancer Discov*, vol. 1, no. 3, pp. 222–235, Aug. 2011, doi: 10.1158/2159-8290.CD-11-0098.

[96] L. Shi, S. An, Y. Liu, J. Liu, and F. Wang, "PCK1 Regulates Glycolysis and Tumor Progression in Clear Cell Renal Cell Carcinoma Through LDHA," *Onco Targets Ther*, vol. 13, pp. 2613–2627, Mar. 2020, doi: 10.2147/OTT.S241717.

[97] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020, doi: 10.3322/caac.21590.

[98] M. S. Litwin and H.-J. Tan, "The Diagnosis and Treatment of Prostate Cancer: A Review," *JAMA*, vol. 317, no. 24, pp. 2532–2542, Jun. 2017, doi: 10.1001/jama.2017.7248.

[99] M. Afshar, F. Evison, N. D. James, and P. Patel, “Shifting paradigms in the estimation of survival for castration-resistant prostate cancer: A tertiary academic center experience,” *Urol. Oncol.*, vol. 33, no. 8, p. 338.e1–7, Aug. 2015, doi: 10.1016/j.urolonc.2015.05.003.

[100] E. Powers *et al.*, “Novel therapies are changing treatment paradigms in metastatic prostate cancer,” *Journal of Hematology & Oncology*, vol. 13, no. 1, p. 144, Oct. 2020, doi: 10.1186/s13045-020-00978-z.

[101] A. J. Templeton *et al.*, “Translating clinical trials to clinical practice: outcomes of men with metastatic castration resistant prostate cancer treated with docetaxel and prednisone in and out of clinical trials,” *Ann. Oncol.*, vol. 24, no. 12, pp. 2972–2977, Dec. 2013, doi: 10.1093/annonc/mdt397.

[102] C.-W. Huang *et al.*, “A richly interactive exploratory data analysis and visualization tool using electronic medical records,” *BMC Med Inform Decis Mak*, vol. 15, p. 92, Nov. 2015, doi: 10.1186/s12911-015-0218-7.

[103] A. Banerjee, “Challenges for learning health systems in the NHS. Case study: electronic health records in cardiology,” *Future Hosp J*, vol. 4, no. 3, pp. 193–197, Oct. 2017, doi: 10.7861/futurehosp.4-3-193.

[104] J. M. Unger, E. Cook, E. Tai, and A. Bleyer, “The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies,” *American Society of Clinical Oncology Educational Book*, no. 36, pp. 185–198, May 2016, doi: 10.1200/EDBK_156686.

[105] J. W. Tukey, *Exploratory Data Analysis by John W. Tukey*. 1656.

[106] R. M. Church, “How to look at data: A review of John W. Tukey’s Exploratory Data Analysis,” *J Exp Anal Behav*, vol. 31, no. 3, pp. 433–440, May 1979, doi: 10.1901/jeab.1979.31-433.

[107] C. Chatfield, “Exploratory data analysis,” *European Journal of Operational Research*, vol. 23, no. 1, pp. 5–13, Jan. 1986, doi: 10.1016/0377-2217(86)90209-2.

[108] F. E. Grubbs, “Sample Criteria for Testing Outlying Observations,” *Ann. Math. Statist.*, vol. 21, no. 1, pp. 27–58, Mar. 1950, doi: 10.1214/aoms/1177729885.

[109] “(PDF) Comparison of outlier detection methods in biomedicaI data,” *ResearchGate*. https://www.researchgate.net/publication/234100794_Comparison_of_outlier_detection_methods_in_biomedicaI_data (accessed Jun. 02, 2020).

[110] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[111] P. H. Sneath, “The application of computers to taxonomy,” *J. Gen. Microbiol.*, vol. 17, no. 1, pp. 201–226, Aug. 1957, doi: 10.1099/00221287-17-1-201.

[112] S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” in *2014 Science and Information Conference*, Aug. 2014, pp. 372–378. doi: 10.1109/SAI.2014.6918213.

[113] G. Boente, A. M. Pires, and I. M. Rodrigues, “Influence Functions and Outlier Detection under the Common Principal Components Model: A Robust Approach,” *Biometrika*, vol. 89, no. 4, pp. 861–875, 2002.

