Department of Computer Science

Computational Analysis and Modeling of High-Throughput Data to Understand T-helper Cell Differentiation

Kari Nousiainen

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 197/2022

Computational Analysis and Modeling of High-Throughput Data to Understand T-helper Cell Differentiation

Kari Nousiainen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall TU1 Saab Auditorium of the school on 13 January 2023 at 12.00.

Aalto University
School of Science
Department of Computer Science
Computational Systems Biology Group

Supervising professor
Professor Harri Lähdesmäki, Aalto University School of Science, Finland

Preliminary examiners
Professor Tero Aittokallio, Institute for Molecular Medicine Finland, Finland and University of Oslo, Norway
Professor Dario Greco, University of Tampere, Finland

Opponent
Professor Tom Michoel, University of Bergen, Norway

Aalto University publication series
DOCTORAL THESES 197/2022

ISBN 978-952-64-1078-4 (printed)
ISBN 978-952-64-1079-1 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-1079-1

Unigrafia Oy
Helsinki 2022

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Author
Kari Nousiainen

Name of the doctoral thesis
Computational Analysis and Modeling of High-Throughput Data to Understand T-helper Cell Differentiation

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL THESES 197/2022

Field of research Computer and Information Science

Manuscript submitted 12 April 2022 **Date of the defence** 13 January 2023

Permission for public defence granted (date) 17 August 2022 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

T-helper cells are essential for adaptive immunity. During immune response T-helper cells are influenced by cytokines which steer cell differentiation into cellular subsets having specific functions. The process is affected by cellular subsystems and causes profound changes in epigenetic modifications as well as gene and protein expressions which can be experimentally observed using high-throughput technologies. This thesis has three objectives. 1) To identify and characterize molecular elements involved in T-helper cell differentiation and immune response through analyzing datasets using bioinformatic tools. 2) To develop computational tool to detect enrichments of trait associated single nucleotide polymorphisms (SNPs) on genomic regions. 3) To develop computational frameworks for characterizing dynamic models for regulatory networks. The goals have been achieved in five studies.

In the first study, proteomes and transcriptomes of Th17 and iTreg cells were profiled and analyzed to understand how they change during early phases of cell differentiation and how the transcriptomes and proteomes of the cell types differ from each other.

The second publication characterized bindings of transcription factor (TF) STAT3 genome-wide during Th17 cell differentiation. As a SNP can alter binding affinity of a TF, we investigated whether SNPs associated with immune diseases co-localize in STAT3-binding sites. The analysis applied publicly available information on SNPs and empirical statistical methods.

The third publication proposes a computational tool snpEnrichR implemented in R language for facilitating co-localization analyses of SNPs and genomic regions. Co-localization analysis of SNPs associated to various traits and STAT6 binding sites of cells differentiating toward Th2 type showed that incorporating proxies of the tag-SNPs enhances co-localization detection.

The fourth publication introduced a method to infer dynamically evolving regulatory networks from time-course data. The method couples mechanistic ordinary differential equation (ODE) models with a latent process that approximates the network structure rewiring process. When applied to Th17 RNA-seq data the method predicted lineage specific subnetworks that are activated sequentially and control the differentiation process in an overlapping manner.

The fifth publication studies the dynamic interplay of histone modifications signaling enhancer activity and transcription factor binding modeling them using systems of ODEs and simulated time-course data focusing on the parameters of the models and model selection. The method is able to find the correct model when measurement noise level is reasonable and the number of measurement time points is adequate.

The datasets generated and the analyses performed as part of this thesis help to understand of T-helper cell differentiation better. The developed computational frameworks and tools available

Keywords T-helper cells, bioinformatics, SNP, dynamical systems

ISBN (printed) 978-952-64-1078-4 **ISBN (pdf)** 978-952-64-1079-1

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2022

Pages 150 **urn** http://urn.fi/URN:ISBN:978-952-64-1079-1

Tekijä
Kari Nousiainen

Väitöskirjan nimi
Computational Analysis and Modeling of High-Throughput Data to Understand T-helper Cell Differentiation.

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL THESES 197/2022

Tutkimusala Informaatiotekniikka

Käsikirjoituksen pvm 12.04.2022 **Väitöspäivä** 13.01.2023

Väittelyluvan myöntämispäivä 17.08.2022 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

T-auttajasolut ovat tärkeitä hankitulle immuniteetille. Immuunivasteen aikana T-auttajasolut erilaistuvat sytokiinien vaikutuksesta alasolutyypeiksi, joilla on kullakin omat tehtävänsä. Prosessissa solunsisäiset järjestelmät muuttavat solun epigenetiikkaa sekä proteiinien ja geenien ilmentymiä. Solun molekyyleistä voidaan saada havaintoja suurikapasiteettimittaustekniikoilla. Tässä väitöskirjassa on kolme tavoitetta: 1) Tunnistaa ja luonnehtia erilaistuvien T-solujen molekyylien toimintaa analysoimalla aineistoja bioinformatiikan menetelmin. 2) Kehittää laskennallinen menetelmä, joka tunnistaa SNPien rikastumat annetuilla genomialueilla. 3) Kehittää laskennallisia menetelmiä solun dynaamisten säätelyjärjestelmien mallintamiseen. Väitöskirjan tavoitteet saavutettiin viidellä tutkimuksella.

Ensimmäisessä tutkimuksessa Th17 ja iTreg solujen transkriptomit ja proteomit profiloitiin ja analysoitiin biometriikan menetelmillä. Työssä selvitettiin solujen muutoksia erilaistumisen alkuvaiheessa ja solutyyppien geeni-ilmentymien eroja transkriptomi- ja proteomitasoilla.

Toisessa tutkimuksessa selvitettiin koko genomin kattavasti transkriptiotekijä STAT3:n sitoutumista genomiin soluilla, jotka ovat erilaistumassa Th17 soluiksi. Sitoutumispaikkojen yhden emäksen monimuotoisuudet (SNP) vaikuttavat sitoutumusaffiniteettiin, minkä vuoksi tutkittiin, ovatko immuunisairauksiin liitetyt SNPit rikastuneet STAT3:n sitoutumispaikoissa. Analyysissa hyödynnettiin yleisesti SNPeistä saatavilla olevaa tietoa sekä empiirisiä tilastollisia menetelmiä.

Kolmannessa työssä esiteltiin R-ohjelmointikielellä toteutettu laskennallinen työkalu snpEnrichR helpottamaan eri ominaisuuksiin liittyvien SNPien genomialueilla rikastumisen analysointia. STAT6:n sitoutumispaikkojen sekä eri sairauksiin liittyvien SNPien tutkiminen osoitti, että edustaja-SNPien hyödyntäminen tunniste-SNPien lisäksi auttaa tunnistamaan ko-lokalisaatiota.

Neljännessä työssä kehitettiin menetelmä, jolla päätellään dynaamisesti kehittyvä säätelyverkko aikasarja-aineistosta. Menetelmä yhdistää mekanistisen tavallisen differentiaaliyhtälöryhmämallin piiloprosessiin, joka puolestaan approksimoi mallin rakenteen uudelleen kytkentöjä. Sovellettaessa Th17 RNA-seq dataan menetelmä ennusti solutyypille ominaiset alaverkot, jotka aktivoituvat peräkkäin ja kontrolloivat erilaistumisprosessia päällekkäin.

Viidennessä työssä tutkittiin enhansereiden aktiivisuutta ilmaisevien histonimuutosten transkriptiotekijän genomiin sitoutumisen vuorovaikutusta mallintamalla niitä tavallisilla differentiaaliyhtälöillä (DY) hyödyntämällä simuloitua aikasarja dataa keskittyen mallin parametreihin ja mallinvalintaan. Menetelmä kykenee löytämään datan generoivan mallin, kun aineiston kohina on maltillista ja datapisteitä on riittävä määrä.

Tämän väitöskirjan sisältämät analyysit auttavat ymmärtämään paremmin T-solujen erilaistumista. Kehitetyt laskennalliset viitekehykset ja työkalut ovat vapaasti käytettävissä.

Avainsanat T-auttajasolu, bioinformatiikka, SNP, dynaamiset järjestelmät

ISBN (painettu) 978-952-64-1078-4 **ISBN (pdf)** 978-952-64-1079-1

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2022

Sivumäärä 150 **urn** http://urn.fi/URN:ISBN:978-952-64-1079-1

Preface

I have done this work in the Computational Systems Biology Group at the Department of Computer Science at the Aalto University School of Science. I am deeply grateful for all people in the group and especially to Dr. Jukka Intosalmi, Dr. Kartiek Kanduri and my supervisor Prof. Harri Lähdesmäki for kind support. Furthermore, great part of this work has been done in close collaboration with biologists and medical people. I am deeply thankful to Dr. Jane Chen, Dr. Imran Mohammad, Dr. Ricaño-Ponce and Prof. Riitta Lahesmaa as well as other co-workers who have supported me with valuable data sets and biological insights. Finally, I thank everybody else who have supported this work.

Espoo, Finland, November 28, 2022,

Kari Nousiainen

Contents

List of Publications

This thesis consists of an overview of the following publications which are referred to in the text by their Roman numerals.

I Imran Mohammad, Kari Nousiainen, Santosh D. Bhosale, Inna Starskaia, Robert Moulder, Anne Rokka, Fang Cheng, Ponnuswamy Mohanasundaram, John E. Eriksson, David R.Goodlett, Harri Lähdesmäki, Zhi Chen. Quantitative proteomic characterization and comparison of T helper 17 and induced regulatory T cells. *PloS Biology*, 16.5 (2018): e2004194, May 2018.

II Subhash K Tripathi, Zhi Chen, Antti Larjo, Kartiek Kanduri, Kari Nousiainen, Tarmo Äijo, Isis Ricaño-Ponce, Barbara Hrdlickova, Soile Tuomela, Essi Laajala, Verna Salo, Vinod Kumar, Cisca Wijmenga, Harri Lähdesmäki, Riitta Lahesmaa. Genome-wide Analysis of STAT3-Mediated Transcription during Early Human Th17 Cell Differentiation. *Cell Reports*, 19.9 (2017): 1888-1901, May 2017.

III Kari Nousiainen, Kartiek Kanduri, Isis Ricaño-Ponce, Cisca Wijmenga, Riitta Lahesmaa, Vinod Kumar, Harri Lähdesmäki. snpEnrichR: analyzing co-localization of SNPs and their proxies in genomic regions. *Bioinformatics*, 34.23 (2018): 4112-4114, December 2018.

IV Jukka Intosalmi, Kari Nousiainen, Helena Ahlfors, Harri Lähdesmäki. Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks. *Bioinformatics*, 32.12 (2016): i288-i296, June 2016.

V Kari Nousiainen, Jukka Intosalmi, Harri Lähdesmäki. A Mathematical Model for Enhancer Activation Kinetics During Cell Differentiation. *In International Conference on Algorithms for Computational Biology*, pp. 191-202, May 2019.

Author's Contribution

Publication I: "Quantitative proteomic characterization and comparison of T helper 17 and induced regulatory T cells"

Nousiainen contributed to the data curation and management. He designed, implemented, and ran the computational pipeline in order to perform RNA-seq and proteomics data analyses. He participated in designing and implementing the visualization. Nousiainen wrote methods parts of the manuscripts and contributed to writing and revising the whole manuscript.

The following list contains the contributions of other authors. 1) Proteomic and transcriptomic experiments and methodology: I.M., S.B.D., I.S., A.R. and J.E.E. 2) Data analysis and curation: H.L. and C.Z. 3) Biological validation and further experiments: I.M., I.S., F.C., P.M., J.E.E. and D.R.G. 4) Expertise: H.L., R.M., J.E.E. and D.R.G. 5) Supervision: H.L., R.M., J.E.E., D.R.G. and Z.C. 6) Funding and Resources: H.L., Z.C., and J.E.E. 7) Original draft and conceptualization: Z.C. 8) Leadership: Z.C. 9) Writing and reviewing of the manuscript: I.M., S.B.D., I.S., R.M., A.R., F.C., P.M., J.E.E., D.R.G., H.L. and Z.C.

Publication II: "Genome-wide Analysis of STAT3-Mediated Transcription during Early Human Th17 Cell Differentiation"

Nousiainen had the main role in developing and implementing the SNP enrichment analysis method. He wrote parts of the methods section and took part to writing and revising the whole manuscript.

S.K.T. designed, performed the experiments, analyzed data, prepared figures, and wrote the manuscript. Z.C. designed and performed the experiments, provided expertise, analyzed data, prepared figures, and wrote the manuscript. A.L., K.K., K.N., and T.Ä. analyzed data, prepared figures, and wrote parts of the manuscript. I.R.-P. analyzed data for the SNP study. B.H. contributed to SNP analysis. S.T. provided expertise. E.L. performed eQTL analysis. V.S.

performed cell cultures. C.W. and V.K. provided expertise and guidance. H.L. provided expertise and guidance and supervised A.L., K.K., K.N., T.Ä., and E.L. R.L. designed and supervised the study and wrote the manuscript.

Publication III: "snpEnrichR: analyzing co-localization of SNPs and their proxies in genomic regions"

Nousiainen participated in the development of the software package for the computational method initially proposed in Publication II. He implemented the R package with Kartiek Kanduri. He carried out all computational and data analyses as well as interpreted the results with other co-authors. R.L. provided the biological question, data, interpretation of the results and the need for implementing an R package to address the question. She co-supervised K.K. C.W., V.K., and I.R. were involved in the optimization process of the algorithm in defining LD structure, proxy SNP selection, and colocalization. They participated in interpreting the results. Nousiainen wrote the manuscript together with other co-authors. H.L. supervised the study.

Publication IV: "Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks"

Nousiainen participated in computational method development and performed the parameter identifiability analysis. Nousiainen contributed to writing and revising the manuscript. H.A. provided expertise. J.I. is the first author and he developed the method with H.L. J.I. run most of the experiments. J.I. wrote the manuscript together with other co-authors.

Publication V: "A Mathematical Model for Enhancer Activation Kinetics During Cell Differentiation"

Nousiainen created the methodology together with Intosalmi and Lähdesmäki. Nousiainen implemented the framework computationally and carried out all computational simulations and statistical analyses. Nousiainen performed the visualizations and he wrote the paper together with other co-authors.

Abbreviations

A adenine

APC antigen presenting cell

APL adjusted profile likelihood

BCR B cell receptor

BDF backward differentiation formula

bp base pair

C cytosine

CD4 cluster of differentation 4

CD8 cluster of differentation 8

cDNA complementary DNA

CRE cis-regulatory element

DGE differential gene expression

DNA deoxyribonucleic acid

EDF empirical distribution function

ES enrichment score

FDR false discovery rate

FSA forward sensitivity analysis

FWER family-wise error rate

GSEA gene set enrichment analysis

GLM generalized linear model

GWAS genome–wide association studies

G guanine

iTreg inducible T-regulatory

LC-MS mass spectrometry coupled with liquid chromatography

LD linkage disequilibrium

LFQ label-free quantification

MAF minor allele frequence

MHC major histocompatibility complex

mRNA messanger RNA

MS mass spectrometry

MLE maximum likelihood estimate

NGS next-generation sequencing

NB negative binomial

ODE ordinary differential equation

pmf probability mass function

qPCR quantitative polymerase chain reaction

pre-mRNA precursor mRNA

RNA ribonucleic acid

RNA-seq RNA sequencing

SNP single nucleotide polymorphism

SNV single nucleotide variant

STAT3 signal transducer and activator of transcription 3

T thymine

TCR T cell receptor

TF transcription factor

Th T helper

Th17 T helper type 17

TSS transcription start site

U uracil

1. Introduction

The smallest biological, functional and structural unit of a living organism is a cell. A cell is alive when it is able to grow, metabolize, respond to stimuli, adapt and reproduce. Genetic information needed to accomplish the activities characterizing life is encoded in DNA (deoxyribonucleic acid) sequence from which it flows through RNA (ribonucleic acid) to proteins that are responsible for many of cellular functions. The biomolecules interact with each other through sophisticated chemical pathways that control the expression of the molecules and activities they perform. Many branches of biological or life sciences have been devoted to study biomolecules either one by one or at system-wide level. During a couple of previous decades, technological advances have enabled more and more large scale studies in the field by lowering the costs and speeding up the experimental processes. Mass-spectometry, microarray and next generation sequencing (NGS) (successor to the Sanger sequencing) based studies produce huge amounts of high-throughput data from proteins, DNA, RNA and other biomolecules for various applications which can be used to gain deep knowledge of cells. Often such kind of data is called with names ending with -omics suffix like genomics, transcriptomics and proteomics.

Prerequirement for applying microarray or next generation sequencing based experimental procedures is completely sequenced genome (or transcriptome). For example, publishing of the initial sequences of human, important animal models and other genomes (Venter et al., 2001; Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002; Huttenhower et al., 2012) has been vital for biomedical studies by providing reference sequences. Microarray and NGS both can be used to, for example, measure gene expression, identify copy number polymorphism, genotype single nucleotide polymorphisms (SNPs), and capture sequences of interest for downstream analysis. Emperimental efficacy has enabled immense efforts like 1000 genome project or ENCODE that aimed to list genetic variation between individuals and functional elements in the human genome, respectively (Birney and Soranzo, 2015; ENCODE Project Consortium et al., 2004). Such projects as well as many databases provide useful data sources for further use. The other high-throughput technology, mass spectometry (MS), is used in proteomics and metabolomics (Aebersold and Mann,

2003; Clish, 2015). MS has been applied successfully for many purposes, such as identifying peptides, proteins, and post-translational modifications, measuring protein amounts, characterizing the structure of the protein, and identifying protein interactions with proteins or nucleic acids (Sidoli et al., 2016). Despite the potential of mass spectrometry, it has not been applied as widely as NGS perhaps due to the fact that both biological and statistical methods are still maturing (Sidoli et al., 2016).

As high-throughput technologies have become standard tools in life sciences, data processing and storage must be updated to meet the requirements of the experimental data which is noisy, have high dimension and is often derived from relatively small sample size (Muir et al., 2016). Bioinformatics and computational biology combine computer science, mathematics and statistics with biology in order to translate data into biological insight. The field is rapidly evolving directed by the need for novel methods and software as well as by creative application of existing tools. Because meaningful computational accomplishments require both high-quality experimental data and specialists' ability to evaluate new biological findings, the work is done in close collaboration with experimental biologists.

The biological context of this dissertation is cell differentiation. More specifically, the biological questions relate to adaptive cell-mediated immunity and T cell development. Cell differentiation refers to process in which progenitor cells, that are able to develop into different kinds of specialized cells, develop into some specific cell type steered by external stimuli. During differentiation cells manifest changes in various subcellular systems which can be screened with high-throughput methods. This thesis has a dual point view into the subject. On one hand, the studies involve a lot of data analysis and bioinformatics. Such work consists of a variety of tasks such as quality control of the data, hypothesis testing or model selection and calibration, pathway enrichment, functional analysis, and data visualization. Each of them requires the choosing appropriate methodology and tools accomplishing the task. On the other hand, this thesis introduces novel methodology and computational tools. This part of the work focuses on detecting significant co-localization of genomic regions and genetic variation as well as on the mechanistic modeling of subcellular systems.

This thesis consists of five peer-reviewed journal articles, which are parceled out in three themes. The first theme concerns the protein expression profiling of Th17 and inducible T–regulatory (iTreg) cells. It catalogues proteomic and transcriptional changes between the cell types and the changes that occur during respective cell differentiation. The functional analysis of the findings is naturally incorporated. The work covered in Publication I. The second theme concerns deciphering the importance of genomic regions, such as transcription factor (TF) binding sites, through genetic variation represented by single nucleotide polymorphisms (SNP). Publication II shows that STAT3 regulates human Th17 cell differentiation and reveals that SNPs associated with disorders in immunity affect the binding affinity of STAT3 to DNA. As a part of the study, we utilized

co-localization analysis of SNPs and TF binding sites. It provided insight that otherwise hidden co-localizations may be revealed when the tag-SNPs are augmented with other SNPs in their linkage disequilibrium (LD). The method, the practical toolkit for such analysis and an illuminating test case is documented in Publication III. The third theme focuses on finding suitable mechanistic models for subcellular systems when only a limited amount of data and a little or conflicting prior knowledge of the system is available. Publication IV covers modeling regulatory systems that change when time passes by due to the actions of latent molecules which is a likely scenario in differentiating cells. The study represents a computational framework for model selection and parameter tuning for such a model. The framework found the correct model for a simulated setting and a promising model for the gene regulatory network steering Th17 cell differentiation using time series RNA-seq data. Publication V concerns the enhancer activation regulating cell differentiation. The work aims to discover what is required from the data in order to make reliable model selection.

The overall aim of this dissertation is to develop methods for multi-omics data analysis, data integration and system modelling by applying diverse computational approaches ranging from statistical analysis to ordinary differential equations (ODEs). This work introduces applicable methods and implements them as usable software. The objective of these contributions is to facilitate further experimental research in immunology as well as other areas of life sciences.

The chapters of this thesis are organized followingly. To begin with, I briefly introduce the biological concepts relevant to this thesis in Chapter 2. Then, I continue with describing the main statistical, computational and mathematical methods that have been applied in this dissertation in Chapter 3. The main results of the publications are discussed in Chapter 4 to Chapter 6. Finally, Chapter 7 contains the discussion and the main conclusions of this dissertation.

2. Biological background

A multicellular organism can be functionally complex incorporating various specialized systems, tissues and cell types. For example, Bianconi et al. (2013) have estimated that an adult human body consists of 37.2 trillion cells which can be divided into roughly 200 different cells types (DeSanctis and Loreti, 2017). Despite this complexity, all cells originate from stem cells or already slightly specialized progenitor cells that are able to develop into some mature cell type (Seaberg and van der Kooy, 2003). This thesis considers mainly cell differentiation in the context certain human and mouse cells involved in adaptive immunity which is part of the immune system. Cell differentiation is a profound process thoroughly affecting the molecular composition and cellular subsystems of the cells. This dissertation traces and models them and additionally considers the possible impacts of heritable differences in DNA to cells. The main classes of molecules are DNA, RNA and protein. This chapter considers them from the global point of view representing them as genome, transcriptome and proteome.

2.1 Genome

Nucleotides are essential monomers for encoding hereditary information. Nucleotides participating this purpose are composed of three components: 1) a deoxiribose monosaccharide, 2) a phosphate group and 3) one of the nitrogeous nucleobases adenine (A), cytosine (C), guanine (G) or thymine (T). A nucleotide forms one chemical bond with the sugar and another with the adjoining nucleotide. These chemical bindings create larger molecules characterized by the resulting nucleotide sequences. DNA consists of two nucleotide sequence in reversed order forming double helix structure (Watson and Crick, 1953) as shown in Figure 2.1. The nucleotide sequences are known as the strands of DNA. Hydrogen bonds between the nucleotides in the complementary strands connects the strands together. A nucleotides bond to T in the opposite strand as well as C to G. The bonded nucleotides form base pairs used as the basic unit measuring the length of DNA sequences. DNA constitutes the complete genetic material of an organism known as genome. The definition enables the molecular

study of genes despite apparently constant speculation about the quiddity of a gene (Pearson, 2006; Gerstein et al., 2007; Portin and Wilkins, 2017).

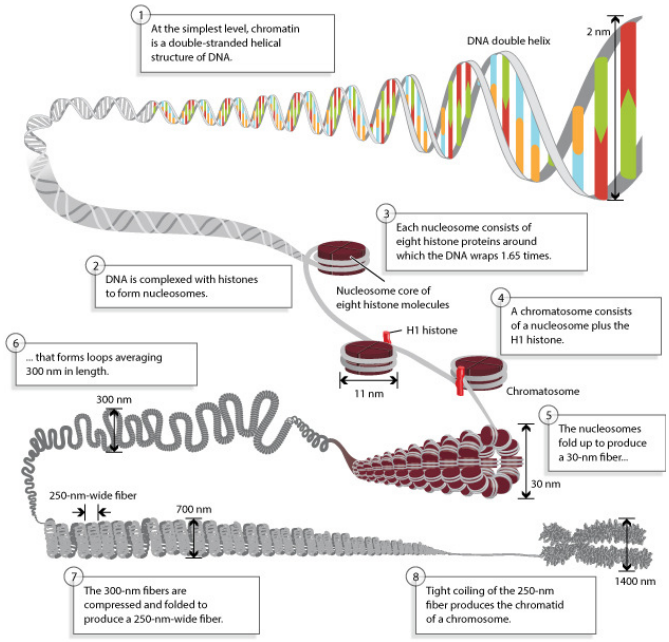

Figure 2.1. Well-known structure of DNA is nucleotide pairing in double helix form. The chromatin structure formed with histones allows very compact packaging for DNA when needed. Usually, DNA structure is uncoiled enabling gene transcription. The view reflects central dogma of molecular biology which describes information flows through the cells via stages as DNA replication, transcription and translation(Crick, 1958, 1970). This work studies dynamics of the information flow focusing on the actions of the following cellular systems 1) single-nucleotide variations, 2) chromatin modifications, 3) changes in transcriptomes and proteomes as well as 4) biological pathways and processes in T cell differentiation. Figure from Annunziato (2008)).

2.1.1 Chromatin and histone modifications

In cell nuclei, DNA molecules are packed compactly in chromosomes with the help of chromosomal proteins. The structure consisting of chromosomal proteins and DNA molecules is called chromatin as illustrated in Figure 2.1. Chromatin has many functions. For example, it regulates gene expression, is involved in mitosis and protects DNA against damage (Olins and Olins, 2003). Depending on the stage of the cell cycle, chromatin can be more or less densely packed. Sparsily packed chromatin is called heterochromatin and densely packed euchromatin (Gaspar-Maia et al., 2011). The nucleosome is the basic repeating subunit of chromatin. It is formed by an octamer which consists of two copies of each histone proteins H2A, H2B, H3 and H4 or some noncanonical variant of H2A and a stretch of double stranded DNA (Bönisch and Hake, 2012). The adjacent nucleosome particles are separated by linker DNA. Linker histones H1 and H5 located outside of nucleosome participate in condensing chromatin fiber (Gilbert

et al., 2005).

Each of the core histones consists of a long N-terminal amino acid tail and a histone fold regions. In histone octamers, the fold regions are tightly connected to each other while the tails extend out of the DNA-histone core. The histone tails are subject to many kinds of covalent modifications that control chromatin structure (Alberts et al., 2002). Modifications are abundant. There are over 60 different sites on histone tails where modifications have been detected (Kouzarides, 2007). By convention, post-translational histone modifications are named by abbreviation starting with the affected histone, continuing with the specification of the amino acid denoted by its abbreviation and location number and ending with the type of modification and its count. For example, H3K4me3 stands for trimethylation of fourth lysine residue from the beginning of N-terminal of histone H3 (Turner, 2005). Histone modifications affect accessability of DNA and thus cellular processes such as transcription as well as DNA replication and repair (Zentner and Henikoff, 2013). This thesis considers only certain post-translational histone modifications. Acetylation in general and methylation of lysine 4 have been associated with active transcription (Strahl and Allis, 2000).

2.1.2 Genetic variation

Genetic variation is the difference in DNA sequences within a species. A common source of the differences is point mutations which are permanent changes in DNA sequence mapping to a single base pairs. This thesis focuses on those. New point mutations are caused by errors in DNA replication or DNA damages due to exposure to specific chemicals or radiation that are not corrected by DNA repair enzymes (Lodish et al., 2000). Once the error is copied and fixed in the DNA, it is considered as a mutation. Mutations can occur in both germ cells and other cells which are called somatic cells. When mutations occur in germ cells, they may become heritable and can over the time affect subsequent populations (Griffiths et al., 2000).

A variant is a specific region of the genome which differs between two genomes. Different versions of the same variant are called alleles. If there are different alleles at a single position in DNA sequence, it is called as single nucleotide variant (SNV). If it is sufficiently frequent within a population, it is called single nucleotide polymorphism (SNP) (Brookes, 1999; Karki et al., 2015).

Genetic variations are often studied at a genome scale. In such case, samples are compared with a reference genome that consists of reference alleles. A base that does not match the reference allele at any given locus is caller alternative allele. There may be several alternative variants per variant. The most common allele in a given population is called major allele and the second most common is minor allele. The associated (relative) allele frequencies are usually represented fractions or percentages. Especially, minor allele frequency (MAF) is a commonly represented measure for alternative alleles.

Sexual reproduction involves a process called meiosis. It is a specialized type of cell division that cuts the number of chromosomes half, creating four haploid cells distinctive from the original cell. Generating a recombinant gamete after a round of meiosis leads to the concept of linkage disequilibrium (LD). It measures the association of alleles at loci in the same chromosome that are not random. Let A and B be alleles at two loci nearby each other, AB be their combination and p_A, p_B and p_{AB} be the respective frequencies. Then quantity known as coefficient of linkage disequilibtion is $D_{AB} = p_{AB} - p_A p_B$, combination AB is called haplotype and p_{AB} is haplotype frequency. LD is commonly quantified using

$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}$$

which is a correlation coefficient indicating the presence of alleles A and B. Haplotypes can be found a setting a proper threshold for r^2 or D which is specific for every allele pair. $D = 0$ means that alleles are statistically independent and thus in linkage equilibrium. An LD block consists of loci that are in LD. The name orginates from a common practise to graphically distinguish pairs of loci with high levels of LD (Slatkin, 2008).

2.2 Gene and protein expression

The central dogma of molecular biology describes how information flows through the cells via stages as DNA replication, transcription and translation (Crick, 1958, 1970). It emphasizes that DNA contains information cell needs to produce its functional molecules such as RNA and proteins.

DNA is transcribed into RNA by a molecular mechanism called transcription. In the process, enzymes called RNA polymerases produce new RNA molecules from the template strand of DNA. The other strand of DNA known as coding strand has the same nucleotide sequence as generated RNA except thymines in DNA are replaced by uracils (U) in RNA. Other main differences between the molecules are: RNA is single stranded and it incorporates ribose sugar instead of deoxyribose sugar.

Transcriptome is the ensemble of all RNA transcripts in cells composing the sample with associated transcription levels (Velculescu et al., 1997; Ozsolak and Milos, 2011). RNA transcripts can be classified into numerous categories based for example on their size, interaction partner or functional role (Kim et al., 2009). A basic distinction is between non-protein coding RNA and protein coding RNA (Mattick and Makunin, 2006). The latter can be translated into proteins by ribosomes. The regions of DNA that encode a protein or some other functional molecule is called a gene. When a gene is part of transcriptome, it is called expressed.

A gene consists of genomic regions called exons and introns that alternate with each other (Gilbert, 1978). Exons encode the coding regions while intron regions

have other functions (Chorev and Carmel, 2012). Transcribed protein coding gene sequences form first precursor messenger RNA (pre-mRNA). Subsequent splicing removes introns from pre-mRNA and joins exons together to form mature mRNA (Kornblihtt et al., 2013). Alternative splicing assembles exons in various combinations which are translated into amino acids sequences linked by peptide bonds in protein synthesis leading into many protein isoforms from a gene (Smith et al., 1989). Alternative splicing of non-protein coding genes is harder to interpret but is reckoned as universal phenomenon (Deveson et al., 2018).

Only part of genome is expressed at a given time. Gene expression can be regulated in any RNA processing stage from transcription initiation to translation. This thesis focuses on aspects affecting transcription regulation. It is complicated including several highly orchestrated processes involving interplays of various molecules (Sims et al., 2004; Saunders et al., 2006). Particularly central regulator molecules are the ones belonging to a special class of DNA binding proteins called transcription factors (TF) after their positive or negative impact on transcription through DNA binding (Karin, 1990). TFs contain DNA binding domains having affinity to bind particular DNA sequences called transcription factor binding sites (Latchman, 1997). Genomic regions that contain at least one TF binding site in the same chromosome as the gene whose transcription it affects are cis-regulatory elements (CRE) (Wittkopp and Kalay, 2012). An activating CRE near the transcription start site (TSS) of a gene is called a promoter (Butler and Kadonaga, 2002). TF binding to a promoter enables initiation of the gene transcription. Co-activating CREs that are farther away from TSSs are called enhancers (Blackwood and Kadonaga, 1998). Their function does not depend on the distance from or the orientation relative to the promoter (Maston et al., 2006; Pennacchio et al., 2013). TFs and enhancers govern cell type specific gene expression (Spitz and Furlong, 2012; Heinz et al., 2015). Certain chromatin modifications are associated with enhancers (Heintzman et al., 2007, 2009; Rada-Iglesias et al., 2011). This thesis considers both gene expression changes and interplay between TFs and enhancer chromatin modifications during cell differentiation.