[114] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br J Cancer*, vol. 89, no. 2, Art. no. 2, Jul. 2003, doi: 10.1038/sj.bjc.6601118.

[115] V. Bewick, L. Cheek, and J. Ball, “Statistics review 12: Survival analysis,” *Crit Care*, vol. 8, no. 5, pp. 389–394, 2004, doi: 10.1186/cc2955.

[116] D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, Jan. 1972.

[117] J. Fan and R. Li, “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001, doi: 10.1198/016214501753382273.

[118] R. Tibshirani, “The Lasso Method for Variable Selection in the Cox Model,” *Statist. Med.*, vol. 16, no. 4, pp. 385–395, Feb. 1997, doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

[119] H. H. Zhang and W. Lu, “Adaptive Lasso for Cox’s proportional hazards model,” *Biometrika*, vol. 94, no. 3, pp. 691–703, Aug. 2007, doi: 10.1093/biomet/asm037.

[120] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J Royal Statistical Soc B*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[121] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[122] T. Hastie, R. Tibshirani, and J. Friedman, “Boosting and Additive Trees,” in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, T. Hastie, R. Tibshirani, and J. Friedman, Eds. New York, NY: Springer, 2009, pp. 337–387. doi: 10.1007/978-0-387-84858-7_10.

[123] T. G. Dietterich, “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization,” *Machine Learning*, vol. 40, no. 2, pp. 139–157, Aug. 2000, doi: 10.1023/A:1007607513941.

[124] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[125] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, 1 edition. Boca Raton: Chapman and Hall/CRC, 1984.

[126] L. Abarca-Gómez *et al.*, “Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults,” *The Lancet*, vol. 390, no. 10113, pp. 2627–2642, Dec. 2017, doi: 10.1016/S0140-6736(17)32129-3.

[127] S. E. Kahn, R. L. Hull, and K. M. Utzschneider, “Mechanisms linking obesity to insulin resistance and type 2 diabetes,” *Nature*, vol. 444, no. 7121, pp. 840–846, Dec. 2006, doi: 10.1038/nature05482.

[128] L. F. Van Gaal, I. L. Mertens, and C. E. De Block, “Mechanisms linking obesity with cardiovascular disease,” *Nature*, vol. 444, no. 7121, pp. 875–880, Dec. 2006, doi: 10.1038/nature05487.

[129] A. G. Renehan, M. Zwahlen, and M. Egger, “Adiposity and cancer risk: new mechanistic insights from epidemiology,” *Nat Rev Cancer*, vol. 15, no. 8, pp. 484–498, Aug. 2015, doi: 10.1038/nrc3967.

[130] M. Juonala, M. Raitakari, J. S. A. Viikari, and O. T. Raitakari, “Obesity in youth is not an independent predictor of carotid IMT in adulthood: The Cardiovascular Risk in Young Finns Study,” *Atherosclerosis*, vol. 185, no. 2, pp. 388–393, Apr. 2006, doi: 10.1016/j.atherosclerosis.2005.06.016.

[131] M. O. Goodarzi, “Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications,” *The Lancet Diabetes & Endocrinology*, vol. 6, no. 3, pp. 223–236, Mar. 2018, doi: 10.1016/S2213-8587(17)30200-0.

[132] M. Juonala, J. S. A. Viikari, and O. T. Raitakari, "Main findings from the prospective Cardiovascular Risk in Young Finns Study," *Curr. Opin. Lipidol.*, vol. 24, no. 1, pp. 57–64, Feb. 2013, doi: 10.1097/MOL.0b013e32835a7ed4.

[133] M. Juonala *et al.*, "Childhood environmental and genetic predictors of adulthood obesity: the cardiovascular risk in young Finns study," *J. Clin. Endocrinol. Metab.*, vol. 96, no. 9, pp. E1542-1549, Sep. 2011, doi: 10.1210/jc.2011-1243.

[134] A. E. Locke *et al.*, "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, pp. 197–206, Feb. 2015, doi: 10.1038/nature14177.