Proteins consist of precise amino acid sequences that have folded into particular three-dimensional shapes, or conformations, that determine their function. All proteins are able to bind other molecules. Such interactions change the chemical and physical properties of proteins enabling dynamic functions cells perform. There are various kinds of proteins with many kinds of functions. In addition to already mentioned TFs and histones, the proteins can be for example signaling molecules such as cytokines or hormones, antibodies to foreign substances or enzymes that catalyse chemical events such as kinases catalyzing phosphorylation of other proteins. Proteins can form complex interaction networks that underlie cellular function involving accommodation in circumstances which can depend on both intra- and extracellular conditions. Proteins do most of the work of the cells. They are important in biological pathways such as signal transduction and

metabolic pathways as well as gene regulatory networks.

2.2.1 Data from nucleotide sequences

Many experimental technologies, such as many forms of high-throughput sequencing, DNA microarrays and real-time quantitative polymerase chain reaction (qPCR), are based on tracking fluorescently dyed deoxynucleotide triphosphate sequences (dNTPs) (Goodwin et al., 2016). The methods concern DNA but can be applied to RNA when it is converted to complementary DNA (cDNA) using reverse transcriptase (Lowe et al., 2017). Nearly all RNA expression levels of this study are quantified using Illumina sequencing with a few notable exceptions. Publication I used published transcriptomics data obtained with microarrays in addition to its original sequencing data and Publication II used microarrays in STAT3 silencing experiments. In addition to gene expression data, this dissertation considers genotypes and genetic variation, chromatin modifications, transcription factor binding sites and disease associated SNPs that are commonly quantified using these kinds of measurement techniques. Often data is obtained from public databases.

In general, PCR uses an enzyme called thermostable DNA polymerase to amplify short specific sections of DNA called amplicons. The process evolves in cycles, and each of them doubles the amplicon. qPCR monitors fluorescently labelled amplicon in real-time. During some cycle fluorescence is detected. Earlier detection indicates higher initial amount of the target DNA.

In DNA microarrays, single-stranded DNA (ssDNA) probes are attached to a solid surface. Target DNA present in sampled is labelled with a fluorophore and hybridized to the probes on the array. The intensities of the fluorescent signals measured by microarray scanners are used to quantify the bound molecules (Heller, 2002). The signal strengths must be carefully pre-processed in order to obtain values that can be treated as proxies for molecule abundances (Draghici, 2003). The methods for both preprocessing and data analysis are well established as the technology has been widely used for a long time. In addition to transcriptomics, DNA microarrays have been applied in studies focusing on biological questions resembling haplotyping and protein binding (Hoheisel, 2006).

The main difference between microarray and next-generation sequencing (NGS) is that NGS measures the nucleotide composition of DNA fragments directly instead of utilizing predefined probes. There are various sequencing platforms with specific properties and technical details affecting the experimental procedure (Goodwin et al., 2016). Illumina sequencer shown in Figure 2.2 processes raw images for sequenced DNA fragments and produces nucleotide sequences, called reads, with associated PHRED quality scores in FASTQ format (Cock et al., 2010; Ewing et al., 1998). This allows removing bad quality reads, adapter sequences and assessing possible biases in nucleotide distribution etc. (Andrews, 2010). When a proper reference is available, let it be transcriptome

Figure 2.2. Illumina sequencing involves preparing the DNA library, generating DNA clusters on a solid surface by bridge amplification generation and sequencing by synthesis (Mardis, 2013) (Figure taken from Illumina (2010)).

or genome, reads are convenient to align against it. Reference sequences and annotations such as UCSC known genes (Hsu et al., 2006), Ensembl (Zerbino et al., 2018) and gencode (Harrow et al., 2012) are publicly available and well-maintained. Numerous software tools can be used to perform alignment (Fonseca et al., 2012). For example, a pipeline for differential gene expression analysis proposed by Trapnell et al. (2012) incorporates software tools Bowtie (Langmead et al., 2009) and TopHat (Trapnell et al., 2009) based on Burrows-Wheeler transformation (Burrows and Wheeler, 1994) in its alignment steps. Alignments are usually stored in human readable SAM format or corresponding compressed binary BAM format. Both file types can be manipulated with samtools software (Li et al., 2009), used with genome browsers such as IGV (Thorvaldsdóttir et al., 2013) and use with feature counting software HTSeq-count (Anders et al., 2015). The following differential gene expression analysis is based on the read counts mapping to the genes.

2.2.2 High-throughput protein expression data

In general it is harder to quantify proteome than transcriptome from biological samples. However, mass spectrometry (MS) instruments have become more sensitive and have achieved improved resolution enabling proteome-wide characterization of proteins. Now, it is possible to quantify abundances of the proteins and protein turnover rates or identify isoforms, post-translational modifications and sub-cellular localizations of the proteins. (Aebersold and Mann, 2016; Breker and Schuldiner, 2014; Larance and Lamond, 2015). Proteomics studies apply either bottom-up or top-down approach. In bottom-up methods, proteins are first cut into peptides, which are separated by liquid chromatography (LC) to reduce the complexity of the sample before MS is used (Hosp and Mann, 2017). In top-down methods no peptide cutting is applied (Toby et al., 2016). This facilitates determination of post-translational modifications and protein isoforms (Choudhary and Mann, 2010). The procedures continue with ionizing the molecules. MS measures their responses to electric and magnetic forces. There are plethora of different tools and procedures to do MS analysis in practise (Savaryn et al., 2016). Often they produce tandem mass spectra (MS/MS) containing information needed to quantify the peptides and identify the proteins (Hosp and Mann, 2017).

A basic distinction is between quantification protocols that label the samples chemically with some stable isotope and label-free quantification (LFQ) (Ahmad and Lamond, 2014; Zhang et al., 2013). The latter quantifies proteins either by MS signal intensities or fragment spectra identifying peptides of a given protein (Bantscheff et al., 2007). Intensity determination and normalization is challenging, because different peptides are present in different samples. Delaying normalization and maximazing peptide ratio extraction attempts to overcome the problem (Cox et al., 2014). In practice, computational workflows detecting proteins and quantifying them from raw MS data are implemented in widely used platforms such as MaxQuant (Tyanova et al., 2016) which we use in this thesis.

2.3 Immune system

An organism's immune system cooperatively protects it from pathogens such as viruses or bacteria. It includes various biological structures to identify and remove harmful agents threatening the system. Non-specific innate immunity is composed of barriers such as skin or acidic environment that are hostile to pathogens. Moreover, innate immunity incorporates various kinds of specific immune cells having specific abilities to control inflammatory responses or otherwise destroy pathogens as well as infected or cancerous cells (Janeway et al., 2001).

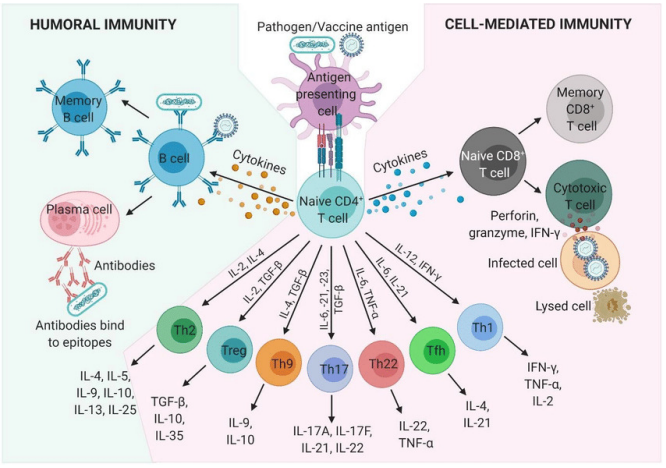

Figure 2.3. $CD4^+$ cells are central in orchestrating adaptive immune responses. When an antigen on a Naïve Th cell is recognized, it begins to polarize into specific Th cell subtypes based on affecting cytokines. Different kinds of Th cells produce specific cytokine profiles which accordingly stimulate other immune cells (Figure reproduced from Schijns et al. (2021) with licence CC BY 4.0 https://creativecommons.org/licenses/by/4.0/.)

Adaptive immunity involves specialized cells for clearing pathogens in a systemic way. The two types of adaptive immunity are humoral immunity mediated by B cells as well as cell-mediated immunity incorporating T cells illustrated

in Figure 2.3. The surfaces of these cells have specific receptors for substances that cause immune responses commonly known as antigens. Receptors for T and B cells are, well, T cell receptors (TCR) and B cell receptors (BCR). (Alberts et al., 2002). Based on presence of co-receptors, cluster of differentation 4 (CD4) and 8 (CD8), T cells have been subclassified into $CD4^+$ T helper (Th) cells or $CD8^+$ cytotoxic T cells. Major histocompatibility complex (MHC) on antigen presenting cell (APC) interact with CD4 or CD8 co-receptor making T cells response. Cytotoxic T cells kill infected cells and T helper cells secrete specific cytokine proteins other immune cells use for their purposes.

2.3.1 Th cells

Th cells first develop in thymus, then they move to spleen and lymph nodes where they mature into naive T cells. When a naive T cell recognizes an antigen represented by MHC class II protein on APC, it becomes activated. Complete activation, that leads to actual T cell proliferation and polarization, requires two additional conditions to hold. Co-stimulatory protein CD28 must have sufficient stimulation from its counterparts on APC and T cell must have proper cytokine stimulus (Corthay, 2006). Cytokine environment determines which of four major T cell subtypes (Th1, Th2, Th17 or Treg) it begins to differentiate into (Zhu et al., 2010). Th1 and Th2 were the first recognized Th cell types (Mosmann et al., 1986). Later discovered Th17 and Treg form another pair of Th cell types. Former has been associated with tissue inflammation and autoimmune disorders (Korn et al., 2009) while latter does the opposite suppressing the immune reactions (Bettelli et al., 2006). Cell type is recognized by specific signature cytokines they produce or the gene expression of regulatory transcription factors (Zhu et al., 2010). However, already differentiated Th cells are able to change their function according to changing circumstances (DuPage and Bluestone, 2016).

3. Statistical inference

The research problems of this thesis are data-driven requiring the selection of suitable statistical inference methods for each of them. For example, detecting whether phenotype-associated genetic variations are significantly co-localized or not in given genomic regions calls for hypothesis testing on an empirical distribution. Assessing differences in gene expression both at transcription and post-translational level between different kinds of cells calls for regression analysis and associated hypothesis testing. Such differential gene expression (DGE) screen reveals differentially expressed genes which can be further divided into up-regulated or down-regulated genes or proteins. Downstream analyses such as pathway enrichment or gene set enrichment analysis (GSEA) pursue profound insight into the actual cellular functions occurring in cells. Furthermore, mechanistic modeling presumes linking models to data, assessing the mathematical properties of a dynamic model, and a model selection procedure to estimate the suitability of models to describe data. This chapter introduces the key concept of the applied statistical techniques used in Publications I-V.

3.1 Statistical hypothesis testing and empirical p-value

Statistical inference commonly utilizes a method called statistical testing. The underlying idea is that observed data are an outcome of random variables and a statistical model can be constructed to describe the probability distribution of the random variables. Often a statistical model is given as the probability density function f. If f is fully determined using a particular mathematical model with parameters ψ, the model and associated methods and techniques are called parametric. If a model does not have a fixed number of parameters but the effective number of parameters increase with the number of data points instead, then the model is called non-parametric. In any case, it is possible to set a conservative null hypothesis H_0 to suggest that a simple model is responsible for the observations. Test statistic T is selected to measure the discrepancy between the data and H_0. By convention, large values of T are considered evidence against H_0. The level of evidence against H_0 is measured by the p-value or the

significance probability

$$p = \Pr(T \geq t | H_0), \tag{3.1}$$

where t is the observed value of T. Let t_p be such that from $t_p \geq t$ follows that H_0 is rejected at level p, or $100p\%$, then t_p is called critical value for t. Formal definition of t_p, $\Pr(T \geq t_p | H_0) = p$ results in an interpretation that p is the error rate of the test. The distribution under H_0 is called the null distribution of T.

The p-value 3.1 has a uniform distribution on $[0,1]$ under H_0 if T is continuous. Hence, corresponding random variable P has a distribution

$$\Pr(P \leq p | H_0) = p. \tag{3.2}$$

Sometimes, it is accentuated that this is not exactly true if T discrete. Still, Equation 3.2 can be considered as a valid estimation even if T is continuous. Equations 3.1 and 3.2 are equivalent which highlights the mentioned interpretation of p-value being the error rate of the test.

It is important to note that a p-value comes from a single test whereas typically in bioinformatics multiple comparisons are applied to the same data simultaneously. When many tests are performed, it becomes increasingly likely that significant p-values occur by random chance. Therefore, p-value is not a feasible measure of statistical significance when simultaneous multiple comparisons are performed. There are many ways to control this problem such as family-wise error rate (FWER) and false discovery rate (FDR) which is the expected proportion of the accepted tests, when the null hypothesis should have been rejected, to all tests found significant. In this thesis, false discovery rates are controlled using Benjamini-Hochberg method. It assumes the tests are independent. First, the p-values are ordered in ascending order. The smallest p-value has a rank of $i = 1$, then next smallest has $i = 2$, etc. Then, each individual p-value is compared to its Benjamini-Hochberg critical value, $\left(\frac{i}{m}\right) q^*$, where i is the rank, m is the total number of tests, and q^* is the set false discovery rate. The largest p-value that has $p \leq \left(\frac{i}{m}\right) q^*$ is significant, and so are all other p-values that are smaller than it (Benjamini and Hochberg, 1995). Even though the method as such does not produce adjusted p-values, the procedure is commonly interpreted to define adjusted p-value as $p_{\text{adjusted}} = p_{\text{original}} \frac{m}{i}$ which produces a list of adjusted p-values in an ascending order. In practice, the order of adjusted p-values is enforced by setting $q_{i+1} = q_i$ whenever $q_i > q_{i+1}$. Such list enables a more flexible inspection of multiple comparisons than setting a fixed q^* in advance. The method is straightforward and hence widely applied, even when it is not realistic to assume that data is independent.

Consider a sample of data $y_1, \ldots, y_n$ and assume it is an outcome of independent and identically distributed random variables $Y_1, \ldots, Y_n$. It is possible to base a statistical test either on a theoretical model for the data or just on the actual observed — or empirical — data. In the latter case, the data is said to follow an empirical distribution that puts equal probabilities n^{-1} at each sample value y_j. The probability distribution is described nonparametrically using the empirical

distribution function (EDF), $\widehat{F}$, which is defined by

$$\widehat{F}(y) = \frac{\#\{y_i \leq y\}}{n},$$

or more formally

$$\widehat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} H(y - y_j), \tag{3.3}$$

where H is the Heaviside function

$$H(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0. \end{cases}$$

A suitable test statistic T is selected and null distribution H_0 is formed in order to apply EDF in statistical testing. The empirical distribution $\widehat{F}$ of test statistic T following null hypothesis H_0 is denoted by $\widehat{F}_0$. In practice, it may be difficult to calculate the exact empirical p-value

$$p = P(T \geq t|\widehat{F}_0). \tag{3.4}$$

Hence, it is approximated by comparing observed test statistic t to R independent replicate samples $t_1^*, \ldots, t_R^*$ drawn from $\widehat{F}_0$. The p-value is then approximated by simple proportion of the drawn test statistics t^* reaching or exceeding t

$$p\text{emp} = \frac{\#\{t^* \geq t\}}{R}. \tag{3.5}$$

Sometimes in literature approximation (3.5) is modified by adding ones to both numerator and denominator (Davison et al., 1997, p. 148). By definition, empirical significance testing involves drawing repeated samples of empirical null distribution of test statistic. However, test statistics can also be obtained indirectly without drawing random numbers from $\widehat{F}_0$ directly. In such case, one computes test statistics for the original observed data set y and for data samples $y_1, \ldots, y_R$ that are drawn from the same empirical distribution as y is drawn. This approach is adopted to the SNPs and genomic regions co-localization analysis. The observed data set and the corresponding randomly drawn data sets were all composed of equally many SNPs. Computed test statistic were the counted overlaps between the SNPs in the sets and given genomic regions.

3.2 Linear and generalized linear models for detecting differentially expressed genes

Consider detection of differential gene expression (DGE) both at transcriptional and at post-translational level between two sample groups or biological conditions. DGE and the related downstream analysis connecting it to actual biological functions occurring in the cells motivates the transcriptomics and proteomics screens in this thesis. The analysis is based on statistical regression

where the models take the distinctive character of the data obtained by different high-throughput technologies into account. Mass spectrometry quantifies proteomics data which is summarized as protein intensities considered normally distributed after logarithmic transformation. In this case, the standard linear model is a natural selection (Hoyle et al., 2002; Kammers et al., 2015). RNA sequencing (RNA-seq) read data are mapped to the reference transcriptome and summarized as read counts mapping to the genes thought to follow a distribution that can be approximated well with a distribution from the exponential family and hence generalized linear model is well-justified (Anders et al., 2013). Nevertheless, this is not the only option. Huang et al. (2015), for example, reviews the methods of differential expression analysis for RNA-seq and acknowledges several alternative approaches that make different assumptions of data distribution. For example, Poisson, beta binomial, full Bayesian, empirical Bayesian and nonparametric models have been used.

Assuming data vector $\mathbf{y}_i$ for a gene i is normally distributed, linear model is defined by expectation of $\mathbf{y}_i$ being a linear combination of the predictor variables, given as design matrix $\mathbf{X}$, and regression coefficients $\boldsymbol{\beta}$

$$E(\mathbf{y}_i) = \mathbf{X}\boldsymbol{\beta}. \tag{3.6}$$

Thus, $\mathbf{y}_i \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$. Generalized linear models (GLMs) extend the concept by relaxing the normality assumption by introducing an invertible link function g to Equation (3.6),

$$E(\mathbf{y}_i) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{3.7}$$

and a function which tells how the variance depends on the mean.

The data analysis applies two software tools based on these model families. Limma (linear models for microarray data) is used when the summarized data values are considered as normally distributed on logarithmic scale (Smyth, 2005). Othwerwise, the data is summarized as read counts and treated as generalized linear model with edgeR (Robinson et al., 2010). Both tools are widely used in bioinformatics as they are implemented as part of Bioconductor project in R language (Gentleman et al., 2004; Ihaka and Gentleman, 1996). The benefits of Bioconductor packages include they are well developed as well as maintained and documented actively. It is worth to note that edgeR is not the only software package in Bioconductor that approaches DGE with similar statistical framework. Especially, DESeq (Anders and Huber, 2010) and later DESeq2 (Love et al., 2014) have been somewhat popular in similar studies.

Originally, limma was developed to facilitate construction of linear models to enable DGE from microarray data under various experimental designs (Smyth, 2005). This is accomplished by constructing individual linear models of the form of Equation (3.6) through two matrices. Design matrix $\boldsymbol{X}$ indicates the conditions affecting the samples, and contrast matrix $\boldsymbol{C}$ indicates which comparisons are made between the samples. The fitted models yield gene-specific contrasts of coefficients $\widehat{\boldsymbol{\beta}}_i$. Importantly, analysis is based on an empirical Bayes model whose global hyperparameters connects the gene-wise models together. This enables

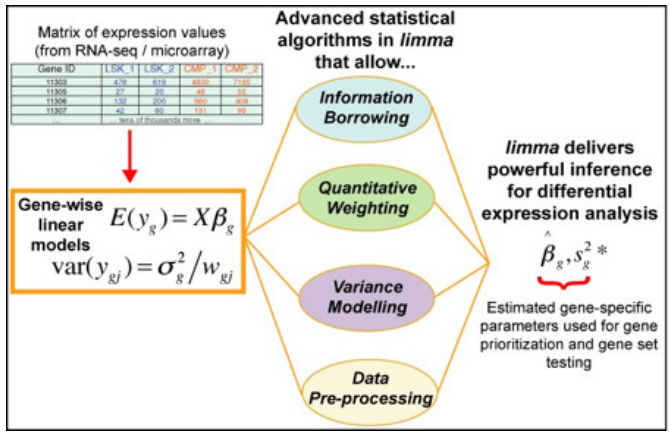

Figure 3.1. Limma provides statistical tools to manage experiments with small sample sizes and variant quality gene signals for more powerful statistical testing (Figure from Ritchie et al. (2015).)

information borrowing between the genes to obtain posterior variance estimator s_i^{*2} for each gene as shown in Figure 3.1 which aims to improve statistical testing especially when experiment has a small size. Modified variance estimation is used to build a test statistic t for null hypothesis $H_0 : \beta_{ij} = 0$, for a gene i and constrast of coefficients j under F-distribution. More recently, application of limma is spead from microarray studies into proteomics and RNA-seq data analysis (Kammers et al., 2015; Law et al., 2014; Liu et al., 2015). In this thesis, limma is used in detecting differentially abundant proteins whereas RNA-seq analysis is based on generalized linear models as implemented in edgeR.

EdgeR models count-based data with the negative binomial generalized linear model using log link function. A reasoning for this approach can be derived considering read counts as random variables. The number of reads in a particular gene (or any genomic region) i in a sample j can be considered as random variable Y_{ij} which follows binomial distribution $\mathscr{B}(N_j, p_{ij})$ where N_j is the total number of reads in the sample and success probability p_{ij} is the unknown relative expression of gene i. If G is the total number of genes in j, then $\sum_{i=1}^{G} p_{ij} = 1$. In this case, the probability p_{ij} is usually quite small because of the large size of a genome even if N_j is large. Thus, binomial distribution can be approximated as Poisson distribution Pois(λ) , where parameter λ represents both the mean and the variance that are equal (Hodges and Le Cam, 1960). However in practise, Poisson distribution is known to capture technical variation but not biological variation between samples (Marioni et al., 2008). In fact, variation between samples often exceeds the mean. The overdispersion relative to the mean can be modeled with negative binomial (NB) distribution $Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_i)$, where the mean is

$$\mathrm{E}(Y_{ij}) = \mu_{ij} = p_{ij} N_j$$

and the variance is

$$\mathrm{var}(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}\phi_i). \tag{3.8}$$

Log-linear model for each gene is

$$\log\left(\frac{\mu_{ij}}{N_j}\right) = \mathbf{x}_j^T \boldsymbol{\beta}_i,$$

where $\mathbf{x}_j^T$ is the vector of covariates that specifies the treatment conditions applied to sample j in the experimental design, and $\boldsymbol{\beta}_i$ is a vector of regression coefficients by which the covariate effects are mediated for gene i. Equation 3.8 embodies quadratic relationship with the mean depending on an additional dispersion parameter ϕ_i (Robinson and Smyth, 2007). This stresses the importance of finding a reliable estimate for ϕ_i as the variance, and so much of the inference rests upon it. Usually in GLM theory, model is fitted iteratively by using Newton-Rhapson algorithm which is equivalent with Fisher scoring $\mathscr{I}_i$ and leads to maximum likelihood estimates (MLE) (McCullagh and Nelder, 1989, p. 40-43). While maximum likelihood estimates are intuitive and easy to form given the data, MLEs for variance parameters are systematically underestimated (Robinson and Smyth, 2008). McCarthy et al. (2012) suggests applying Cox-Read adjusted profile likelihood (APL) for reducing the bias in MLE

$$\mathrm{APL}_i(\phi_i) = l(\phi_i; \mathbf{y}_i, \widehat{\boldsymbol{\beta}}_i) - \frac{1}{2}\log\det\mathscr{I}_i,$$

where l is the log-likelihood function, $\mathbf{y}_i$ is the vector of read counts for gene i, $\widehat{\boldsymbol{\beta}}_i$ is the estimated coefficient vector and $\mathscr{I}_i$ is Fisher information matrix. APL can be used to share information across genes in dispersion estimation in three alternative ways: (i) estimate a common dispersion, so that $\phi_S = \phi$, by maximizing the shared likelihood function $\mathrm{APL}_S(\phi) = \frac{1}{G}\sum_{i=1}^{G}\mathrm{APL}_i(\phi)$, (ii) estimate trended dispersion ϕ_{Si} using local shared log-likelihood $\mathrm{APL}_{Si}\left(\phi_i\right)$ defined as a weighted average of the APLs for gene i and its neighbouring genes by average read counts, or (iii) estimate genewise ϕ_i by maximizing $\mathrm{APL}_i(\phi_i) + G_0\mathrm{APL}_{Si}(\phi_i)$ where G_0 is the weight given for the local shared log-likelihood. Likewise in microarray transcriptomics, information sharing in dispersion estimation aims to more robust inference where number of biological replicates per condition of RNA-seq experiments is small. Use of contrast matrices enables formulating variously complex null hypotheses which can be tested using either quasi-likelihood (QL) F-test or likelihood ratio test.

3.3 Functional analysis of genes and proteins

A list of differentially expressed genes - or abundant proteins - may be a valuable resource for biologists interested in particular molecules helping them to direct their further work. However in general, such list is seldom a satisfactory end point for a high-throughput screen. More often, the interest is in the physiological or functional meaning of the detected molecules. This leads to investigating molecules as groups with an underlying idea that, if a cellular process is atypical, then the related molecules should emerge as a group. Nowadays, many public

databases provide functional annotations for genes. This enables utilization of statistical testing to decide whether an annotation is overpresented given genes detected in the study. There are numerous computational tools using different statistical tests for functional enrichment analysis (Huang et al., 2008). This thesis applies two widely applied methods. Hypergeometric test corcerns only genes that are differentially expressed whereas gene set enrichment analysis (GSEA) covers all genes with positive read count (Tavazoie et al., 1999; Subramanian et al., 2005).

The hypergeometric distribution considers sampling without replacement. It is commonly explained by a metaphor of drawing marbles from an urn or a jar. Obviously, there can be only a finite number of marbles N in the urn. Assuming the marbles are either blue or white implies that, if K marbles are blue, then the rest $N-K$ are white. Say we are interested in blue ones. That defines drawing a blue marble as a success and white as a failure. Let the random variable X be number of successes in the sample consisting n draws. Now, probability mass function (pmf) of the hypergeometric distribution of observing exactly k successes is defined by

$$f(k,N,K,n) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where n is the number of draws. This can be used to decide whether or not the observed k number of successes (i.e., blue marbles) is statistically significant or not. Let us set a conservative null hypothesis that n observations do not correlate with the specific class composed of K objects. Observing k overlaps between the observations and class members implies p-value $p = 1 - \sum_{i=0}^{k-1} f(i,N,K,n)$. Let us return to bioinformatics from the illustrative urn parable. Now, the hypergeometric test is applied separately for each pathway as shown in Figure 3.2. Variable n is the number of differentially expressed genes (DEGs), N is the number of genes in an organism, K is the number of genes in the annotation, $N-K$ is the number of genes not in the annotation, k is the number of DEGs in the annotion and $n-k$ DEGs in the sample that are not in the annotation. The choice of gene universe affects heavily hypergeometric test results. In this thesis, gene universe consists of genes with successfully quantified (i.e., non-zero) signal in the high-throughput screen. This kind of straightforward post-processing of DEGs gives an intuitive view on the altered cellular functions between the biological conditions. However, it misses subtler differences not detected by DGE analysis in the first place. GSEA pursuits to detect coordinated but possibly modest changes in gene expression on some prespecified sets of related genes such as pathways by considering all genes in DGE screen instead of a DEG list (Mootha et al., 2003).

GSEA is a nonparametric method. It involves all quantified genes in study. That means all proteins quantified by MS, genes having probes in a microarray, or genes having positive read counts in a RNA-seq experiment. The method is based on a running sum statistic going through the ranked gene list from

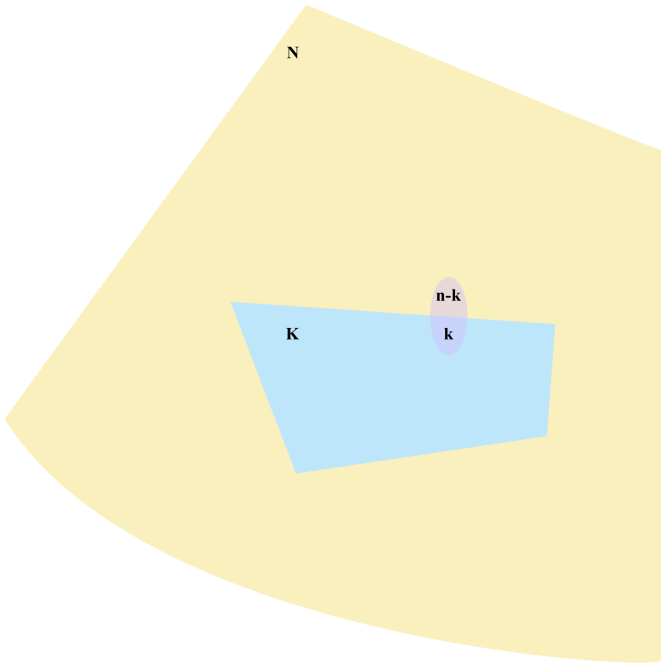

Figure 3.2. Hypergeometric test shown as a Venn diagram. Yellow area and the other colored areas surrounded by it depict the set of all genes in the gene universe. Purple shows the genes involved in a given annotation while blue represents the set of DEGs. Different shades of purple emphasize that some of the genes in the annotation have been found as DEGs in the preceding DGE screen whereas some others have not met the statistical significance level.

top to bottom. The sum is increased when the gene is in the annotation and decreased when not. The size of the increment depends on the correlation of the gene with the phenotype. The enrichment score (ES) is the approximated weighted Kolmogorov–Smirnov-like statistic. The score aims to reflect how much an annotation is overrepresented at the top or bottom of the ranked list. P-value for a gene set is derived using a permutation test. When multiple gene sets are evaluated, ES for each gene is normalized to account for the size of the set and FDRs are used to control multiple testing problem (Subramanian et al., 2005).

3.4 Modeling biological systems

Biological systems can be viewed as deterministic dynamic systems which are often modeled by systems of ordinary differential equations (ODEs)

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{x}}, t) \tag{3.9}$$

where the state vector $\mathbf{x} \in \mathbb{R}^{n_{\mathbf{x}}}$ often consists of the concentrations of the molecules at time t and $\boldsymbol{\theta}_{\boldsymbol{x}} \in \mathbb{R}^{n_\theta}$ is the parameter vector. The initial value $\mathbf{x}_0 = \mathbf{x}(t_0)$ is the state at initial time t_0. Initial values can be considered as parameters as

well. Vector field $f : \mathbb{R}^{n_{\mathbf{x}}} \times \mathbb{R}^{n_{\boldsymbol{\theta}}} \rightarrow \mathbb{R}^{n_{\mathbf{x}}}$ defines the dynamics of the biochemical species. It describes the structure of the model by specifying the relations between molecules as well as their production and degradation mechanisms. If the modeled system is linear and sufficiently small, the initial value problem $\mathbf{x}(t, \boldsymbol{\theta}_x) : [t_0, t_n] \times \mathbb{R}^{n_{\boldsymbol{\theta}_x}} \rightarrow \mathbb{R}^{n_{\mathbf{x}}}$ may be both possible and feasible to be solved in a closed form

$$\mathbf{x} = \mathbf{x}_0 + \int_{t_0}^{t_n} f(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{x}}, t)\,\mathrm{d}t. \tag{3.10}$$

Generally in computational biology this is not the case and solutions for the initial value problems are approximated by means of numerical integration. This thesis utilizes highly optimized numerical ODE solver CVODES implemented in the SUNDIALS package (Hindmarsh et al., 2005). The systems of ODEs in this thesis are moderately large compared to some rather ambitious efforts in computational system biology. In practice, this means the number of the model parameters does not overwhelm the number of the model outputs indicating forward sensitivity analysis (FSA) is sufficient in parameter estimation. Some of the candidate models in our studies, may contain wirings that lead to numerically stiff problems. Therefore, backward differentiation formula (BDF) with Newton iteration and dense Jacobians are applied to solve ODEs numerically. The technical details of the methods are explained by Serban and Hindmarsh (2003). Perhaps computational efficiency and flexibility have made CVODES a popular solver. It is incorporated for example in several platforms designed for computational biology (Raue et al., 2015; Fröhlich et al., 2017; Stapor et al., 2017).