[135] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *Trends Genet.*, vol. 17, no. 9, pp. 502–510, Sep. 2001, doi: 10.1016/s0168-9525(01)02410-6.

[136] E. K. Speliotes *et al.*, "Association analyses of 249,796 individuals reveal eighteen new loci associated with body mass index," *Nat Genet*, vol. 42, no. 11, pp. 937–948, Nov. 2010, doi: 10.1038/ng.686.

[137] S. E. Hodge and D. A. Greenberg, "How Can We Explain Very Low Odds Ratios in GWAS? I. Polygenic Models," *Hum. Hered.*, vol. 81, no. 4, pp. 173–180, 2016, doi: 10.1159/000454804.

[138] Y. Yasui, "Why odds ratio estimates of GWAS are almost always close to 1.0," *COBRA Preprint Series*, May 2012, [Online]. Available: https://biostats.bepress.com/cobra/art94

[139] S. Gupta *et al.*, "Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry," *BMJ Open*, vol. 4, no. 3, p. e004007, Mar. 2014, doi: 10.1136/bmjopen-2013-004007.

[140] H. Behravan, J. M. Hartikainen, M. Tengström, V.-M. Kosma, and A. Mannermaa, "Predicting breast cancer risk using interacting genetic and demographic factors and machine learning," *Scientific Reports*, vol. 10, no. 1, Art. no. 1, Jul. 2020, doi: 10.1038/s41598-020-66907-9.

[141] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," *Bioinformatics*, vol. 23, no. 1, pp. 30–37, Jan. 2007, doi: 10.1093/bioinformatics/btl543.

[142] H. J. Kim, H. J. Kim, Y. Park, W. S. Lee, Y. Lim, and J. H. Kim, "Clinical Genome Data Model (cGDM) provides Interactive Clinical Decision Support for Precision Medicine," *Scientific Reports*, vol. 10, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41598-020-58088-2.

[143] M. E. Futschik, M. Sullivan, A. Reeve, and N. Kasabov, "Prediction of clinical behaviour and treatment for cancers," *Appl Bioinformatics*, vol. 2, no. 3 Suppl, pp. S53-58, 2003.

[144] D. Sun, A. Li, B. Tang, and M. Wang, "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 45–53, Jul. 2018, doi: 10.1016/j.cmpb.2018.04.008.

[145] Y. Y. Teo *et al.*, "A genotype calling algorithm for the Illumina BeadArray platform," *Bioinformatics*, vol. 23, no. 20, pp. 2741–2746, Oct. 2007, doi: 10.1093/bioinformatics/btm443.

[146] "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000529 (accessed Jun. 03, 2020).

[147] R. M. Durbin *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, Art. no. 7319, Oct. 2010, doi: 10.1038/nature09534.

[148] S. Steinsbekk, D. Belsky, I. C. Guzey, J. Wardle, and L. Wichstrøm, "Polygenic Risk, Appetite Traits, and Weight Gain in Middle Childhood," *JAMA Pediatr*, vol. 170, no. 2, p. e154472, Feb. 2016, doi: 10.1001/jamapediatrics.2015.4472.

[149] C. E. Elks *et al.*, “Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth,” *PLoS Med.*, vol. 7, no. 5, p. e1000284, May 2010, doi: 10.1371/journal.pmed.1000284.

[150] M. Hakanen *et al.*, “FTO genotype is associated with body mass index after the age of seven years but not with energy intake or leisure-time physical activity,” *J. Clin. Endocrinol. Metab.*, vol. 94, no. 4, pp. 1281–1287, Apr. 2009, doi: 10.1210/jc.2008-1199.

[151] E. López de Maturana *et al.*, “Challenges in the Integration of Omics and Non-Omics Data,” *Genes (Basel)*, vol. 10, no. 3, Mar. 2019, doi: 10.3390/genes10030238.

[152] A. Althubaiti, “Information bias in health research: definition, pitfalls, and adjustment methods,” *J Multidiscip Healthc*, vol. 9, pp. 211–217, May 2016, doi: 10.2147/JMDH.S104807.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama Oy, Turku, Finland 2022

ISBN 978-951-29-8759-7 (PRINT)
ISBN 978-951-29-8760-3 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)