Depending on the modeled system, the structure can be assumed to remain static over time, or it may be subject to temporal evolution. In Publication IV, we modeled the latter case by introducing latent processes $\ell(t, \boldsymbol{\theta}_\ell) : [t_0, t_n] \times \mathbb{R}^{n_{\boldsymbol{\theta}_\ell}} \rightarrow \mathbb{R}^{n_\ell}$ where $n_{\boldsymbol{\theta}_\ell}$ represents the number of parameters $\boldsymbol{\theta}_\ell$ of the latent state ℓ that is able to rewire the system and n_ℓ is the dimension of the latent state. Each component of Equation 3.9 is coupled with latent states via a weight function w which controls the impact of the latent state depending on the concurrent structure and time

$$\frac{dx_i}{dt} = \sum_{j=1}^{J_i} f_{ij}(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{x}}) w(t, \ell, \boldsymbol{\theta}_\ell, z_{j1}, \ldots, z_{jn_\ell}), \tag{3.11}$$

where $i = 1, \ldots, n_{\boldsymbol{x}}$ and J_i is the number of terms affecting x_i inducing $J = \sum_{i=1}^{n_{\boldsymbol{x}}} J_i$ and $Z \in \{0,1\}^{J \times n_\ell}$ which is the incidence matrix indicating relations between variables, and the weight function is

$$w(t, \ell, \boldsymbol{\theta}_\ell, z_{j1}, \ldots, z_{jn_\ell}) = \frac{\sum_{k=1}^{n_\ell} z_k \ell_k(t, \boldsymbol{\theta}_\ell)}{\sum_{k=1}^{n_\ell} \ell_k(t, \boldsymbol{\theta}_\ell)}.$$

The solution for the coupled ODE system 3.11 is $\mathbf{x}(t, \boldsymbol{\theta}_{\mathbf{x}}, \ell(t, \boldsymbol{\theta}_\ell), \mathbf{Z}) \in \mathbb{R}^{n_{\mathbf{x}}}$.

It is vital to link theoretical models and observables to experimental time-series data $\mathscr{D} = \{\{\overline{x}_{ik}\}_{i=1}^{n_{\mathbf{x}}}, t_k\}_{k=1}^{n_t}$ using statistical models that characterize well

the modeled problem and respect the properties of the data. This ensures the relevance of the study as proper statistical models enable weighing the data objectively, disseminating the uncertainty of the measurements through the models and facilitate interpreting the obtained results. In our work, the noise is assumed to have normal distribution. Hence, the experimental data $\mathscr{D}$ is modeled by

$$\bar{x}_{ik} = x_i(t_k, \boldsymbol{\theta}_{\boldsymbol{x}}) + \epsilon_{ik},$$

where $\epsilon_{ik} \sim \mathcal{N}(0, \sigma_{ik}^2)$ and $x_i(t_k, \boldsymbol{\theta}_{\boldsymbol{x}})$ is used as a shorthand to denote the solution $x_i(t, \boldsymbol{\theta}_{\mathbf{x}}, \ell(t, \boldsymbol{\theta}_\ell), \mathbf{Z})$. In addition to the model parameters $\boldsymbol{\theta}_{\boldsymbol{x}}$, a model may contain variables, such as standard deviation σ, that can be considered as parameters. Hence, we denote parameter vector more generally by $\boldsymbol{\theta}$. The modeled observations are calibrated to the experimental data using maximum likelihood estimation. The likelihood of observing $\mathscr{D}$ for the parameters $\boldsymbol{\theta}$ is

$$\mathscr{L}_{\mathscr{D}}(\boldsymbol{\theta}) = \prod_{i=1}^{n_{\mathbf{x}}} \prod_{k=1}^{n_t} \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(\bar{x}_{ik} - x_i(t_k, \boldsymbol{\theta}_{\boldsymbol{x}}))^2}{2\sigma_{ik}^2} \right) \tag{3.12}$$

and MLE $\widehat{\boldsymbol{\theta}}$ for the parameters is

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}}{\arg\min}\, \mathscr{L}(\boldsymbol{\theta}). \tag{3.13}$$

$\mathscr{L}_{\mathscr{D}}(\boldsymbol{\theta})$ depends on the solution of the model via observables. Therefore, estimating $\widehat{\boldsymbol{\theta}}$ is an ODE-constained optimization problem. Generally, ODEs must be solved numerically as they rarely have closed form solutions. In order to obtain as stable numerics as possible, $\widehat{\boldsymbol{\theta}}$ are estimated by minimizing the negative logarithm of the likelihood. The objective function $\mathscr{J}(\boldsymbol{\theta}) = -\ln(\mathscr{L}_{\mathscr{D}}(\boldsymbol{\theta}))$ is

$$\mathscr{J}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{n_{\mathbf{x}}} \sum_{k=1}^{n_t} \left(\ln 2\pi\sigma_{ik}^2 + \left(\frac{(\bar{x}_{ik} - \mathbf{x}(\boldsymbol{\theta}_{\boldsymbol{x}}, t_k))^2}{\sigma_{ik}^2} \right) \right). \tag{3.14}$$

Usually in computational biology, the objective function may have multiple local minima caused due to the optimization problem which is often nonlinear and non-convex. This calls for a global optimization procedure. This thesis applies local multistart optimization with initial values derived from latin hypercupe sampling. In latin hypercube sampling, N samples of parameters $\boldsymbol{\theta}$ are obtained by dividing the ranges of every parameter θ_i into N non-overlapping segments, each having equal probability, and then sample N points of each segments segments so that each sample is the only one in each axis-aligned hyperplane containing it (Owen, 1992). Each multistart is optimized using interior-point algorithm.

Interior-point algorithm demands the gradient of the objective function with respect to the parameters. It is estimated with forward sensitivity analysis known to produce accurate gradients (Hindmarsh et al., 2005). The gradients are computed upon the output trajectories of the system $\mathbf{x}(t, \boldsymbol{\theta}_{\boldsymbol{x}})$ which we

consider as observables. The (first order) forward sensitivities are obtained simply by differentiating the ODEs with respect to its parameters

$$s^{\boldsymbol{x}} = \frac{d}{dt}\frac{\boldsymbol{x}(t,\boldsymbol{\theta}_{\boldsymbol{x}})}{d\boldsymbol{\theta}_{\boldsymbol{x}}}.$$

They are used to obtain the gradient of the objective function 3.14. If variance is fixed and known, the gradient has form

$$\frac{\mathscr{J}}{\partial\theta_j} = \sum_{i=1}^{n_{\mathbf{x}}}\sum_{k=1}^{n_{\boldsymbol{t}}}\frac{\overline{x}_{ik} - x_i(t_k)}{\sigma_{ik}^2}s_j^{x_i}.$$

In some cases variance can be considered as a parameter i.e. a component in the parameter vector $\boldsymbol{\theta}$ and is inferred with the rest of the model. In such cases, we assume data is homoscedastic. Hence, the gradient of variance has form

$$\frac{\mathscr{J}}{\partial\sigma^2} = \frac{1}{2}\sum_{i=1}^{n_{\mathbf{x}}}\sum_{k=1}^{n_{\boldsymbol{t}}}\left(\frac{1}{\sigma^2} - \frac{(\overline{x}_{ik} - x_i(t_k))^2}{\left(\sigma^2\right)^2}\right).$$

Parameters define the dynamic behaviour of the model. Therefore, they may carry a lot of information about the dynamics themselves as well as the model and its relation to the data. In practice, the parameters and specifically their identifiabilities are assessed with profile likelihoods. Profile likelihood of parameter θ_k, $k \in \{1,\ldots,n_\theta\}$ at point c is defined by

$$\mathrm{PL}_{\theta_k}(c) = \max_{c=\theta_k,\boldsymbol{\theta}\in\Omega} -\mathscr{J}(\boldsymbol{\theta}). \tag{3.15}$$

PL is computed by solving a sequence of optimization problems where $\mathrm{PL}_{\theta_k}(c)$ is computed in a grid $\{c_l\}$. For each point c, the negative log-likelihood function 3.14 is minimized subject to $\theta_k = c$. The minimization yields the optimal parameter vector, $\boldsymbol{\theta}_c^*$, the corresponding negative log-likelihood function $\mathscr{J}(\boldsymbol{\theta}_c^*)$ and the profile likelihood $\mathrm{PL}_{\theta_k}(c) = -\mathscr{J}(\boldsymbol{\theta}_c^*)$. The process starts from MLE $\boldsymbol{\theta}^*$ and progresses iteratively in both directions. In each iteration, new value for c is selected adaptively, and parameters are optimized locally.

Two times the negative log-likelihood $2\mathscr{J}$ is also known as χ^2 or the goodness of fit statistic. It is used to derive the confidence intervals for the profile likelihoods. The confidence interval of the k^{th} parameter at the confidence level α for profile likelihood 3.15 is defined asymptotically by

$$\mathrm{CI}_{\theta_k,\alpha} = \left\{ c \mid \mathrm{PL}_{\theta_k}(c) \geq \min_{\theta} -\mathscr{J}(\boldsymbol{\theta}) - \frac{\Delta(\alpha)}{2} \right\} \tag{3.16}$$

where the threshold $\Delta(\alpha)$ is given by α-quantiles of χ^2 distribution with one degree of freedom. The confidence intervals classify parameters in three categories. Identifiable parameters have clear profile likelihoods with unique maximum and a restricted interval for the required threshold. Structurally non-identifiable parameters have flat profile likelihoods indicating infinite confidence intervals for any confidence level α. It means that any change of value of such parameters

does not affect the likelihood as the other parameters are able to compensate the changes. The data provides no information about such parameters. Practically non-identifiable parameters have such profile likelihoods that exhibit unique optimum but do cross not the threshold. In such a case, the data contain information about the parameter, but the parameter range is unlimited. Figure 3.3 illustrates the identifiabilities of two parameters in terms of profile likelihoods.

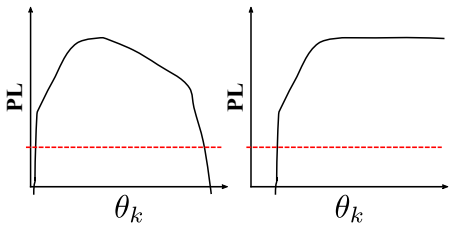

Figure 3.3. An identifiable parameter is on the left and a practically unidentifiable parameter on the right.

This thesis models such cellular subsystems that do not provide sufficient theoretical information to enable specifying a mathematical model a priori without a doubt. Instead, models describing several plausible explanations for the phenotype are considered in order to infer the best possible model from data. Unlike previous sections where inference were based on statistical testing, model selection is based on model ranking. A natural idea is to rank models according to MLEs. However, when comparing MLEs of different models, there tends to be bias that more complex models overfit the data. Bayesian information criteria (BIC) for model $\mathscr{M}_i$ corrects the bias by penalizing the maximum log-likelihood $\ln \mathscr{L}_{\mathscr{D}} \left(\widehat{\boldsymbol{\theta}}_{\mathscr{M}_i} \right)$ with the number of data points $n_{\mathscr{D}}$ and the number of parameters $n_{\boldsymbol{\theta}_{\mathscr{M}_i}}$

$$\mathrm{BIC}_{\mathscr{M}_i} = \ln \mathscr{L}_{\mathscr{D}} \left(\widehat{\boldsymbol{\theta}}_{\mathscr{M}_i} \right) - \frac{1}{2} n_{\boldsymbol{\theta}_{\mathscr{M}_i}} \ln n_{\mathscr{D}}. \tag{3.17}$$

The concept is derived in Bayesian context, by approximating marginal likelihood of the data given model $\Pr(\mathscr{D} \mid \mathscr{M}_i)$ with the Laplace approximation of an integral. Equation 3.17 defines BIC for the regular systems of ODEs of the form Equation 3.9. Precise treatment of coupled models defined by Equation 3.11 involves prior probabilities of the model structures $\Pr(\mathbf{Z})$. In general in model selection holds $\Pr(\mathbf{Z} \mid \mathscr{D}) = \frac{\Pr(\mathscr{D}|\mathbf{Z}) \Pr(\mathbf{Z})}{\Pr(\mathscr{D})}$. However in this work, all model structures are assumed to be equally probable a priori , and $\ln(\Pr(\mathbf{Z}|\mathscr{D}))$ reduces to Equation 3.17. When the number of the candidate models is reasonable small, it is conceivable to compute all BICs and find the highest ranked model. However, when the number of models increases, this it not always possible in practice. In such cases, model selection can be done using a greedy stepwise forward-backward model selection algorithm. It records the best model structure, causes individual changes into it and monitors their impact on BIC. Process is iterated until no improvement occurs.

This thesis applies Bayesian information criteria in model ranking as it is widely used and easy to evaluate. However, it has limitations to be aware of. It

assumes that correct models are under consideration - which may not always hold for every experiment (Kuha, 2004). Moreover, BICs does not contain prior information of parameters $\boldsymbol{\theta}$ as it has been reduced from the formula based on assumption the prior is unknown. In cases where model selection is crucially based on prior information $\Pr(\boldsymbol{\theta})$, Bayesian factors may have an advantage (Vyshemirsky and Girolami, 2007). Even though BIC would be considered inherently a non Bayesian statistic, it requires sufficiently large data sample to tightly constrain the parameter values making them well-determined (Bishop, 2006).

4. Proteomic profiles in Differentiating Th17 and iTreg cells

In this chapter I highlight some of the main results of Publication I.

4.1 Biological objective and experimental design

In this publication, we studied differentiation of Th17 and Treg cells from Naïve T helper cells (Th0). The cells are central in adaptive immunity having opposite responsibilities: Th17 drives the immune responses while Treg suppresses them. The balance of these actions are important for immune homeostasis. Disturbances are related to cancer and autoimmune diseases. Although the polarized cells have contrasting characters, they both originate from the same cell type requiring the same cytokine transforming growth factor β (TGFβ) for their cell fate. Th17 and Treg lineage commitment is mainly driven by presence of cytokines interleukin 6 (IL6) and interleukin (IL2), respectively (see Figure 2.3). The differentiation process involves thorough changes in gene expression which can be observed in transcriptomics as well proteomics level. Previous studies had concentrated on the former revealing the regulatory networks and many TFs orchestrating the cell differentiation by using NGS experiments. This work went further by considering the proteomes of the cells using the experimental procedure summarized in Figure 4.1 and applying statistical methods described in Section 3. We hypothesized that the proteome and the transcriptome profiles differed from each other within and between cell types. The main results were (1) the quantified proteome catalogues of polarizing Th17 and Treg cells, (2) protein expression changes occurred during differentiation, (3) differential protein expression between Th17 and iTreg cells, (4) comparison of gene expression between transcriptomics and proteomics levels, (5) biological verification of some of the findings. The results fit well with the hypotheses.

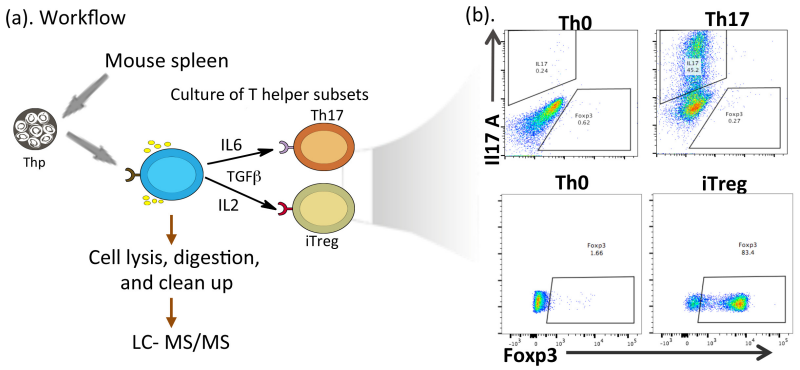

Figure 4.1. (a) The experimental procedure: Naïve $CD4^+$ T (Thp) cells were isolated from the spleens of the animals and were in vitro polarized either towards Th17 or iTreg cells. Thp cells were cultured for three days. Th17 cells were differentiated in 72 hours and iTreg cells in 10 days of culturing. (b) The success of the differentiation was confirmed by detecting expression of interleukin 17 (IL17) for Th17 and forkhead box P3 (Foxp3) for iTreg. The transcriptomes and the proteomes were quantified with Illumina sequencing described in Section 2.2.1 and lliquid chromatography (LC-MS) and label-free quantitation (LFQ) represented in Section 2.2.2, respectively. The quantified data was analyzed applying methods and tools covered in Section 3. For data analyses, experimental design was blocked pairing replicates between cultures.

4.2 Quantified proteomic profiles of polarizing Th17 and iTreg cells

This study is based on experimental work which produced quantitative proteome characterizations of polarizing murine Th17, iTreg, activated (Th0) and Thp cells. The procedure comprising the sample preparation, mass-spectrometry analysis and data quantification produced proteomics data summarized as protein intensities. For this purpose, three independent cultures of each cell types were processed. In total, intensities of 4,287 protein groups were detected.

The data showed that 1) different T cell lineages had distinguishable expression profiles, 2) the protein intensity profiles were consistent across all biological replicates of each cell type, 3) proteins across a wide range of expression were detected, 4) first quarter of the cumulative intensities was mostly attributed to the cytoskeleton and glycolytic enzymes, 5) majority of the proteins were expressed in all T cell lineages including many proteins known to be involved in all T cell types, 6) detected proteins were annotated to almost all cellular compartments, and 7) proteins annotated to extracellular space included cytokines important in deciding fate of polarizing Th cell. Together, these observations suggest the data constitute proteome profiles of T cell subsets at a reasonable coverage and resolution.

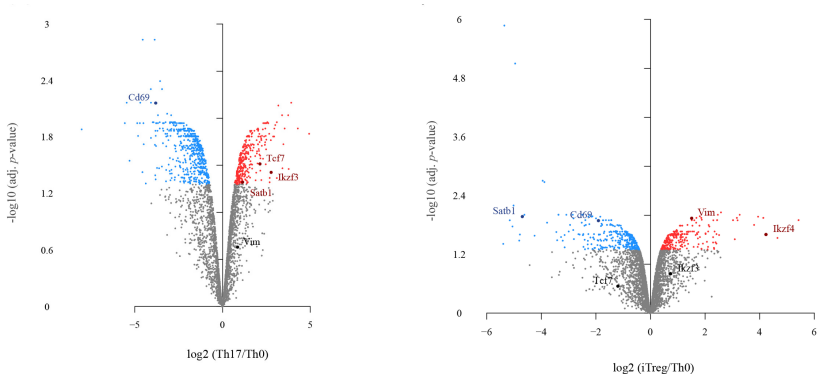

Figure 4.2. Volcano plots visualize signals of biological (fold changes) and statistical (adjusted p-values) significance of each protein with respect to each other. Up- and down-regulated proteins are denoted by red or blue color, respectively. Among DE proteins, cluster of differentiation 69 (CD69) was significantly down-regulated in both Th17 and iTreg cells. It is a marker of early T cell activation with an immune regulatory role. Special AT-rich sequence-binding protein 1 (SATB1) known to be a chromatin organizer was highly expressed in Th17 cells and lowly expressed in iTreg. Moreover, inflammation related transcription factor 7 (Tcf7) was highly expressed in Th17 cells. IKAROS family zinc finger 4 (IKZF4) and 3 (IKZF4), were highly expressed in Th17 and iTreg cell, respectively. Finally, Vimentin (Vim) was found to be highly expressed in iTreg cells.

4.3 Protein expression changes during Th17 and iTreg differentiation

In order to decipher how polarization changes the proteomes of the cells, we examined differential protein expression of Th17 and iTreg cells with respect to TCR–activated Th0 cells. This was done in two phases. Firstly, by listing proteins that were exclusively expressed in polarizing cell type. Secondly, performing differential gene expression analysis at protein level followed by pathway enrichment analysis introduced in sections 3.2 and 3.3.

Th17 cells expressed exclusively 40 proteins. The number included well known Th17 cell signature proteins interleukin 17F (IL17F) and TF retinoic acid receptor–related orphan receptor C (RORC). The other proteins detected only in Th17 cells included aryl hydrocarbon receptor (AHR) and phosphodiesterase 5A (PDE5A) and solute carrier family 4 member 2 (SLC4A2, also known as AE2). These proteins are either active in Th17 cells associated with primary biliary cirrhosis and autoimmune disease of the urogenital tract.

Differential protein expression analysis was performed using limma method as described in Section 3.2 with FDR multiple hypothesis correction explicated in 3.3. Considering FDR<0.05 as threshold for statistically significance, we found 414 up-regulated and 591 down-regulated proteins in Th17 vs. Th0 comparison and 308 up-regulated and 367 down-regulated in iTreg vs. Th0 cell. The results are summarized in Figure 4.2.

T cell differentiation is driven mainly by TFs, provided that TCR is activated with its costimulatory receptors and proper cytokine stimuli are available.

Hence, we scrutinized this protein class more closely. Foxp3 known to govern Treg polarization were consistently detected in iTreg cultures and absent in Th0 cultures. Statistical analysis revealed changes in 75 and 50 transcription regulators and ligand-dependent nuclear receptors during Th17 and iTeg polarization, respectively. The numbers included both well–known TFs involved in the respective T cell subset specific differentiation, such as retinoic acid receptor–related orphan receptor C (RORC) and FOXP3. Moreover, TFs CCR4-NOT transcription complex subunit 2 (CNOT2) and family with sequence similarity 129 member B (FAM129B), were differentially expressed in both Th17 and iTreg cells. Their functions in Th17/iTreg cell differentiation are not well known. In addition to TFs, we focused on kinase proteins as they are both known to be involved in cell fate determination and targets of various medical treatments for inflammation related diseases and autoimmune disorders. We found expression changes in several kinases including mitogen-activated protein kinase 11 (MAPK11), right open reading frame kinase 1 (RIOK1), and cleavage and polyadenylation factor I subunit 1 (CLP1) not previously reported being associated with Th17 or iTreg cell function or differentiation.

In addition to examine individual DE proteins, we investigated biological processes involved in cell differentiation using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis as introduced in Section 3.3. The DE proteins from comparison between iTreg and Th0 cells were enriched in only systemic lupus erythematosus and alcoholism pathways mainly due to expression changes of several histone cluster 1 h2 and h4 family members. Th17 cell differentiation process were associated with ten pathways including antigen processing and metabolic changes, especially oxidative phosphorylation.

The results of the direct differential expression analysis between Th17 and iTreg cells are summarized in Figure 4.3 (a) and (b). In addition, (c) Transcriptional regulatory network of consisting 155 TFs and ligand-dependent nuclear receptors detected as DE proteins is shown.

4.4 The coherence of gene and protein expression changes

We evaluated the concordance between the trancriptomes and proteomes for Th17 cells. In order to obtain transcriptome for DGE analysis between Th17 and Th0 cells, we generated RNA-seq data compatible with the proteomics data using Illumina sequencing. The sequences were processed to read counts mapping to genes using tools and methods described in Section 2.2.1. DGE were extracted from the count data as discussed in Section 3.2.

Exploration of Th17 proteome and transcriptome showed that we were able to detect key molecules IL17F and RORC driving Th17 polarization at both level. Furthermore, 96.7% of the proteins detected by LC-MS/MS were also found in RNA-seq data. We excluded proteins with no corresponding transcripts from further inspection. In order to investigate which protein expression changes

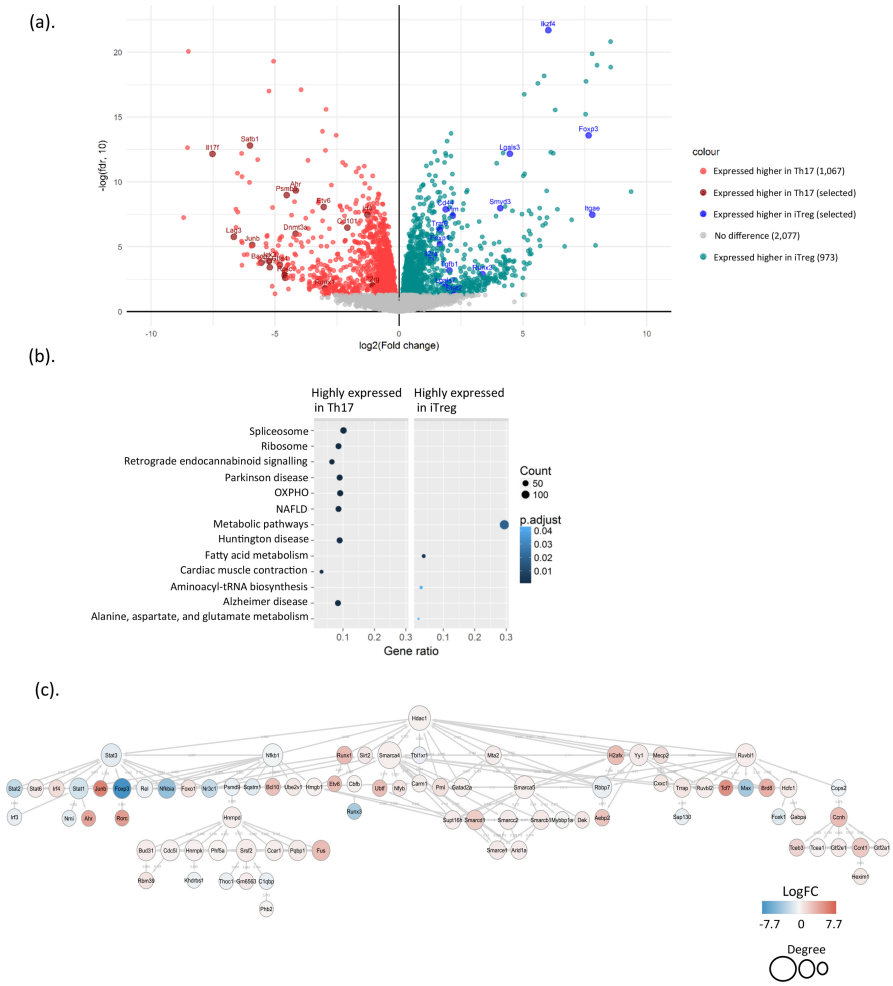

Figure 4.3. (a) There were 2,040 DE proteins from which 1,067 proteins were expressed higher in Th17 cells, and 973 higher in iTreg cells. From DE proteins known to steer polarization towards iTreg lineage, FOXP3, IKAROS family zinc finger 4 IKZF4 and RUNX3 found highly expressed iTreg cells, whereas chromatin organizer SATB1 were lowly expressed as it should when FOXP3 is highly expressed. Importantly, we found that SATB1 was highly expressed in Th17 cells compared to both TCR-activated Th0 cells and iTreg cells. When compared the differentiated cells to TCR–activated Th0 cells, we found that 20 proteins showed pattern of low expression in iTreg cell and high expression in Th17 cells. The opposite pattern was followed by 26 proteins. (b) Spliceosome, ribosome, and oxidative phosphorylation pathways were enriched in Th17 cells were while top enriched pathways in iTreg cells were metabolic pathway and fatty acid metabolism pathways. (c) In TF network, epigenetic regulators histone deacetylase 1 (HDAC1) and SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 4 (SMARCA4) emerged from the network due to their connectivity with cell type specific proteins and high-node degree (Figure from Publication I).

were not seen at mRNA level, we compared the DE genes and DE proteins. In total, there were 963 genes and proteins of which only 284 (29.5%) consistently up– or down–regulated at protein and mRNA levels. Interestingly, most of the DE proteins were differentially expressed at mRNA level. Discrepancy between

DE genes and DE proteins were not due to low mRNA expression as the gene expression levels of the DE genes that encode non-DE proteins and the DE genes that encode DE proteins were similarly distributed. This indicates translation, protein degradation and export process, and maybe post-translation modification are important in regulating Th17 cell protein expression.

We did similar investigation concerning iTreg lineage using published transcriptomics microarray data for iTreg and Thp cells. The microarray data was analyzed with gene-wise linear models using limma as detailed in Section 3.2. The driving TF Foxp3 was expressed both at protein and at mRNA level. The expression changes were consistent at protein and mRNA levels among 757 genes. However, most of the DE proteins (67.8%) had inconsistent RNA expression and 1,313 DE proteins were not differentially expressed at mRNA level. The number includes proteins not previously associated with iTreg such as stathmin 2 (STMN2), prolyl 4-hydroxylase subunit alpha 1 (P4HA1), and reactive oxygen species modulator 1 (ROMO1) as well as proteins such as H3K4 histone methyltransferases, SET and MYND domain containing 3 (SMYD3) whose contributions to iTreg cells have been well characterized.

We confirmed protein expression patterns of some selected proteins using targeted laboratory experiments and three additional cell cultures. Protein expressions of Th17 and iTreg cells were compared with protein expressions of Th0 cells. We validated, for example, that Enolase 3 (ENO3) was repressed in Th17 and induced in iTreg cells, Forkhead box O1 (FOXO1) was increased in both iTreg and Th17 cells, Nuclear factor of activated T cells 2 (NFATC2) was upregulated in Th17 cells, SMYD3 was upregulated in iTreg cells and a intermediate filament protein Vimentin (VIM) was highly expressed iTreg cells.

VIM was particularly interesting as it was highly expressed both at protein and at transcription level in differentiated iTreg cells and differentially expressed in the cells already after three days of polarization. By comparing cells derived from normal mice with cells from mice with no VIM gene under varying cytokine miliue, we found that TGFβ induces VIM expression in iTreg cells and it contributes to Foxp3 expression.

5. Transcription factor regulation on T cell biology

In this chapter, I cover the main findings of Publication II and Publication III focusing mainly on the co–localization of SNPs and TF binding sites (or any set of genomic regions).

The first publication concerns the molecular mechanism of human Th17 cell differentiation. The focus is on direct and indirect targets of Signal transducer and activator of transcription 3 (STAT3) during early phases of the cell differentiation. The work is intergrative by involving new high-throughput experiments and integrating pre-existing data to form a comprehensive view of the subject. Corcerning disease associated single nucleotide polymorphisms (SNP) enrichment on specific genomic regions, we made a hypothesis that using the proxies of the tag-SNPs enhances the sensitivity of the analysis. The second publication generalizes the statistical approach further, represents the computational tool we developed for this kind of analyses, and demonstrates its function. The two use cases of this approach shows that the hypothesis holds. In this chapter, I first highlight the main findings of the first publication and then I give an overview of the statistical approach and the software tool described in the second publication.

RNA interference (RNAi) and RNA-seq techniques were used to verify that STAT3 is indeed an important regulator of Th17 cell differentiation. The silencing experiment showed that inhibition of STAT3 decreased secretion of signature cytokine IL-17 and reduced expression of CCR6, a chemokine receptor highly expressed in Th17 cells. Together, they imply that STAT3 is important in the cell polarization. Next, DGE analysis of polarizing cells with respect to Th0 revealed 2200 and 1500 differentially expressed genes at two hour and 72 hour time points of the initiation of cell differentiation. Moreover, direct DGE screen between STAT3 interfered and ordinary polarizing Th17 cells showed differential expression in all measured time points and differences were highest at the final 72 hour time point. Proportion of STAT3 regulated genes in differentially expressed genes increased during differentiation while the number of the upregulated and the number of the downregulated genes were approximately the remained the same at every time points. This indicates that STAT3 both promotes Th17 cell polarization and prevents the ability of the cells to differ-

entiate to alternative lineages. The conclusion is that STAT3 is an important transcriptional regulator of Th17 cell differentiation. Naturally, DGE studies produced lists of STAT3 regulated genes.

STAT3 chromatin immunoprecipitation sequencing (ChIP-seq) study was performed distinguish which of the genes affected by STAT3 are its direct and indirect targets. The investigation was done for the cells at half an hour time point after the induction of Th17 cell polarization. There were almost 3000 STAT3-binding sites. A fifth of the binding sites were located in the immediate promoter regions. However, two thirds of the sites were located in introns or intergenic regions. The genome types reflect possible regulation mechanism of TF. Former suggests STAT3 regulates gene expression via binding to core promoters of its targets while latter sings to gene expression regulation via binding to distal regulatory elements. Nearly half of the binding sites were localized within 10 kb up– or down–stream of transcription start sites (TSS). STAT3 target genes were classified into four categories. First group contained direct STAT3 targets. Their expression was changed during Th17 cell differentiation with respect to Th0 cells, were found to be regulated by STAT3 in silencing experiments, and had STAT3 binding was detected in ChIP-seq experiments. The second group consisted of indirect STAT3 targets. Their expression was changed during Th17 cell differentiation, were regulated, but not bound by STAT3. Lack of STAT3 binding indicates that STAT3 influences the genes via indirect mechanisms. The third group is called putative targets. Genes in the group were differentially expressed during Th17 cell differentiation, and bound by STAT3, but not regulated. The fourth group contained TFs that were differentially expressed in polarizing Th17 cells, but not regulated or bound by STAT3. This combination of indications suggests that the genes express independently from STAT3.

5.1 STAT3 regulated cell differentiation and the impact of the related SNPs

As Th17 cell are known to be involved in many autoimmune diseases and many SNPs are associated with many diseases (even when located far away from the gene), we studied the possible enrichment of SNP associated with human diseases in genomic regions that where STAT3 target genes or STAT3-binding sites. In first investigation, STAT3–regulated genes were linked to SNPs near their transcription starts sites (TSSs) that were associated to a traits in NHGRI GWAS catalog. Along significantly enriched traits, there were 11 autoimmune diseases. They were crutinized more carefully using method described in section 5.2. There were no significant enrichment in any of the autoimmune–disease–associated SNPs. Interestingly, when diseases were combined and treated as a single trait statistical signal improved a little, but FDR was still 0.11. The next section focuses on the method used to identify co-localizations of SNPs and genomic regions.

5.2 Detecting statistically signifigant SNPs in genomic regions

This section focuses on a method to decipher whether SNPs associated to some phenotype are significantly co–localized with given genomic regions or not. This kind of study is possible because of public catalogues of indivual genomic variation within various population as well associations between SNPs and various biological traits are available. The basic workflow is simple. one fetched tag–SNPs for disease from a database, create sets of corresponding SNPs with similar genomic attributes, and base the co–localization analysis on the empirical statistical method described in section 3.1. In this case, the natural selection for test statistics t is the overlap of the SNPs and the genomic regions and empirical p–value is derived by applying Equation 3.5. We developed this method further Publication II by taking into account that obtained SNPs are affected by linkage disequilibrium (LD) as shown in figure 5.1. We noted that considering all expanding the sets of tag–SNPs with proxy SNPs, i.e. SNPs in LD, improves the sensitivity of the co–localization analysis.

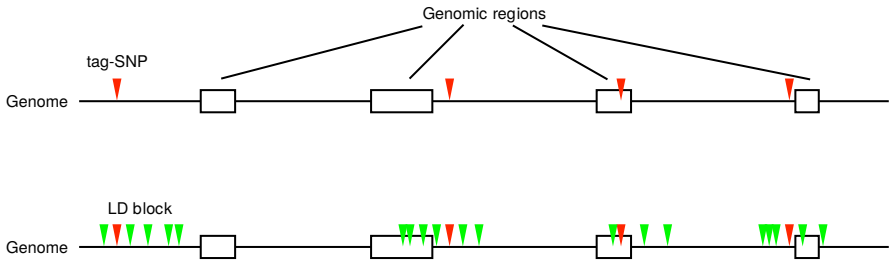

Figure 5.1. In the upper figure, the red wegdes and rectangles highlight the tag–SNPs of a disease derived from GWAS and the binding sites of an interesting TF in the genome, respectively. In this example, only one of the tag–SNPs overlapped with the genomic regions. Below, green wedges show the loci of the proxy SNPs. Now, as tag–SNPs and their proxies combined overlap more with the genomic regions, test statistics t captures more these previously hidden co–localizations.

In Publication III, we represented publicly available sotfware for co-localization analysis. The tools was developed in the form of R package it integrates pre-existing tools, like whole genome association analysis toolset plink and webtool SNPsnap, enabling custom SNP co-localization analysis in a convenient and efficient way. The packages contains functionalities to fetch data from NHGRI-EBI GWAS Catalog, to cleanup SNPs from duplicate SNPs from the same LD, adding all SNPs in the LD to a list of SNPs, connection to SNPsnap server in order to make a retrieval request to generate a specified number of SNP sets matching the query SNP sets, expanding a list of SNP with all their proxies, computing test statistics, empirical p-values and multiple testing corrections.

In Publication III, we represented a test case where we elucidate the functionality the software. In addition, we assessed the importance and feasibility of considering proxy SNPs in the analysis. The test case was performed using publicly available data of STAT6 binding sites in Th2 cells. As Th2 cells are part of adaptive immunity, we tested whether we could find significant co–localization

of the STAT6 binding sites and the SNPs associated to the same 11 immunity-related diseases we studied in Publication II. Moreover, we added three traits with no known association to immunity to the study. We performed the analysis in two ways. First, was done using the tag-SNPs only and second utilized the tag-SNPs together with their proxies. With tag-SNPs, we found significant co-localization in two diseases which were both immune-related. Addition of proxies resulted in detection five more immune-related diseases as significantly co-localized. Importantly, inclusion of proxies did not neither diminish signal of significance in any of the diseases nor indicated that all diseases with some overlap would become artificially significant just by including the proxies. Overall, we conclude that the software facilitates the co-localization analysis and it is feasible to use proxies in the analysis.

6. Mechanistic modeling of transcriptional regulation

This chapters covers the main results and the methodology used in publications Publication IV and Publication V.

6.1 Representations of subcellular system

A subcellular system is visualized either as a cartoon to emphasize biological details to be modeled or as a graph to abstract the representation from them. A graph consists of nodes that denote the molecules and directed edges that express interactions between the nodes. By convention, there is no heading at the beginning of the edge while at the end there is either an arrow of a flat head. An arrow head means the activation of the node at the end of the edge and the consumption of the node at the beginning of the edge. An edge that does not point to any node indicates degradation of the attached molecule and an edge that does not begin from a node implies basal activation that does not depend on any. The flat headed edge means that the node at the beginning inhibits the node at the end. Especially, when the modeled system is large, edges denoting basal activation and degradation processes are not shown or shown with faded color in the graph in order to maintain larger graphs visually pleasing and tractable. Figure 6.1 shows both cartoon and graph representations of a dynamically evolving molecular system. The graphs are converted into systems of ordinary differential equations of the form Equation 3.9 (or Equation 3.11 in the case of dynamically evolving system) where variables corresponds to the nodes and terms to the edges of the network.

The studies consider systems from which detailed knowledge of the underlying molecular mechanisms is not available. The objective was to find likely models from a set of models that are composed of molecules known to be involved in the cellular processes and mechanistically possible interactions between the molecules. Hence, instead of using enzymatically motivated kinetics in the ODEs, we assume that each term in every ODE in the systems is a product of impacting molecules and associated parameter. Formally, for a system with N

variables each ODE is assumed to be of the form

$$\frac{dy_i}{dt} = k_i^{\text{basal}} + \sum_{j=1}^{N} k_{ij}^{\text{independent activation}} y_k + \sum_{\substack{j,i=1 \\ j<k}}^{N} k_{ijk}^{\text{synergistic activation}} y_j y_k + \sum_{j=1}^{N} k_{ij}^{\text{inhibition}} y_j y_i + k_i^{\text{degradation}} y_i.$$

where the name of the parameter indicates influencing variables and type of action. In Publication V, inhibitory terms are not included in the models as the focus is purely on enhancer activation signature.

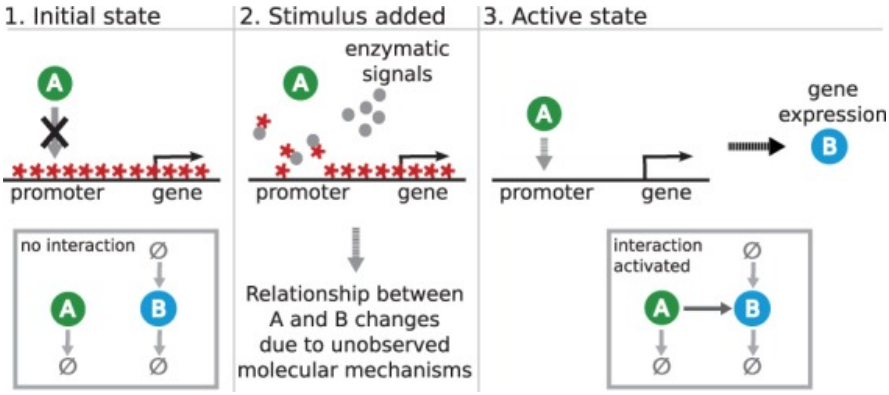

Figure 6.1. The upper panel consists of the cartoons that show molecular actions occurring in the system at three occasions. The lower panel comprises graphs depicting the network structures corresponding to the system before and after it is altered by a hidden enzymatic process. As the network is not static, this kind of system is called dynamically evolving (Figure from Publication IV).

6.2 Model selection for latent effect mechanistic (LEM) model

Publication IV hypothesized that incorporating non-static network structure enables the model capture the impact of transient phenomena impact. Hence it introduced LEM models that adapt systems ODEs to consider switches in the interactions between the molecules, and associated model inference framework.

We applied this approach to infer models for a simulated system and the core Th17 regulatory network. The simulated system consisted of four genes

that interacted differently during three gradually shifting phases. This was consistent with previously reported idea on how idea that Th17 cells differentiate. We created data using fixed parameters and were able to infer correct model successfully. The forward-backward-stepwise selection procedure converged to the correct LEM model, and the inferred parameter values and the model responses were close to the values used to simulate the data.

The core Th17 regulatory network consisted of TFs the retinoic acid receptor-related orphan receptor gamma t (RORC), signal transducer and activator of transcription 3 (STAT3), basic leucine zipper transcription factor (BATF), transcription factor Maf (MAF) and interferon regulatory factor 4 (IRF4). Only experimentally validated interactions between the genes were considered in the models which were calibrated to RNA-seq time-series data. Inference converged to a model which supported interactions experimentally validated with different measurement techniques and experimental conditions. Thus, LEM modeling can used to predict the system incorporating sequentially activated subnetworks. Three transcriptional phases - early induction, intermediate onset of phenotype and amplification, and late stabilization - took place about 0-4 hours, 4-13 hours and 13-72 hours postinduction of differentiation. They all have apparent interpretations. First, cell differentiation is initiated and the cytokine milieu prepares cells for lineage commitment. Next, MAF and IRF4 support differentiation and RORC is activated. Finally, RORC and STAT3 maintain differentiation supported by BATF and IRF4. Especially interesting is that the model highlights how important STAT3 is for Th17 differentiation during two first phases. Moreover, the inferred LEM model fitted to the RNA-seq time-series data much better than the static model embodying all interactions without dynamic rewirings of the network structures. This is in accordance with the hypothesis. Parameter identifiability was analyzed with profile likelihood estimates as described in section 3.4. Most of the parameters were identifiable, there were no structurally unidentifiable parameters, and four were parameters practically unidentifiable. We speculated that unidentifiablity may be removed by reparametrizing the model, but such effort was out of the scope of this study. Altogether, the developed methodology performed well in both simulated setting as well as Th17 differentiation.

6.3 Modeling enhancer activation

Publication V focuses on activation of cis-regulary elements during cell differentiation. Enhancers steer the differentiation by forming three dimensional loops in the chomatin structure leading into complexes of transcription activation molecules and hence impacting the transcriptomics of the cells. Active enhancers are known to associate with post-translational chromatin modifications and binding of context specific transcription factors. In addition to such observable key enhancer activation molecules, the dynamical systems governing

enhancer activity involves other molecules which are either unknown or more difficult to measure. In this work, we modeled enhancer activity as an interplay of the best known enhancer activity signals using systems ODEs. In the context of T cell differentation, the systems include three activation signals. Between them, we considered four simple interactions types: 1) The signals are independent, 2) the signals form cascades, 3) two molecules cause the activity of the signal synergistically, or 4) additively. Figure 6.2 depicts effectively different models.

Our hypothesis was the selection of the correct model is achievable with relatively small amount of experimental data. Hence, we introduced a Bayesianly motivated statistical framework to link the data to the models. The construction was used to survey potentiality of this approach for experimental biology from two perspectives. Is this framework able to set a part the true model from other models and are the models identifiable in term of theirs parameters. We simulated data from one model of each model family. We introduced basic variability in the dynamics by using parameters drawn from normal distribution with fixed mean and five percent coefficient of variation. Aiming to characterize amount of data that are feasible to obtain experimentally we created three different scenarios. There were measurements from time points 0, 4 and 72 hours, 0, 4, 8, and 72 hours or 0, 4, 8, 12 and 72 hours reflecting times scale important cellular changes are known to occur in early T cell differentiation. Simulated data were subjected to increasing level of measurement noise standard deviation were 0.15, 0.25, 0.5, 0.75, 1, 1.25 1.5 or 2.0. Execution of each setting 50 times independently, led to 4800 different independently created data sets. Model selection were done either in deliberately liberal way in the sense that in addition to rate parameters the initial values and measurement noise were inferred from the data or more conservative way that only rate parameters were considered free while the initial values and measurement noise were fixed. In both scenarios, five time point were sufficient to distinguish the correct model given that measurement noise were reasonable with respect to data values (standard deviation was less than 1). Synergistic model were notable more difficult to infer than others this limited amount of data as its partly structurally identifiable parameters. Overall, the framework is able to detect correct interactions between enhancer activation signals with five time points which fits well with the hypothesis.

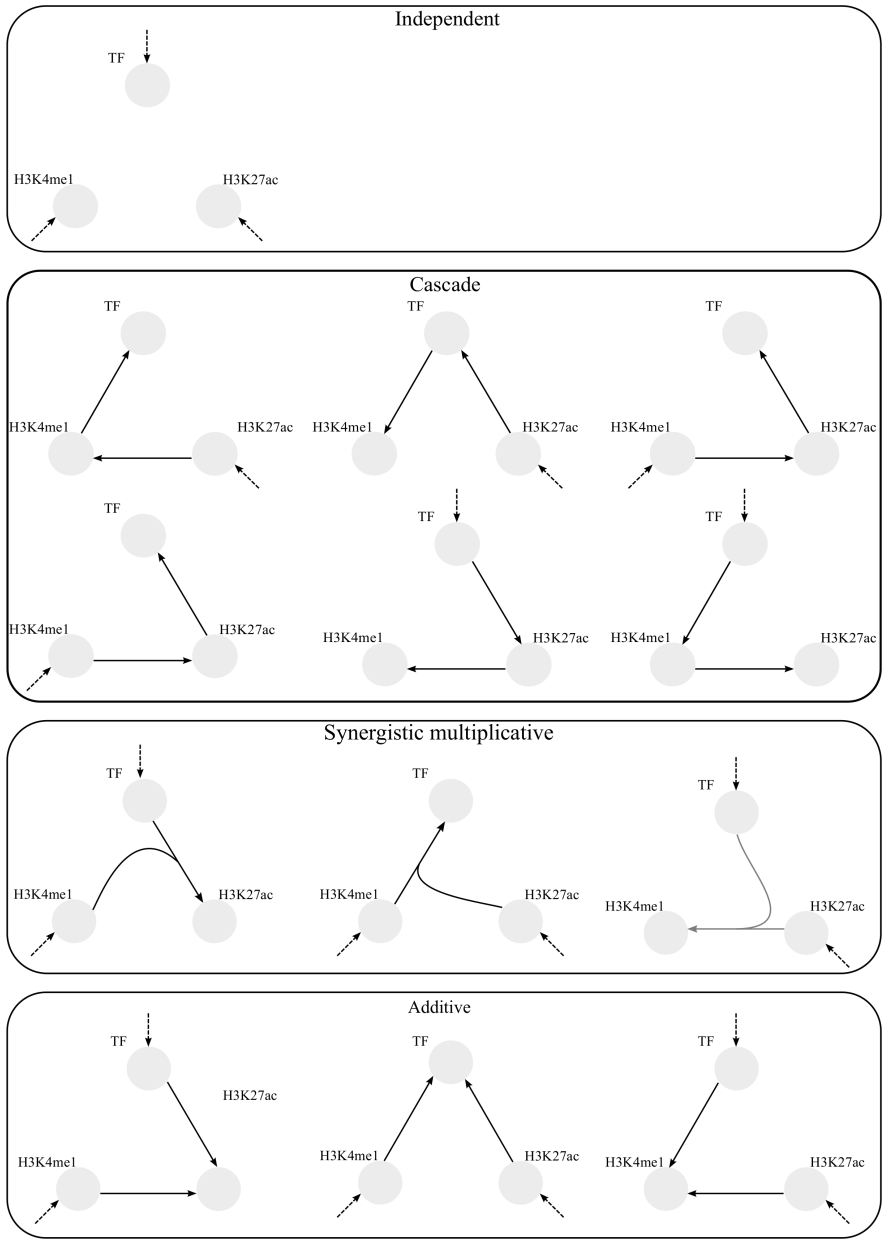

Figure 6.2. There are 13 effectively different enhancer activation models. Activation impacts are shown in solid lines. For simplicity, deactivation each component is not shown.

7. Conclusions

The work represented in this thesis focused on some proteomic, transcriptomic, genomic and epigenetic markers important in progressing cell differentiation. The studies have dual nature. They apply existing bioinformatics methods and tools to map systems levels molecular and functional progressions of the cells. Then again, new analysis methods, modeling frameworks and associated computational tools are developed in order to achieve new biological knowledge.

The proteomic analysis of Th17 and iTreg cells resulted in a catalogue of proteins potentially involved in autoimmune disease in mice as well as an outline of a data analysis pipeline. Both are immediately applicable to further studies in the developing field of proteomics. An obvious further direction could be conducting a similar study focusing on human cells. In addition, the study showed some discrepancies between trancriptomics and proteomis of the cells as well as proposed a functional network based on protein-protein interactions and GSEA analysis. These can be interesting targets for further research as well.

Transcriptional control of Th17 cell differentiation embodies SNPs impact on binding of the key transcription factors. This called for a method for revealing statistical significance of co-localized sets of SNPs and TF binding sites. We developed such method emphasizing importance of respecting LD blocks of the SNPs, tested the method with independent public data and a published tool implementing the method R package snpEnrichR.

Mechanism of transcriptional regulation of Th17 cells were modeled deterministically at two levels using systems of ODEs. First level integrated seemingly conflicting published transcriptomics data and incorporated time-dependent latent effects into the equations able to change the impact of variables during differentiation. The inferred transcription regulatory network fitted the data much better than models without latent effects. Moreover, most of the parameters were well identifiable. The second level focused on determining data requirements for modeling of active cis-regulatory elements. Data from five time points sampled from dynamically active time frame may be sufficient to fix parameters when data comes from a model describing the dynamics well. In such case, the proposed statistical framework is able rank the generative model high.

Together, the research in this dissertation illuminates cellular systems involved in Th17 cell differentiation from many perspectives as well as provides computational methods and tools for further or similar studies on other organisms or other cell types.

References

Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198, 2003.

Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347, 2016.

Yasmeen Ahmad and Angus I Lamond. A perspective on proteomics in cell biology. *Trends in cell biology*, 24(4):257–264, 2014.

Bruce Alberts, Alexandrt Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4 edition, 2002. ISBN 0-8153-4072-9.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, 8(9):1765, 2013.

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

Simon Andrews. Fastqc: a quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`, 2010. [Online; accessed 08-January-2019].

A Annunziato. Dna packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26, 2008.

Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389(4):1017–1031, 2007.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Estelle Bettelli, Yijun Carrier, Wenda Gao, Thomas Korn, Terry B Strom, Mohamed Oukka, Howard L Weiner, and Vijay K Kuchroo. Reciprocal developmental pathways for the generation of pathogenic effector t h 17 and regulatory t cells. *Nature*, 441 (7090):235, 2006.

Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, et al. An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471, 2013.

Ewan Birney and Nicole Soranzo. Human genomics: The end of the start for population sequencing. *Nature*, 526(7571):52, 2015.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Elizabeth M Blackwood and James T Kadonaga. Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63, 1998.

Clemens Bönisch and Sandra B Hake. Histone h2a variants in nucleosomes and chromatin: more or less stable? *Nucleic acids research*, 40(21):10719–10741, 2012.

Michal Breker and Maya Schuldiner. The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nature Reviews Molecular Cell Biology*, 15(7):453, 2014.

Anthony J Brookes. The essence of snps. *Gene*, 234(2):177–186, 1999.

Michael Burrows and David J Wheeler. A block-sorting lossless data compression algorithm. 1994.

Jennifer EF Butler and James T Kadonaga. The rna polymerase ii core promoter: a key component in the regulation of gene expression. *Genes & development*, 16(20): 2583–2592, 2002.

Michal Chorev and Liran Carmel. The function of introns. *Frontiers in genetics*, 3, 2012.

Chunaram Choudhary and Matthias Mann. Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews Molecular cell biology*, 11(6):427, 2010.

Clary B Clish. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies*, 1(1):a000588, 2015.

Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

A Corthay. A three-cell model for activation of naive t helper cells. *Scandinavian journal of immunology*, 64(2):93–96, 2006.

Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular & cellular proteomics : MCP*, 13:2513–26, 2014.

Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

Francis HC Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

Anthony Christopher Davison, David Victor Hinkley, et al. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

Mauro DeSanctis and Pierpaolo Loreti. Data mining of the human being. *Human Bond Communication: The Holy Grail of Holistic Communication and Immersive Experience*, pages 59–70, 2017.

Ira W Deveson, Marion E Brunck, James Blackburn, Elizabeth Tseng, Ting Hon, Tyson A Clark, Michael B Clark, Joanna Crawford, Marcel E Dinger, Lars K Nielsen, et al. Universal alternative splicing of noncoding exons. *Cell systems*, 6(2):245–255, 2018.

Sorin Draghici. *Data analysis tools for DNA microarrays*. Chapman and Hall/CRC, Boca Raton, 2003. ISBN 1-58488-315-4.

Michel DuPage and Jeffrey A Bluestone. Harnessing the plasticity of cd4+ t cells to treat immune-mediated disease. *Nature Reviews Immunology*, 16(3):149, 2016.

ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8 (3):175–185, 1998.

Nuno A. Fonseca, Johan Rung, Alvis Brazma, and John C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012. doi: 10.1093/bioinformatics/bts605. URL http://dx.doi.org/10.1093/bioinformatics/bts605.

Fabian Fröhlich, Barbara Kaltenbacher, Fabian J Theis, and Jan Hasenauer. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS computational biology*, 13(1):e1005331, 2017.

Alexandre Gaspar-Maia, Adi Alajem, Eran Meshorer, and Miguel Ramalho-Santos. Open chromatin in pluripotency and reprogramming. *Nature reviews Molecular cell biology*, 12(1):36, 2011.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

Mark B Gerstein, Can Bruce, Joel S Rozowsky, Deyou Zheng, Jiang Du, Jan O Korbel, Olof Emanuelsson, Zhengdong D Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-encode? history and updated definition. *Genome research*, 17(6): 669–681, 2007.

Nick Gilbert, Susan Gilchrist, and Wendy A Bickmore. Chromatin organization in the mammalian nucleus. *International review of cytology*, 242:283–336, 2005.

Walter Gilbert. Why genes in pieces? *Nature*, 271(5645):501, 1978.

Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6): 333, 2016.

Anthony JF Griffiths, Jeffrey H Miller, Suzuki David T, Richard C Lewontin, and William M Gelbart. *An Introduction to Genetic Analysis*, volume 4 of *10*. W. H. Freeman, New York, 7 edition, 2000. ISBN 0-7167-3520-2. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21766/.

Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311, 2007.

Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108, 2009.

Sven Heinz, Casey E Romanoski, Christopher Benner, and Christopher K Glass. The selection and function of cell type-specific enhancers. *Nature reviews Molecular cell biology*, 16(3):144, 2015.

Michael J Heller. Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153, 2002.

Alan C Hindmarsh, Peter N Brown, Keith E Grant, Steven L Lee, Radu Serban, Dan E Shumaker, and Carol S Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)*, 31(3): 363–396, 2005.

Joseph L Hodges and Lucien Le Cam. The poisson approximation to the poisson binomial distribution. *The Annals of Mathematical Statistics*, 31(3):737–740, 1960.

Jörg D Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics*, 7(3):200, 2006.

Fabian Hosp and Matthias Mann. A primer on concepts and applications of proteomics in neuroscience. *Neuron*, 96(3):558–571, 2017.

David C Hoyle, Magnus Rattray, Ray Jupp, and Andrew Brass. Making sense of microarray data distributions. *Bioinformatics*, 18(4):576–584, 2002.

Fan Hsu, W James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler. The ucsc known genes. *Bioinformatics*, 22(9):1036–1046, 2006.

Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2008.

Huei-Chung Huang, Yi Niu, and Li-Xuan Qin. Differential expression analysis for rna-seq: An overview of statistical methods and computational software: Supplementary issue: Sequencing platform modeling and analysis. *Cancer informatics*, 14:CIN–S21631, 2015.

Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H Badger, Asif T Chinwalla, Heather H Creasy, Ashlee M Earl, Michael G FitzGerald, Robert S Fulton, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207, 2012.

Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

Inc. Illumina. Hiseq™ 2000 sequencing systemredefining the trajectory of sequencing. `https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf`, 2010. Online; accessed: 2021-09-30.

Charles A Janeway, Paul Travers, Mark Walport, and Mark J Shlomchik. Immunobiology: The immune system in health and disease. *New York*, 2001.

Kai Kammers, Robert N Cole, Calvin Tiengwe, and Ingo Ruczinski. Detecting significant changes in protein abundance. *EuPA open proteomics*, 7:11–19, 2015.

M Karin. Too many transcription factors: positive and negative interactions. *The New Biologist*, 2(2):126–131, 1990.

Roshan Karki, Deep Pandya, Robert C Elston, and Cristiano Ferlini. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC medical genomics*, 8(1):37, 2015.

V Narry Kim, Jinju Han, and Mikiko C Siomi. Biogenesis of small rnas in animals. *Nature reviews Molecular cell biology*, 10(2):126, 2009.

Thomas Korn, Estelle Bettelli, Mohamed Oukka, and Vijay K Kuchroo. Il-17 and th17 cells. *Annual review of immunology*, 27:485–517, 2009.

Alberto R Kornblihtt, Ignacio E Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews Molecular cell biology*, 14(3):153, 2013.

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

Jouni Kuha. Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229, 2004.

Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

Mark Larance and Angus I Lamond. Multidimensional proteomics for cell biology. *Nature reviews Molecular cell biology*, 16(5):269, 2015.

David S Latchman. Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312, 1997.

Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2): R29, 2014.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Ruijie Liu, Aliaksei Z Holik, Shian Su, Natasha Jansz, Kelan Chen, Huei San Leong, Marnie E Blewitt, Marie-Liesse Asselin-Labat, Gordon K Smyth, and Matthew E Ritchie. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic acids research*, 43(15):e97–e97, 2015.

Harvey Lodish, Arnold Berk, SL Zipursky, P Matsudaira, D Baltimore, and J Darnell. *Molecular cell biology*. W. H. Freeman, New York, 4 edition, 2000. ISBN 0-7167-3136-3. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK21475/`.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

Rohan Lowe, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. Transcriptomics technologies. *PLoS computational biology*, 13(5):e1005457, 2017.

Elaine R Mardis. Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303, 2013.

John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 2008.

Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

John S Mattick and Igor V Makunin. Non-coding rna. *Human molecular genetics*, 15 (suppl_1):R17–R29, 2006.

Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

P McCullagh and JA Nelder. Generalized linear models: monographs on statistics and applied probability. vol. 37, 1989.

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

Timothy R Mosmann, Holly Cherwinski, Martha W Bond, Martin A Giedlin, and Robert L Coffman. Two types of murine helper t cell clone. i. definition according to profiles of lymphokine activities and secreted proteins. *The Journal of immunology*, 136(7):2348–2357, 1986.

Mouse Genome Sequencing Consortium et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520, 2002.

Paul Muir, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J Spakowicz, Leonidas Salichos, Jing Zhang, George M Weinstock, Farren Isaacs, Joel Rozowsky, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17(1):53, 2016.

Donald E Olins and Ada L Olins. Chromatin history: our view from the bridge. *Nature reviews Molecular cell biology*, 4(10):809, 2003.

Art B Owen. A central limit theorem for latin hypercube sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(2):541–551, 1992.

Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87, 2011.

Helen Pearson. Genetics: what is a gene? *Nature*, 441:398–401, 2006.

Len A Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4):288, 2013.

Petter Portin and Petter Wilkins. The evolving definition of the term "gene". *Genetics*, 205(4):1353–1364, April 2017. doi: 10.1534/genetics.116.196956.

Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A Brugmann, Ryan A Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279, 2011.

Andreas Raue, Bernhard Steiert, Max Schelker, Clemens Kreutz, Tim Maiwald, Helge Hass, Joep Vanlier, C Tönsing, L Adlung, R Engesser, et al. Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, 2015.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

Mark D. Robinson and Gordon K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008. doi: 10.1093/biostatistics/kxm030. URL `http://dx.doi.org/10.1093/biostatistics/kxm030`.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Abbie Saunders, Leighton J Core, and John T Lis. Breaking barriers to transcription elongation. *Nature reviews Molecular cell biology*, 7(8):557, 2006.

John P Savaryn, Timothy K Toby, and Neil L Kelleher. A researcher's guide to mass spectrometry-based proteomics. *Proteomics*, 16(18):2435–2443, 2016.

Virgil Schijns, Dragomira Majhen, Peter Van Der Ley, Aneesh Thakur, Artur Summerfield, Rita Berisio, Cristina Nativi, Alberto Fernández-Tejada, Carmen Alvarez-Dominguez, Sveinbjörn Gizurarson, et al. Rational vaccine design in times of emerging diseases: The critical choices of immunological correlates of protection, vaccine antigen and immunomodulation. *Pharmaceutics*, 13(4):501, 2021.

Raewyn M Seaberg and Derek van der Kooy. Stem and progenitor cells: the premature desertion of rigorous definitions. *Trends in neurosciences*, 26(3):125–131, 2003.

Radu Serban and Alan C Hindmarsh. Cvodes: An ode solver with sensitivity analysis capabilities. Technical report, Technical Report UCRL-JP-200039, Lawrence Livermore National Laboratory, 2003.

Simone Sidoli, Katarzyna Kulej, and Benjamin A Garcia. Why proteomics is not the new genomics and the future of mass spectrometry in cell biology. *J Cell Biol*, pages jcb–201612010, 2016.

Robert J Sims, Rimma Belotserkovskaya, and Danny Reinberg. Elongation by rna polymerase ii: the short and long of it. *Genes & development*, 18(20):2437–2468, 2004.

Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477, 2008.

Christopher WJ Smith, James G Patton, and Bernardo Nadal-Ginard. Alternative splicing in the control of gene expression. *Annual review of genetics*, 23(1):527–577, 1989.

Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613, 2012.

Paul Stapor, Daniel Weindl, Benjamin Ballnus, Sabine Hug, Carolin Loos, Anna Fiedler, Sabrina Krause, Sabrina Hroß, Fabian Fröhlich, and Jan Hasenauer. Pesto: Parameter estimation toolbox. *Bioinformatics*, 34(4):705–707, 2017.

Brian D Strahl and C David Allis. The language of covalent histone modifications. *Nature*, 403(6765):41, 2000.

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

Saeed Tavazoie, Jason D Hughes, Michael J Campbell, Raymond J Cho, and George M Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281, 1999.

Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

Timothy K Toby, Luca Fornelli, and Neil L Kelleher. Progress in top-down proteomics and the analysis of proteoforms. *Annual review of analytical chemistry*, 9:499–519, 2016.

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

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562, 2012.

Bryan M Turner. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nature structural & molecular biology*, 12(2):110, 2005.

Stefka Tyanova, Tikira Temu, and Jürgen Cox. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, 11(12):2301, 2016.

Victor E Velculescu, Lin Zhang, Wei Zhou, Jacob Vogelstein, Munira A Basrai, Douglas E Bassett, Phil Hieter, Bert Vogelstein, and Kenneth W Kinzler. Characterization of the yeast transcriptome. *Cell*, 88(2):243–251, 1997.

J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

Vladislav Vyshemirsky and Mark A Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2007.

J D Watson and F H Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1): 59, 2012.

Gabriel E Zentner and Steven Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259, 2013.

Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2018.

Yaoyang Zhang, Bryan R Fonslow, Bing Shan, Moon-Chang Baek, and John R Yates III. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4):2343–2394, 2013.

Jinfang Zhu, Hidehiro Yamane, and William E Paul. Differentiation of effector cd4 t cell populations. *Annual review of immunology*, 28:445–489, 2010.

Aalto-DT 197/2022

ISBN 978-952-64-1078-4 (printed)
ISBN 978-952-64-1079-1 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES