Department of Information and Computer Science

Computational Methods for Analysis of Dynamic Transcriptome and Its Regulation Through Chromatin Remodeling and Intracellular Signaling

Tarmo Äijö

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 149/2014

Computational Methods for Analysis of Dynamic Transcriptome and Its Regulation Through Chromatin Remodeling and Intracellular Signaling

Tarmo Äijö

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 AS1 of the school on 31 October 2014 at 12 noon.

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Prof. Harri Lähdesmäki

Thesis advisor
Prof. Harri Lähdesmäki

Preliminary examiners
Prof. Alexander J. Hartemink, Duke University, USA
Prof. Liisa Holm, University of Helsinki, Finland

Opponent
Prof. Richard A. Bonneau, New York University, USA

Aalto University publication series
DOCTORAL DISSERTATIONS 149/2014

ISBN 978-952-60-5884-9
ISBN 978-952-60-5885-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5885-6

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Tarmo Äijö

Name of the doctoral dissertation
Computational Methods for Analysis of Dynamic Transcriptome and Its Regulation Through Chromatin Remodeling and Intracellular Signaling

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 149/2014

Field of research Information and Computer Science

Manuscript submitted 10 June 2014 **Date of the defence** 31 October 2014

Permission to publish granted (date) 9 September 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Transcription is the first step in gene expression in which genetic information is transferred from DNA to RNA. Gene expression is highly controlled through transcriptional regulation at many steps. Transcriptional regulation in eukaryotes occurs, e.g., through binding of transcription factors and chromatin remodeling via various epigenetic pathways. Additionally, dysregulated transcription has been reported in various diseases. Thus, transcription and transcriptional regulation are of great interest for research.

In this work, we study the transcriptome and its regulation using bioinformatic and computational biology approaches. We propose computational methods, LIGAP and DyNB, for analysis of temporal gene expression profiles measured using microarrays and RNA-seq, respectively. LIGAP is a methodology based on Gaussian processes for simultaneous differential expression analysis between an arbitratory number of time series microarray data sets. DyNB, is an extension of the Gaussian-Cox process in which the Poisson distribution is replaced by the negative binomial distribution. Additionally, DyNB enables the study of systematic differences, such as differential differentiation efficiencies, between conditions. Sorad, is a modeling framework based on differential equations and Gaussian processes for analysis of intracellular signaling transduction through phosphoprotein activities. We also propose and demonstrate how the *in silico* models inferred using Sorad can be used in estimating modulation strategies to obtain desired signaling response. Finally, we study the determinants of nucleosome positioning and subsequent effects on gene expression. All the proposed methods are benchmarked against existing methods and, in addition, they are applied to real-life problems. The comparison studies validate the applicability of the presented methods and demonstrate their improved performance relative to existing methods. Our transcriptome studies led to increased knowledge on the early differentiation of human T cells, and provided a valuable resource of candidate genes for future functional studies of the differentiation process. Our nucleosome study revealed that within loci important for T cell differentiation only 6% of the nucleosomes are differentially remodelled between T helper 1 and 2 cells and cytotoxic T lymphocytes. The remodelled nucleosomes correlated with the known differentiation program, chromatin accessibility, transcription factor binding, and gene expression. Finally, our data supports the hypothesis that transcription factors and nucleosomes compete for DNA occupancy.

Keywords Bioinformatics, computational biology, gene expression, transcriptional regulation, Gaussian processes

ISBN (printed) 978-952-60-5884-9 **ISBN (pdf)** 978-952-60-5885-6

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 174 **urn** http://urn.fi/URN:ISBN:978-952-60-5885-6

A? Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Tarmo Äijö

Väitöskirjan nimi
Transkription, kromatiinin ja soluviestinnän vuorovaikutusten analysointi tilastollisilla menetelmillä

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 149/2014

Tutkimusala Tietojenkäsittelytiede

Käsikirjoituksen pvm 10.06.2014 **Väitöspäivä** 31.10.2014

Julkaisuluvan myöntämispäivä 09.09.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Geenin transkriptiossa kopioidaan DNA:ssa olevaa geneettistä koodia, joka johtaa geenien ilmentymiseen. Geenien ilmentymiseen johtava transkriptioaskel on tarkasti säädelty biologinen tapahtuma. Transkriptiota eukaryoottisoluissa säädellään muun muassa transkriptiotekijöiden sitoumisen promoottori- ja tehostaja-alueille ja epigeneettisten tekijöiden kautta. Geenien transkription säätelyn parempi ymmärtäminen on tärkeää, koska esimerkiksi transkription virheellinen säätely voi johtaa erilaisiin sairauksiin.
Tämän väitöskirjan artikkeleissa on kehitetty laskennallisia menetelmiä geenien ilmentymisen ja ilmentymisen säätelyn tarkempaan tutkimiseen. LIGAP ja DyNB ovat gaussisiin prosesseihin perustuvia menetelmiä mikrosiruilla tai RNA-sekvenssoinnilla mitattujen aikasarja-aineistojen analysointiin. LIGAP-menetelmä soveltuu geenien ilmentymiserojen havainnointiin mielivaltaisessa määrässä biologisia näytteitä. DyNB-menetelmän tilastollinen malli voidaan nähdä Gaussin-Coxin prosessin laajennukseksi jossa Poissonin jakauma korvataan negatiivisella binomijakaumalla. DyNB-menetelmällä on mahdollista estimoida systemaattisia eroja näytteiden välillä. Solunsisäisten signaalinvälitysten tarkasteluun ja mallinnukseen kehitimme Sorad-menetelmän, joka perustuu differentiaaliyhtälöiden ja gaussisten prosessien yhdistämiseen. Sorad-menetelmä mahdollistaa myös analyysin, jossa estimoidaan miten ennalta määrättyjen komponenttien tulisi käyttäytyä jotta saadaan haluttu vaste aikaan. Väitöskirjan viimeisessä artikkelissa paikannamme nukleosomit tarkasti tutkiaksemme kromatiinin tilan vaikutusta geenien ilmentymiseen.
Tekemämme vertailut aiempiin menetelmiin osoittivat kehitettyjen menetelmien edut. Tämän lisäksi kehitettyjä menetelmiä sovellettiin käytännön biologisiin ongelmiin. Geenien ilmentymisiä tarkastelleissa tutkimuksissa keskityimme napaverestä eristettyjen T-auttajasoluihin. Tuloksemme geenien ilmentymisestä T-auttajasolujen varhaisissa erilaistumisissa tarjoavat hyvän lähtökohdan tarkemmille jatkotutkimuksille. Nukleosomitutkimuksessamme osoitimme, että sytotoksisten T-solujen ja tyypin 1 ja T-auttajasolujen välillä ainoastaan kuudessa prosentissa nukleosomeista nähdään eroja. Havaitut muutokset nukleosomeissa korreloivat erilaistumisohjelman, avoimen kromatiinin, transkriptiotekijöiden sitoutumisen, ja geenien ilmentymisen kanssa. Lisäksi havaintomme tukevat hypoteesia nukleosomien ja transkriptiotekijöiden välisestä kilpailusta DNA:han sitoumisessa.

Avainsanat Bioinformatiikka, laskennallinen biologia, geenien ilmentyminen, transkription säätely, gaussiset prosessit

ISBN (painettu) 978-952-60-5884-9 **ISBN (pdf)** 978-952-60-5885-6

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 174 **urn** http://urn.fi/URN:ISBN:978-952-60-5885-6

Preface

This work has been done with the Computational Systems Biology groups at the Department of Signal Processing at Tampere University of Technology, Finland and the Department of Information and Computer Science at Aalto University School of Science, Finland, over the years 2009–2014. In addition, I spent a year between 2012 and 2013 visiting the laboratory of Prof. Anjana Rao in the La Jolla Institute for Allergy & Immunology, La Jolla, USA.

The work has been supported generously by the Finnish Doctoral Programme in Computational Sciences (FICS), Academy of Finland's Centre of Excellence in Molecular Systems Immunology and Physiology Research (SyMMyS), EU FP7 Systems Biology of T-cell Activation in Health and Disease (SYBILLA), the Inkeri Kantele foundation, and the Emil Aaltonen foundation. I would like to thank all parties for their generous financial support.

First and foremost, I want to express my gratitude to the instructor and supervisor of this thesis Prof. Harri Lähdesmäki. Throughout the process, but especially during challenging moments, his tireless guidance and support have been invaluable.

I am grateful to all co-authors for their contributions. To all friends and colleagues in Tampere and Espoo, thanks for the fun, collaboration and support. Particularly, I want to thank Timo Erkkilä, Kirsi Granberg, Jukka Intosalmi, Antti Larjo, Henrik Mannerstöm, Matti Nykter, Sini Rautio, Antti Rantamäki, Pekka Ruusuvuori, and Antti Ylipää.

I warmly thank Prof. Anjana Rao for giving me the opportunity to visit her laboratory. I thank all the members of her laboratory for making my visit so enjoyable; particularly, Prof. Matthew E. Pipkin, Dr. Gustavo J. Martinez, Dr. Angeliki Tsangaratou, Dr. Sara Trifari and B.Sc. Ryan B. Hastie. A special mention goes to Ryan for her assistance with the

English in this thesis. Although the visit only spanned a fourth of my doctoral studies, it will undoubtedly have a long-lasting influence on my life.

I thank Prof. Riitta Lahesmaa for introducing me to the fascinating field of molecular immunology. Additionally, I am grateful for the pleasure of working with many members of her laboratory; I especially want to warmly thank Dr. Sanna M. Edelman, Dr. Tapio Lönnberg, Dr. Subhash Tripathi, and Dr. Zhi Chen.

I would like to thank the pre-examiners of my thesis, Prof. Alexander J. Hartemink and Prof. Liisa Holm, for providing constructive and valuable comments for improving the thesis.

Last but not least, I wish to thank my family.

Tampere, September 18, 2014,

Tarmo Äijö

Contents

List of Publications

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

I Tarmo Äijö, Sanna M. Edelman, Tapio Lönnberg, Antti Larjo, Henna Järvenpää, Soile Tuomela, Emilia Engström, Riitta Lahesmaa and Harri Lähdesmäki. An integrative computational systems biology approach identifies lineage specific dynamic transcriptome signatures which drive the initiation of human T helper cell differentiation. *BMC Genomics*, 13:572, October 2012.

II Tarmo Äijö, Kirsi Granberg and Harri Lähdesmäki. Sorad: A systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements. *Bioinformatics*, 29(10):1283-91, May 2013.

III Tarmo Äijö, Vincent Butty, Zhi Chen, Verna Salo, Subhash Tripathi, Christopher B. Burge, Riitta Lahesmaa and Harri Lähdesmäki. A time-series analysis of RNA-seq data gives insights into the early human Th17 cell differentiation efficiency. *Bioinformatics*, 30(12):i113-i120, June 2014.

IV Matthew E. Pipkin, Tarmo Äijö, Erbay Yigit, Ivana Djuretic, Quanwei Zhang, Liqun Xi, Ji-Ping Wang, Bjoern Peters, Harri Lähdesmäki and Anjana Rao. Gata3 and Runx3 binding underlies differential nucleosome organization in helper and cytolytic T cells. *Nature Communications*, March 2014.

Author's Contribution

Publication I: "An integrative computational systems biology approach identifies lineage specific dynamic transcriptome signatures which drive the initiation of human T helper cell differentiation"

Äijö developed the methodology together with Lähdesmäki and carried out the computational analyses of the microarray data. Äijö, Edelman and Lähdesmäki wrote the paper.

Publication II: "Sorad: A systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements"

Äijö developed the methodology together with Lähdesmäki and carried out the computational analyses. Äijö wrote the paper together with Granberg and Lähdesmäki.

Publication III: "A time-series analysis of RNA-seq data gives insights into the early human Th17 cell differentiation efficiency"

Äijö developed the methodology together with Lähdesmäki and carried out the computational analyses. Äijö wrote the paper together with Lähdesmäki.

Publication IV: "Gata3 and Runx3 binding underlies differential nucleosome organization in helper and cytolytic T cells"

Äijö developed the novel data analysis methodologies and carried out all the computational and data analyses. Äijö and Pipkin wrote the paper together with Lähdesmäki and Rao.

Acronyms

ANOVA	analysis of variance
BEM-seq	bacterial artificial chromosome enriched mononucleosomal DNA sequencing
BMA	Bayesian model averaging
CaM	calmodulin
CBP	cAMP-response element-binding protein
CDK	cyclin-dependent kinase
ChIP-chip	chromatin immunoprecipitation followed by microarray
ChIP-seq	chromatin immunoprecipitation followed by sequencing
circRNA	circular RNA
CPSF	cleavage and polyadenylation specificity factor
CsA	cyclosporin A
CstF	cleavage stimulation factor
CTCF	CCCTC-binding factor
CTL	cytotoxic T lymphocyte
DBD	DNA binding domain
DNA	deoxyribonucleic acid
DNase I	deoxyribonuclease I
DNase-chip	deoxyribonuclease I digestion of chromatin followed by microarray
DRB	5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole

DREAM	Dialogue on Reverse Engineering Assessment and Methods
DRN	differentially remodelled nucleosome
DSIF	DRB sensitivity inducing factor
EGG	electroencephalography
ELL2	elongation factor, RNA polymerase II, 2
GP	Gaussian process
GRF	Gaussian random field
H3K27ac	histone H3 lysine 27 acetylation
H3K27me3	histone H3 lysine 27 trimethylation
H3K36me3	histone H3 lysine 36 trimethylation
H3K4me1	histone H3 lysine 4 monomethylation
H3K4me2	histone H3 lysine 4 dimethylation
H3K9me2	histone H3 lysine 9 dimethylation
H3K9me3	histone H3 lysine 9 trimethylation
HepG2	hepatocellular carcinoma cell line
HMT	histone methyltransferase
LCR	locus controlling region
limma	linear models for microarray data
lincRNA	large intergenic noncoding RNA
lncRNA	long noncoding RNA
MCMC	Markov chain Monte Carlo
miRNA	microRNA
MLE	maximum likelihood estimator
MNase	micrococcal nuclease
MNase-chip	micrococcal nuclease digestion of chromatin followed by microarray
MVUE	minimum variance unbiased estimator
ncRNA	noncoding RNA
NELF	negative elongation factor
NFAT	nuclear factor of activated T cells
NGS	next-generation sequencing

NRON	noncoding RNA repressor of NFAT
ODE	ordinary differential equation
PE	paired-end
PIC	preinitiation complex
piRNA	piwi-interacting RNA
PRC	polycomb repressive complex
pre-mRNA	precursor messenger RNA
PTEFb	positive transcription elongation factor b
PTM	post-translational modification
RMA	robust multichip average
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RNAP	RNA polymerase
RNAPII	RNA polymerase II
rRNA	ribosomal RNA
SE	single-end
shRNA	short hairpin RNA
siRNA	short interfering RNA
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
SOLiD	sequencing by oligonucleotide ligation and detection
sRNA	short non-coding RNA
SWI/SNF	SWItch/Sucrose NonFermentable
TAD	trans-activating domain
TAF1	TBP-associated factor 1
Th0	activated T helper
Th1	T helper 1
Th17	T helper 17
Th2	T helper 2
TIC	transcription initiation complex
tRNA	transfer RNA

TSS	transcription start site

1. Introduction

The work presented in this thesis resides in the interface of applied statistics and genomics, while focusing on the computational side of the research. Specifically, novel methodologies are presented for preprocessing and downstream analysis of genomic and proteomic data. As a result, the work here can be described by the following overlapping terms, bioinformatics and computational biology.

Transcriptome analysis was revolutionized by the development of microarrays as they made it possible to simultaneously screen the expression of tens of thousands of genes using a hybridization approach. Another revolution in this field was the development of next-generation sequencing technology providing an unbiased and high-throughput tool for measuring absolute quantities of DNA and RNA molecules. In order to utilize the large data sets produced by these experimental techniques, specific computational techniques have been developed with the help of the processing power of computers. At the same time, this era of genome-wide study has required additional changes in the conceptual thinking needed to draw conclusions from the data.

The initial sequencing of the human genome published in 2001 (Venter *et al.*, 2001; Lander *et al.*, 2001) and the subsequent sequencing of the genomes of various other organisms has revolutionized genomics research (Green *et al.*, 2010; Human Microbiome Project Consortium, 2012; Huang *et al.*, 2012a; Nystedt *et al.*, 2013). For instance, in the 1000 genome project (The 1000 Genomes Project Consortium, 2012) the goal is to shed light on variation in the human genome between individuals instead of relying on a reference human genome. Importantly, genome-wide studies have demonstrated the complexity of the intertwined systems of various biological pathways. The ENCODE project demonstrated that over 80%, of the human genome is associated with biochemical activ-

ity. This strongly implies that the amount of so called "junk DNA" (Ohno, 1972) has been overestimated (The ENCODE Project Consortium, 2012). Of course, it is important to note that observed biochemical activity does not necessarily imply functional significance (Graur *et al.*, 2013). In addition, a study in the carnivorous bladderwort plant, *Utricularia gibba*, demonstrated that the development and reproduction of a complex organism does not require a huge amount of nongenic DNA. The authors argue that there could be a species-specific bias towards either nongenic DNA deletion or nongenic DNA insertion and duplication (Ibarra-Laclette *et al.*, 2013). Nevertheless, the functional mapping of the genome remains challenging due to the complexity of regulatory mechanisms, limitations in measurement assays, and cell-type specificities.

Various computational and statistical analyses play an important role in the current molecular biology research. Bioinformatics is a research field in which biology and informatics are combined in order to store, retrieve, organize and analyze biological data. For instance, sequence analysis has been useful in both defining evolutionary relationships between organisms using computationally derived phylogenetic trees (Blanchette and Tompa, 2002; Boffelli *et al.*, 2003), and predicting RNA and protein folding based on the nucleotide (Sharma *et al.*, 2008) and amino sequences (Rost and Sander, 1994), respectively. Moreover, bioinformatic approaches have been successfully applied to annotate genomes. These approaches have been used to identify protein-coding genes (Delcher *et al.*, 1999; Zhang, 2002) and interaction sites between DNA and transcription factors by matching the DNA sequence and the binding domains of the transcription factors (Stormo, 2000; Barash *et al.*, 2003). Additionally, bioinformatics approaches have also been used to extract information by computational literature analysis (Scherf *et al.*, 2005).

Computational biology focuses on deriving mathematical models and rules to describe the behaviour of various biological systems. For example, the use of computational biology for modeling and simulation of various cell processes has gained popularity (Noble, 2002). The bioinformatics and computational biology fields overlap in their application. In the beginning of the millenium, an approach, termed as "systems biology" or "computational systems biology", was proposed (Kitano, 2002b,a). In this approach, the interactions within and between intrinsically complex biological systems are studied together instead of focusing on the separate subunits that exist within that system (Kitano, 2002b,a). The successful

use of a systems biology approach requires diverse high-throughput measurements and elegant computational approaches to extract biologically meaningful and valuable information.

The clinical importance of genomics is increasing as these high-throughput techniques are employed more frequently in clinics and hospitals. Next-generation sequencing can replace and complement the current gene-specific clinical tests. This is the fundamental step towards the long-awaited "personalized medicine", where the treatments are specified for the individual patients (Schilsky, 2010; Tursz *et al.*, 2011; Hood and Friend, 2011). These advances will result in a major shift in the current drug discovery paradigm (Emilien *et al.*, 2000; Kramer and Cohen, 2004; Hall *et al.*, 2010; Woollard *et al.*, 2011).

The transcriptional program of a cell largely determines its function and fate. The transcriptional process is an important regulatory mechanism in eukaryotic cells for ensuring a proper response to a stimulus. Therefore, understanding how cells can control and fine tune their transcriptional programs, is a clinically important and motivated task. In this work, the goal was to gain broad insights into gene transcription and its regulation by approaching scientific questions from multiple perspectives. For instance, we have studied intracellular signal transduction networks and the role of chromatin structure regulation in transcription initiation and regulation. Finally, we present computational methodologies for identifying biologically meaningful differences in temporal gene expression landscapes.

This thesis consists of four peer-reviewed articles published in international journals. Publication I and Publication III present dynamic Gaussian process models for analyzing kinetic data sets obtained using either microarray or next-generation sequencing assays, respectively. The LIGAP tool described in Publication I allows comparison between an arbitrary number of time series microarray experiments using Bayesian analysis. The DyNB methodology described in Publication III is a method for analysis of time series RNA-seq data. In addition, it is applicable for studying systematic differences between replicates. Sorad presented in Publication II is a computational method for studying dynamic signal transduction based on experimental phosphoprotein data. Moreover, it can be used to estimate the optimal perturbations for producing desired behaviour in the signaling cascade. Lastly, in Publication IV we studied transcriptional regulation from the perspective of nucleosome position-

ing. To accomplish this, we produced high-resolution nucleosome maps for a subset of the mouse genome using BEM-seq (bacterial artificial chromosome enriched mononucleosomal DNA sequencing). These nucleosome maps were overlaid with transcription factor binding maps to show that the displacement of individual nucleosomes coincides with the binding of transcription factors in a highly cell-type specific manner.

First, a brief introduction to the necessary biological concepts is given in Chapter 2, followed by a short explanation that will connect genomic experiments with statistical analysis in Chapter 3. An introduction to the computational methodologies used in the publications is given in Chapter 4. The main results of the publications are presented in Chapters 5 through 7. Lastly, the discussion and conclusion are in Chapters 8 and 9, respectively.

2. A Primer on Transcription and its Regulation for a Computational Biologist

In this chapter we provide brief descriptions of the biological concepts covered in this thesis. In summary, we discuss some of the mechanisms cells use to modulate their responses to environmental changes. Our focus is on the transcriptional response, so transcription and its regulation are discussed in more detail. Unless otherwise stated, the concepts are introduced and explained in the context of eukaryotic cells.

In this thesis, the two most important classes of molecules are DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). The genetic code, DNA, is composed of two chains of nucleotides in a double-helix structure. A single nucleotide is composed of a nucleobase, a five-carbon sugar and a phosphate group. The human genome consists of around 3 billion nucleotide base pairs in 23 chromosomes (International Human Genome Sequencing Consortium, 2004). Substructures of DNA, genes, are stretches of biologically meaningful DNA, although "gene" does not have a clear and widely-accepted definition. The Sequence Ontology Consortium reportedly gave the following rather broad definition for the term gene: "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions" (Pearson, 2006). Transcription is the process whereby parts of genetic information coded into DNA are read and copied into RNA molecules by RNA polymerases (RNAPs). Often, these transcribed RNA molecules are also referred to as transcripts. In eukaryotic cells there are three different RNAPs. RNA polymerase I and III transcribe ribosomal and transfer RNAs, respectively; whereas, RNA polymerase II (RNAPII) transcribes the protein-coding genes (Cramer *et al.*, 2008). The transcribed molecules have many important roles in cell function via various pathways. For instance, a subset of transcripts (protein-coding) are translated into proteins. Next we will briefly cover the tran-

scription process in more detail.

2.1 Transcription Process

The transcription process has the following five phases: preinitiation, initiation, promoter clearance, elongation and termination (Shandilya and Roberts, 2012). The preinitiation complex (PIC) is the assembly of general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) and RNAPII along with additional cofactors at the core promoter (Sims *et al.*, 2004). Initiation is driven by the transcription factors at the promoter (about 200 bases upstream from the transcription start site (TSS)) by assembling the transcription initiation complex (TIC) (Nikolov and Burley, 1997). At the same time, an ATP-dependent opening of the double-helix structure takes place at the TSS enabling the transcription of the gene (Nikolov and Burley, 1997). Promoter clearance is the transition of RNAPII from the initiation to active elongation phase during which RNAPII loses contact with the initiation factors. Phosphorylation of serine 5 of RNAPII by the general transcription factor TFIIH enhances the promoter escape of the TIC. Meanwhile, positive transcription elongation factor b (pTEFb) phosphorylates serine 2, thereby enhancing active elongation (Sims *et al.*, 2004) as depicted in Figure 2.1. Transcription is terminated upon the identification of a termination sequence by the protein complexes CPSF (Cleavage and polyadenylation specificity factor) and CstF (Cleavage stimulation factor), which leads to the cleavage of nascent RNA and release of the newly transcribed transcript (Shandilya and Roberts, 2012).

The nascent RNAs undergoing transcription, called pre-mRNAs (precursor messenger RNAs), are subject to post-transcriptional processing. After the cleavage of the nascent RNA, the transcript is modified by the addition of a 5' cap and a poly(A) tail (Lewis and Izaurralde, 1997; Proudfoot, 2011). Furthermore, the RNA molecule undergoes alternative splicing in which it is cut and reassembled in a highly regulated manner (Matlin *et al.*, 2005). After alternative splicing of RNA, which takes place during transcription in the cell's nucleus, the processed RNA chain is called mature mRNA, often abbreviated simply as mRNA (McManus and Graveley, 2011).

Cells transcribe various RNA molecules, besides protein-coding RNAs, with different characteristics whose roles are not yet fully understood. In

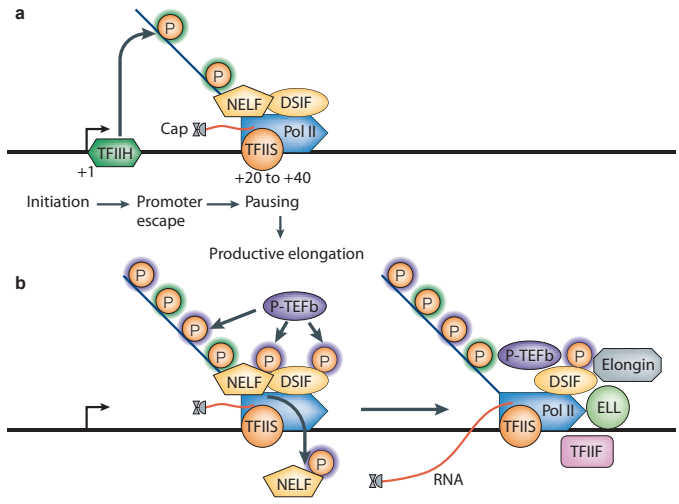

Figure 2.1. Regulation of RNA polymerase II transcription initiaton and elongation. From the initiation phase the phosphorylation of serine 5 of RNAPII by TFIIH leads to promoter escape. The phosphorylation of serine 2 by positive transcription elongation factor b (pTEFb) drives RNAPII to active elongation and transcription. Adapted from (Saunders *et al.*, 2006) with permission from Macmillan Publishers Ltd.

the next section we will cover some of the known RNA molecule families with important regulatory functions.

2.2 Transcriptome

The transcriptome of a cell is its catalogue of RNA molecules, along with the corresponding transcript levels (Ozsolak and Milos, 2011). Transcriptomes are important for cellular behaviour because transcriptomes of different cell types are variable and they are dynamic in nature during cell differentiation (Rhodes and Chinnaiyan, 2005; Cloonan *et al.*, 2008; Wilhelm *et al.*, 2008; Wu *et al.*, 2010; Sharov *et al.*, 2003). Although transcriptomes are highly variable between cell types, the genome is highly stable between different cell types. This emphasizes the importance of understanding regulation of the transcriptome and the consequent effects on cell fate.

Only one-fifth of the RNAs in the transcriptome are protein-coding (Kapranov *et al.*, 2007). Several different functional families of RNA molecules have been described and are distinguished by their size, structure, and function (see Figure 2.2). The RNAs belonging to the major protein synthesis complex, the ribosome, are referred as rRNAs (ribosomal RNAs).

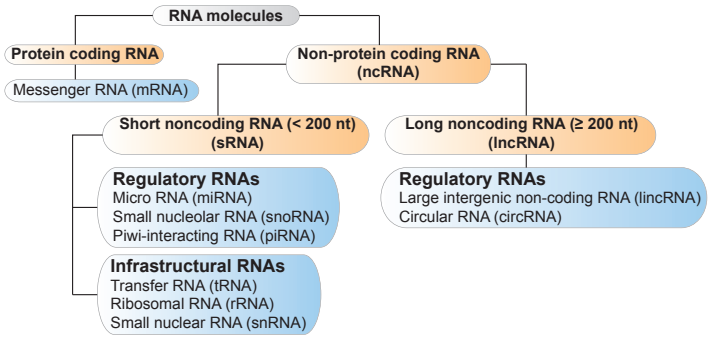

Figure 2.2. Classification of selected RNA molecules by function and size. RNA molecules can be separated based on whether they are translated to proteins or not. Moreover, the noncoding RNAs can be categorized based on the size and function.

Therefore, RNAs have a major role in translation (Noller, 1991). The short RNAs, e.g., tRNAs (transfer RNAs), snoRNAs (small nucleolar RNAs), piRNAs (piwi-interacting RNAs), and miRNAs (microRNAs), are ofted referred by the umbrella term of short noncoding RNAs (ncRNAs) (Kim *et al.*, 2009). Importantly, miRNA, siRNA (short interfering RNA), and shRNA (short hairpin RNA), are capable of regulating gene expression via the RNAi (RNA interference) pathway (Wilson and Doudna, 2013). For example, miRNAs bind to the 3'UTR regions of mRNAs and thereby promote mRNA degradation and prevent translation (He and Hannon, 2004). Recently, there has been an increased interest in the study of lincRNAs (large intergenic noncoding RNAs) (Guttman *et al.*, 2009) and circRNAs (circular RNAs) (Memczak *et al.*, 2013; Hansen *et al.*, 2013). For example, lincRNAs have been shown to play a role in pluripotency and differentiation of embryonic stem cells (Guttman *et al.*, 2011). The lincRNA, noncoding RNA repressor of NFAT (NRON), has been identified as having an inhibitory effect on nuclear factor of activated T cells (NFAT) (Willingham *et al.*, 2005; Sharma *et al.*, 2011).

2.3 Transcription Regulation

The focus of this section is transcription regulation. First of all, transcription initiation and elongation are highly regulated processes (Sims *et al.*, 2004; Maston *et al.*, 2006; Saunders *et al.*, 2006). The transcribed mRNAs are post-transcriptionally regulated at different levels, including processing, stability, translation and transport (Maston *et al.*, 2006). Moreover,

we will briefly describe chromatin and chromatin remodeling molecules, and cover some of the regulatory mechanisms of chromatin in modulating transcription. More detailed descriptions of these concepts can be found, e.g., in the reviews on transcription regulation and on the role of the chromatin structure in gene accessibility (Li *et al.*, 2007; Weake and Workman, 2010; Beisel and Paro, 2011; Spitz and Furlong, 2012)

2.3.1 Regulation of RNA Polymerase II Elongation

Active elongation of RNAPII can be paused by interaction between negative elongation factors NELF (negative elongation factor) and DSIF (DRB sensitivity inducing factor) by increasing the sensitivity to DRB (5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole) (Lee and Young, 2000). DRB is an inhibitor of members of the family of cyclin-dependent kinases (CDKs) (Yankulov *et al.*, 1995), including CDK9. By inhibiting CDK9, DRB inhibits RNAPII elongation by suppressing the effect of the positive transcription elongation factor b (PTEFb) (Ip *et al.*, 2011). PTEFb is a cyclin-dependent kinase formed by CDK9 and other cyclin subunits with the ability to phosphorylate NELF and DSIF, thereby inducing active RNAPII elongation by releasing the stalled RNAPII (Kohoutek, 2009). There are additional positive elongation factors; for example, TFIIH has a positive effect on the initial unwinding of DNA (Sims *et al.*, 2004). TFIIF remains attached with the elongating RNAPII exerting a positive effect on elongation (Sims *et al.*, 2004), and ELL2 (elongation factor, RNA polymerase II, 2) prevents RNAPII backtracking by enhancing the alignment of the 3' OH of nascent RNA with the catalytic site (Martincic *et al.*, 2009). These mechanisms are illustrated in Figure 2.1.

2.3.2 DNA-binding Proteins

A subset of proteins have DNA binding domains (DBDs) and are thereby capable of highly specific binding directly to DNA. They are referred to as transcription factors if they have positive or negative effects on transcription through DNA binding. Transcription factors have been shown to influence transcription by binding to distal regulatory regions and core and proximal promoters. The DBDs on the proteins are used to classify transcription factors into different families; for a review on transcription factor families see (Pabo and Sauer, 1992; Latchman, 1997). Recently, the binding specificities of 830 human transcription factors were measured

systematically using the SELEX technique (reviewed by Stoltenburg *et al.*, 2007) yielding 239 different binding specificity models (Jolma *et al.*, 2013). Other proteins without DBDs also regulate transcription, such as coactivators, corepressors, chromatin remodelers, histone acetylases, and histone deacetylases. The coacting proteins, i.e., coactivators and corepressors, are proteins which form complexes with transcription factors through trans-activating domains (TADs) without direct interaction with DNA. All the possible mechanisms by which proteins regulate transcription are not yet fully understood. For instance, transcription factors can either positively or negatively regulate the recruitment of RNAPII to the promoter, chromatin accessibility for other proteins and transcription machinery, and RNAPII elongation in gene body. Distal regulatory regions are further divided into enhancers, silencers, insulators, and locus controlling regions (LCRs) based on their functions (Maston *et al.*, 2006). Distal regulatory regions have been shown to mediate their regulatory effect through mediators and DNA looping, which emphasizes the importance of studying the three-dimensional structure of chromatin. In murine embryonic stem cells, the protein complexes mediator and cohesin have been shown to connect enhancers and core promoters (Kagey *et al.*, 2010). Moreover, CTCF (CCCTC-binding factor), a protein reported to form an insulator, together with cohesin, has been reported to have a similar function (Merkenschlager and Odom, 2013).

2.3.3 Chromatin

The chromatin is the structure of DNA packed into cell nucleus (see Figure 2.3). The different chromatin structure levels, from large to small, are chromosome, 30 nm fibre, and 10 nm fibre (Olins and Olins, 2003). Chromatin has the following four major functions: packing the DNA into the cell nucleus, regulating gene expression, allowing mitosis, and protecting DNA from damage (Olins and Olins, 2003). The basic repeating unit of chromatin is the nucleosome (Olins and Olins, 2003). The nucleosome core particle is an octamer composed of histone proteins H2A, H2B, H3, and H4 where each one is present in two copies, and approximately 147 base pairs of DNA. The DNA wrapped around the histone octamer in an 1.65 superhelical turn is termed mononucleosomal DNA (Andrews and Luger, 2011). In addition to these four canonical histone proteins, there are histone protein variants H2AZ and H2AX (Redon *et al.*, 2002), whose function on chromatin activity is not yet widely studied. There are studies

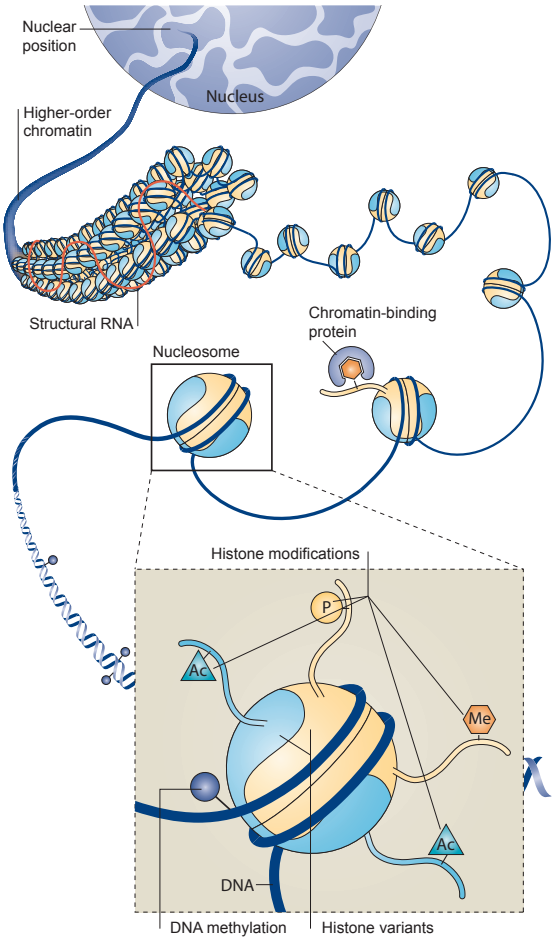

Figure 2.3. The packaging of DNA into chromatin structure. Nucleosomes composed by the histone proteins are the basic repeating units of chromatin. The beads on string form of DNA associated with nucleosomes leads to the packaging of DNA into higher-order chromatin structure in cell nucleus. Adapted from (Probst *et al.*, 2009) with permission from Macmillan Publishers Ltd.

showing that H2AX alters nucleosome instability (Zlatanova and Thakar, 2008) and phosphorylation of H2AX has been associated with DNA damage (Sharma *et al.*, 2012). The linker histone protein H1 does not belong to the nucleosome core particle, as it sits on the top of the structure keeping the wrapped DNA in place. Around 80% of DNA is associated with the core nucleosome particles, the remaining unassociated DNA between the nucleomes is termed as linker DNA (Workman and Kingston, 1998; Alberts, 2007). Individual nucleosomes are packed into a 10 nm fibre and a higher level packing of multiple nucleosomes produces a 30 nm fibre (Belmont *et al.*, 1999; Tremethick, 2007). Based on the packing density chromatin is classified into euchromatin and heterochromatin. In

the euchromatin state the chromatin is sparsely packed; whereas, in the heterochromatin state the chromatin is tightly packed (Grewal and Jia, 2007; Gaspar-Maia *et al.*, 2011). In most cases, the DNA wrapped around the nucleosomes is protected from binding by transcription factors. The binding is blocked because of the inaccessible conformation of DNA, which prevents the identification of the target DNA by DBD. Finally, the determinants governing nucleosome positioning are reviewed in (Struhl and Segal, 2013).

2.3.4 Chromatin Remodeling and Post-translational Modifications on Histone Tails

Dynamic chromatin regulation is a pathway for regulating chromatin structure and accessibility. For instance, chromatin accessibility is regulated at the promoters in order to control the binding of transcription machinery or transcription factors. Different mechanisms driving dynamic chromatin remodeling are not fully understood yet. The SWI/SNF (SWItch/Sucrose NonFermentable) protein complex has been shown to regulate DNA accessibility. It has an ATP-dependent mechanism to destabilize the contact between the DNA molecule and the histone protein complex (Vignali *et al.*, 2000). Additionally, histone acetyltransferases and deacetylases are proposed to a have role in controlling chromatin structure by modifying the acetylation state of histone proteins. Acetylated histones have weakened contact with the DNA segment (Struhl, 1998). The knowledge of different chromatin remodeling pathways are summarized, e.g., in the following reviews (Cosgrove *et al.*, 2004; Zentner and Henikoff, 2013). LincRNAs have also been linked to chromatin remodeling (reviewed in Bergman and Cedar, 2013). They bind to chromatin remodelers, thereby modulating their catalytic activity and controlling their chromatin targets.

The tails of histone proteins have several residues associated with post-translational modifications (PTMs) (see Figure 2.3). PTMs of the tails include methylation, acetylation, phosphorylation, ubiquitination, SUMOylation, citrullination, and ADP-ribosylation (Suganuma and Workman, 2011). Histone methyltransferases (HMTs) are the proteins which can catalyze one or multiple methyl groups to the specific residues on the histone tails (Wood and Shilatifard, 2004; Greer and Shi, 2012).

Commonly, these histone tail PTMs are described by giving the name of the affected histone, the specific amino acid abbreviation with the position, the type of modification and the number of modifications (Turner,

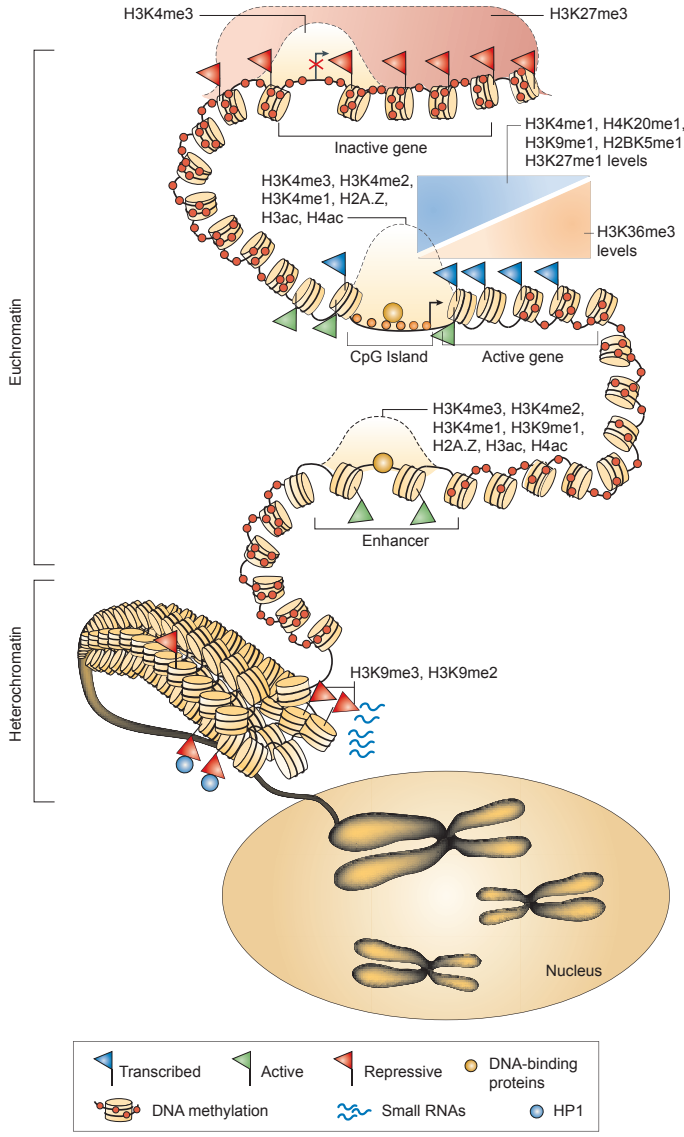

Figure 2.4. Examples of post-translational modifications. A selected list of post-translation modifications on histone tails, which are associated, for instance, with euchromatin and heterochromatin, repressed and active transcription and enhancer activity. For example, H3K36me3 is often found within the actively transcribed genes. Adapted from (Schones and Zhao, 2008) with permission from Macmillan Publishers Ltd.

2005). For example, H3K4me2 denotes the dimethylation of the 4th lysine residue from the beginning of the N-terminal in the H3 protein. Some of the post-translational modifications have been associated with active and poised transcription and different chromatin states as the schematic in Figure 2.4 illustrates. For example, H3K36me3 is found in the gene body of the actively transcribed genes; whereas, H3K27me3 is found in genes

which are not transcribed (Kooistra and Helin, 2012). Heterochromatin is associated with the marks H3K9me2 and H3K9me3 (Rosenfeld *et al.*, 2009). The proteins containing the JmcJ domain are shown have the ability to demethylate residues on histones tails. For example, JHDM1 and JMJ3 are identified to be H3K36 and H3K27 demethylases, respectively (Tsukada *et al.*, 2006; Xiang *et al.*, 2007). Two members of polycomb repressive complexes (PRCs), PRC1 and PRC2, are known to play a role in regulating gene expression by repressive actions (Schwartz and Pirrotta, 2007). For instance, PRC1 monoubiquitylates the histone H2A and PRC2 catalyzes the methylation (di- and tri-) of H3K27 (Margueron and Reinberg, 2011).

The H3K4me1 and H3K4me2 histone marks associated with open chromatin have been identified to be enriched in putative enchancers characterized by the binding of p300 (Heintzman *et al.*, 2007, 2009). The proteins p300 and CBP (cAMP-response element-binding protein) compose a coactivator family, and they have also been shown to act as histone acetyltransferases (Ogryzko *et al.*, 1996). Active enhancers, defined to have RNAPII and a subunit of transcription initiation factor TBP-associated factor 1 (TAF1) (Ong and Corces, 2011), allegedly have different histone marks compared to inactive or poised enhancers. The distinct histone mark of active enhancers is H3K27ac, which is speculated to be present due to the acetyltransferase activity of p300 and CBP (Creyghton *et al.*, 2010).

As a result, transcription is hypothesized to be regulated partly by the post-translational modifications of the histone tails (Strahl and Allis, 2000; Jenuwein and Allis, 2001). That is, these post-translational modifications of histone tails may regulate chromatin structure and transcription directly or, alternatively, via recruitment of additional protein effectors (Suganuma and Workman, 2011). The interpretation of this putative biological regulatory mechanism is complicated because it has been hypothesized that the histone modifications are crosstalking and likely to have combined effects (Suganuma and Workman, 2011).

2.3.5 Extracellular and Intracellular Signaling

In the previous sections we covered various direct intracellular mechanisms of transcription regulation without focusing on the influential indirect mechanisms. First of all, the cells and their behaviour are not isolated from the environment. For example, the cells are communicating with

each other and with the environment using various signal transduction mechanisms. Through these signal transduction mechanisms cells modulate their metabolism, function and development. The signal transduction from the signaling cell to the target cell is mediated by the signaling molecules secreted by the signaling cell. These secreted signal molecules are recognized by the corresponding receptors on the target cell surface.

There are various classes of signaling molecules with different biochemical characteristics, for instance, hormones, cytokines and neurotransmitters. In addition, different receptors on the cell surface can be classified into four subclasses: ion channel, enzyme, tyrosine kinase and G protein-coupled receptors. All of these subclasses have extracellular, membrane and intracellular domains (Alberts, 2007). From the cell surface, the signal produced by the binding of a signaling molecule to a corresponding receptor is further transduced by various secondary messenger molecules which can translocate to the cell cytoplasm and nucleus.

A signal transduction process can be described as endocrine, paracrine, juxtacrine, autocrine or intracrine, where the classification is based on the domain where signal transduction takes place (Alberts, 2007). Signal transduction between cells is endocrine if the transduction happens between distal cells, paracrine if the communicating cells are spatially close, and juxtarine if the cells are touching (Alberts, 2007). Signal transduction within a cell is autocrine if it takes place through the membrane receptors, and intracrine if it happens within the cell (Alberts, 2007).

As an example, we quickly describe the calcineurin/NF-AT pathway partially depicted in Figure 2.5, which leads to the NFAT translocation and consequently regulation of transcription (Shaw *et al.*, 1988). The cell receives the extracellular signal when the T cell receptor on its surface encounters an antigen-presenting cell. Upon T cell activation, PLCγ1 is phosphorylated leading to the production of IP3 and the release of intracellular calcium (Ca^{2+}) stores, and an overall increase of Ca^{2+} levels in the cytoplasm (Smith-Garvin *et al.*, 2009). The increase in the level of cytoplasmic calcium is sensed by the STIM1 and STIM2 sensor molecules, which have a role in activating the Ca^{2+} release activating channels (Liou *et al.*, 2005). Consequently, the calmodulin (CaM) Ca^{2+} sensor protein activates the inhibitory protein complex calcineurin, which is composed of the subunits calcineurin A and calcineurin B, where calcineurin A interacts with calmodulin in a Ca^{2+}-dependent manner and binds Ca^{2+} (Klee *et al.*, 1979). Activated calcineurin will dephosphorylate NFAT, leading to

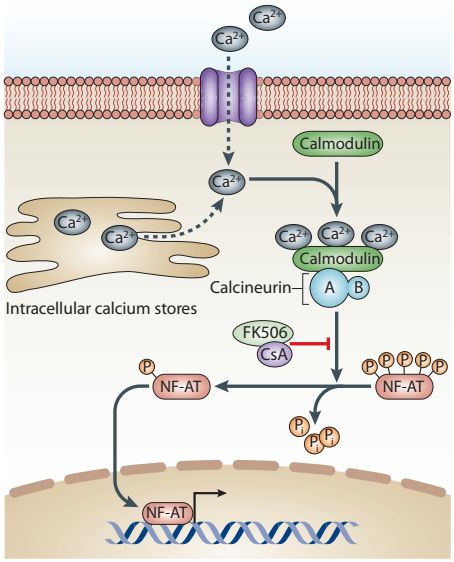

Figure 2.5. The calcineurin/NF-AT pathway drives the translocation of the NFAT transcription factor to the cell nucleus. The T cell activation releases the Ca^{2+} stored in the intracellulalr calcium stores. The increase in the cytoplasmic Ca^{2+} levels activates the Ca^{2+} release channels, which leads to the activation of calcineurin by calmodulin. The activated calcineurin inhibitory protein complex dephosphorylates NFAT allowing it to be translocated into the cell nucleus to drive the T cell transcription program. Adapted from (Steinbach *et al.*, 2007) with permission from Macmillan Publishers Ltd.

a rapid translocation of NFAT from the cytoplasm into the cell nucleus (Okamura *et al.*, 2000). After the translocation to the nucleus, NFAT drives T cell activation by cooperating with various transcription factors, e.g., by forming a complex with AP-1 which will lead to IL-2 production (Jain *et al.*, 1992). The immunosuppressant drugs FK506 and cyclosporin A (CsA) and have been shown to block the effect of the calcineurin/NF-AT pathway by inhibiting NFAT translocation (Rühlmann and Nordheim, 1997).

3. Bridging Experiments and Statistical Analysis

In this chapter we briefly cover the main experimental and computational factors one has to take into account when analyzing genomic data obtained using high-throughput technologies. Two important high-throughput measurement technologies in genomics are DNA microarrays and next-generation sequencing. The introduction of DNA microarray technology revolutionized genomics research by allowing genome-wide studies. DNA microarrays are still widely used for standard genotyping and gene expression profiling because of their simplicity, cost-efficiency, and adaptability of automation.

In general, DNA microarrays and next-generation sequencing are methods for measuring DNA and RNA contents of biological samples. The complete cataloging of the biological material within a cell is not currently possible, so studies must focus on a certain aspect by enriching a subset of the biological material. For example, a study can focus on measuring the DNA reverse transcribed from miRNAs or mRNAs, which are size-selected or selected based on the poly(A) tail, respectively. It is important to include proper controls for the enrichment steps, which should be taken into account in the data analysis. For instance, in antibody-based applications where genomic DNA is fragmented prior to the immunoprecipitation step, it is highly recommended to include a control samples that has not been subjected to immunoprecipitation in the analysis. This control will be beneficial in accounting for fragmentation bias or other biases.

Different computational analysis methods for microarray and next-generation sequencing data are needed for various reasons. First, they are necessary because of the vast amount of generated data. Second, they can provide a statistically sound framework for taking into account the biases in the data by modeling the measurement process and combining replications. Third, they can provide a quantitative framework for identifying

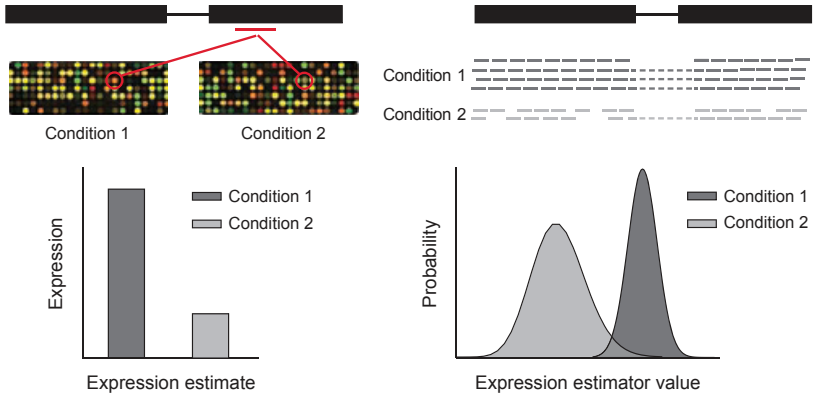

Figure 3.1. Gene expression estimation using microarrays and RNA-seq. A schematic of a gene consisting two exons is depicted in the top panel. A probe in a microarray is targeting a specific short genomic region within a exon. The probe intensity is used as a proxy for the gene expression. In the RNA-seq approach fragmented mRNA molecules are sequenced, which are then mapped against the exons and exon-exon junctions. After normalizing the different sequencing depths the number of aligned sequencing reads within the gene can be used in the gene expression estimation. Finally, the read coverage information can be useful in estimating the distribution of the gene expression estimates. Adapted from (Garber *et al.*, 2011) with permission from Macmillan Publishers Ltd.

biologically significant signals from the data.

3.1 DNA Microarrays

A DNA microarray is a dense collection of predefined DNA oligos, often referred to as probes, which are used to measure the DNA content of a sample based on the hybridization of the probe to a DNA target sequence as depicted in Figure 3.1 (Heller, 2002). Initially, microarrays were used mostly for probing genotypes and gene expression (Schena *et al.*, 1995; Hoheisel, 2006), but they have been usefully applied for other purposes, including ChIP-chip (chromatin immunoprecipitation followed by microarray) (Ren *et al.*, 2000), protein detection (Haab, 2001) and chromosome conformation capture (Simonis *et al.*, 2006; Zhao *et al.*, 2006). Probes are designed in such a way that they cover the sequences of the regions of interest in the genome while maximizing specificity (Heller, 2002). The probe lengths in different microarray designs usually vary between 14 to 60 nucleotides. For instance, the probes in the Affymetrix gene expression arrays and Agilent microarrays are 25 and 60 nucleotides in length, respectively (Heller, 2002; Hardiman, 2004).

In the preparation step, the DNA fragments are fluorescently labeled, which allows them to be detected using a laser excitation. The quantification of the emitted light is done using either a single- (single color) or two-sample (two colors) microarray design. Importantly, this type of quantification does not provide absolute expression estimates of the targeted genomic regions. Therefore, the differences between the conditions being studied are quantified relative to another sample on the same array (two-sample array) or computationally (single-sample array). Moreover, the strength of the signal cannot be treated as a proxy for the abundance of the target sequence, e.g., due to the differences between the probe affinities (as A/T-rich probe sequences have shown to have lower hybridization intensities than probe sequences with high G/C content) (Heller, 2002). In addition, microarrays have been shown to have sensitivity and specificity issues (Draghici *et al.*, 2006). Because probe-based technologies have been in the market for a while and are widely used, the basic analyses, preprocessing and secondary data analysis, of the most common microarray data types are well-established.

After imaging the preprocessing of the raw data begins, the spots are identified and their fluorescence intensitity values are estimated using a selected image quantification software, followed by removal of bad quality spots (Quackenbush, 2002). For gene expression arrays, the previous steps are commonly followed by the application of the robust multi-chip average (RMA) method, which consists of a background adjustment, quantile normalization, and summarization steps. There are alternative normalization techniques available, but the importance of doing the normalization in a proper way should be stressed. For making the probes comparable between arrays, the data has to be transformed and normalized within an array and between arrays. This has to be done in order to correct differences in the quantities of the starting biological material, hybridization preferences, and in labeling and detection (Quackenbush, 2002). One can assess the quality of the data by generating various plots: density plots of probe intensities within an array, boxplots illustrating the probe intensity distributions across a set of arrays, and ratio-intensity plots from two-channel arrays. In addition, the ComBat software (Johnson *et al.*, 2007) can be used to remove batch effects from the arrays, such as systematic differences between the arrays caused by preparing them in different hybridization runs. Finally, the data is usually log transformed in order to make the up- and down-regulated ratios comparable (Quack-

enbush, 2002).

Commonly, after preprocessing the data, limma (linear models for microarray data) (Smyth, 2005) is used for a statistical identification of the differentially expressed genes between a set of conditions. Importantly, the number of replicates is usually limited; thus, estimating the variation in the data can be difficult. Therefore, limma uses an approach which assumes that the probe sets with similar values have similar variation and thus estimates the variances robustly by sharing information between probes (Smyth *et al.*, 2005). The limma software allows for the definition of sophisticated study designs, such as, paired samples study designs, which can be useful in detecting modest changes (Smyth, 2005).

In practice, the probe-based approach is limited by the fact the one can only detect what is included in the probes, making identification of novel coding or noncoding transcripts or fusion genes practically impossible. Another limitation is that not all sequences can be targeted reliably using short probes (Okoniewski and Miller, 2006). Similarly, identification of single nucleotide polymorphisms (SNPs) is limited to the known or predefined ones, and the expression estimation relies on short RNA stretches. Spatial signal patterns can be captured using probes targeting adjacent sequence segments. Obviously, the spatial resolution of microarrays can be improved by using highly overlapping probes. For example, tiling microarrays have been used to detect interactions between DNA-binding proteins and DNA (ChIP-chip) (Cawley *et al.*, 2004), DNase I (deoxyribonuclease I) hypersensitivity sites (DNase-chip) (Crawford *et al.*, 2006), and nucleosome positions (MNase-chip) (Song *et al.*, 2008).

3.2 Next-generation Sequencing

To get a glimpse of next-generation sequencing, a reader may refer to the reviews by Shendure and Ji (2008), Metzker (2010) and Mardis (2013). Instead of using probes to capture the fragments, as in the microarray-based approaches, sequencing directly measures the nucleotide composition of the fragments. This is the biggest conceptual difference between microarrays and next-generation sequencing as depicted in Figure 3.1 (Mardis, 2013). Often the sequencing approaches are divided into Sanger and next-generation sequencing (NGS) technologies. Unlike the Sanger sequencing used in the initial human genome sequencing project (Lander *et al.*, 2001), the NGS approaches produce millions of reads with shorter

read lengths (Mardis, 2013). Importantly, sequence count is a direct measure of abundance with a high dynamic range instead of the relative measure obtained using the microarray approaches (Mardis, 2013). Various NGS platforms are available (see Lam *et al.* (2012); Liu *et al.* (2012)), but the two most widely used platforms are SOLiD (Sequencing by Oligonucleotide LIgation and Detection) and Illumina sequencing (sequencing by ligation and synthesis) (Liu *et al.*, 2012). Sequencing libraries can be constructed and sequenced in different ways. For instance, the fragments can be sequenced from one of the ends or from both ends, which are referred as single-end (SE) and paired-end (PE) sequencing, respectively (Mardis, 2013). Depending on how the sequencing library is designed, PE sequencing can be subdivided into standard PE sequencing (short fragments) and mate-pair sequencing (long fragments) (Mardis, 2013). Recently, several studies have been published to identify the effects, such as PCR artifacts and GC bias, of the measurement technology and library construction protocols on the results (Malone and Oliver, 2011; Nookaew *et al.*, 2012; Kogenaru *et al.*, 2012; Giorgi *et al.*, 2013; Ross *et al.*, 2013).

An advantage of NGS approaches over microarray approaches is the ability to reformulate the hypothesis afterwards, e.g., one can check the existence of a novel fusion gene from old data sets. In addition, NGS approaches have, in general, higher resolution than microarrays due to direct sequencing of the fragments. However, there are drawbacks: sequencing takes more time, sequencing library construction is more tedious, sequencing is more expensive, and the sequencing library construction might require more biological starting material.

3.2.1 Sequence Aligment

For each sequenced fragment, the sequencer analyzes the raw images and outputs a nucleotide sequence together with the quality scores. These quality scores are interpretable as a probability that the base was sequenced correctly in widely-accepted format, such as, FASTQ (Cock *et al.*, 2010; Metzker, 2010). The quality scores, often in the Phred format (Ewing *et al.*, 1998), are useful in assessing if the sequencing was successful and in removing bad quality reads (Patel and Jain, 2012). Additionally, sequencing adapters are removed and an existence of unexpected bias in the nucleotide distribution is checked (Horner *et al.*, 2010; Patel and Jain, 2012). Moreover, one should check the average quality score as a function of read position and the level of sequence duplication caused by low

complexity of the sequencing library.

Often the next step is to identify the sequenced fragments in respect to a reference genome. In other words, to align or map them against the reference genome, which has been sequenced and constructed previously using appropriate sequencing techniques (Metzker, 2010). The importance of the alignment step should be emphasized since, in practice, it will define the starting point and constraints for the analysis. The choice of the reference genome depends on the biological application, e.g., in the case of mRNAs or bisulfite-treated DNA one has to take into account the splicing or the conversion of unmethylated cytosine residues into uracils, respectively (Trapnell *et al.*, 2009; Krueger and Andrews, 2011). The Smith-Waterman algorithm (Smith and Waterman, 1981) does solve the local sequence alignment problem in a general form, but it is not applicable for aligning NGS data sets due to the large size of genomes and great number of sequences. Therefore, various approximative alignment algorithms based on efficient hashing or Burrows-Wheeler transformation have been proposed, such as MAQ (Li *et al.*, 2008), BWA (Li and Durbin, 2009), Bowtie (Langmead *et al.*, 2009), and Bowtie 2 (Langmead and Salzberg, 2012). The SOLiD method differs from other sequencing methods in that the sequencing results are read in the color-base, i.e., it uses the 2-base code (Liu *et al.*, 2012). Therefore, instead of directly converting the reads from color-space into base-space, which does cause a frame shift upon a read error, it is advised that the reads are aligned in the color-space (Li and Homer, 2010). The aligments are reported independently based on the choice of the aligment software, either in the SAM (human readable) or BAM (compressed binary version) formats. The SAM and BAM formats are easily manipulated using the samtools software (Li *et al.*, 2009). In general, PE sequencing is more informative than SE sequencing; for instance, it provides the actual fragment lengths and eases the identification of unique alignments (Metzker, 2010). However, PE sequencing is complicated by various and higly abundant repetitive elements (de Koning *et al.*, 2011), especially because of the short read length (Treangen and Salzberg, 2012). If an appropriate reference genome or transcriptome is not available, then one can attempt to use the reads to construct a candidate genome, i.e., combining overlapping sequence reads in order to identify longer sequence contigs. Many methodologies are available for carrying out de novo genome and transcriptome assembly; for instance, Velvet (Zerbino and Birney, 2008), AbySS (Simpson *et al.*, 2009), Cufflinks

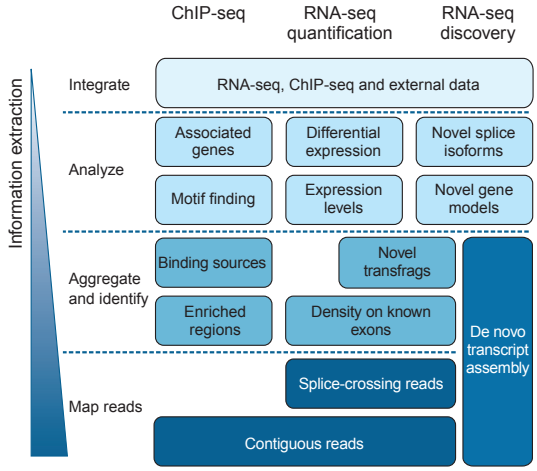

Figure 3.2. Information extraction from ChIP-seq and RNA-seq experiments. The ChIP-seq and RNA-seq quantification and discovery analysis workflows are proceeding from the bottom to top. The analysis starts by aligning the reads against the reference genome or alternatively by constructing a de novo transcriptome. This is immediately followed by a step where the signal is locally quantified by detecting enriched or depleted regions or within certain predefined genomic regions. The implications of the findings are studied from different perspectives and finally integrated across the data types. Adapted from (Pepke *et al.*, 2009) with permission from Macmillan Publishers Ltd.

(Trapnell *et al.*, 2010), and Trinity (Grabherr *et al.*, 2011).

3.2.2 Downstream Analysis

There are several reviews describing the basic analysis of ChIP-seq (chromatin immunoprecipitation followed by sequencing) and RNA-seq (RNA sequencing) data depicted partially in Figure 3.2 (Pepke *et al.*, 2009; Oshlack *et al.*, 2010; Ghosh and Qin, 2010). Moreover, a protocol paper by Trapnell *et al.* (2012) describes the use of the Tuxedo suite pipeline consisting of Bowtie (Langmead *et al.*, 2009; Langmead and Salzberg, 2012), Tophat (Trapnell *et al.*, 2009), and Cufflinks (Trapnell *et al.*, 2010) softwares for transcriptomics analysis. Often the next step after the alignment is the generation of the read coverage signal, which allows visual inspection of the signal. Read coverage signals can be stored using using binary and indexed formats, bigWig and bigBed (Kent *et al.*, 2010), which allow a fast and remote display of the data using, e.g., the UCSC genome browser (Kent *et al.*, 2002).

The counts of read fragments and aligned sequencing reads vary between experiments. Thus, the different sequencing depths of the samples have to be made comparable using normalization. Different normaliza-

tion techniques have been proposed for specific applications. For instance, the total number of reads (Mortazavi *et al.*, 2008) or the median of the ratios of observed read counts (Anders and Huber, 2010) has been used as a normalization factor in sequencing depth normalization. Moreover, a regression-based within- and between-lane normalization procedure has been proposed for GC content normalization (Risso *et al.*, 2011).

Usually some sort of a local quantification of the detected signal has to be done to ease the analysis and interpretation of the data. For instance, calculating the number of reads arising from each gene or exon, focusing on the gene promoters or identifying interesting regions throughout the genome. These methods can be distinguished by whether they identify the interesting regions automatically or use predefined region annotations. For many organisms the protein-coding genes are well annotated (Pruitt *et al.*, 2009; Flicek *et al.*, 2013). Thus, one can simply use these annotations while quantifying the signal, or as a basis for detecting novel transcripts (Anders and Huber, 2010; Trapnell *et al.*, 2010). Unfortunately, the annotation-based approach is not always practical, e.g., the gene-level quantification does not work well if the signal is intra- and intergenic. In those cases, the common approach is to identify either the regions where the signal is enriched or depleted compared to a control or another biological sample. Several approaches based on an identification of signal enrichments have been successful; for instance, mirDeep2 identifies novel RNA molecules (Friedländer *et al.*, 2012), MACS identifies interactions between DNA-binding proteins and DNA (Zhang *et al.*, 2008), and SICER detects histone modifications (Zang *et al.*, 2009).

The statistical models used in microarray data analysis are not statistically sound for NGS data due to the different nature of the data, i.e., arbitrary intensity values versus discrete read counts. Therefore, many of the analysis methods are built based on the assumption of Poisson or negative binomial distribution, such as, DESeq (Anders and Huber, 2010), edgeR (Robinson *et al.*, 2010), MACS (Zhang *et al.*, 2008), SICER (Zang *et al.*, 2009), and cn.MOPS (Klambauer *et al.*, 2012).

More challenging than the analysis of individual data types, is the integration of various data types to get a systematic view that allows for the interpretation of stronger biological conclusions. Even a simple analysis, where the goal is to detect a causal relationship between binding of transcription factor and induced transcription, turns out to be difficult when the binding site is located outside of the promoter. Since various

histone PTMs have been associated with inactive and active enhancers various machine learning approaches, such as support vector machines (Fernández and Miranda-Saavedra, 2012) and random forests (Rajagopal *et al.*, 2013), have been proposed to combine the PTM signals to improve the sensitivity and accuracy of enhancer identification. Moreover, instead of focusing solely on enhancers, various chromatin states have been annotated by integrating the histone PTM information from different cell types by searching for patterns using hidden Markov models (Ernst *et al.*, 2011; Ernst and Kellis, 2012).

Due to the young age of the NGS technology, the analysis pipelines for different NGS applications are not yet well-established. However, the Galaxy project (Goecks *et al.*, 2010) aims to build an online platform for doing biological data analysis. The main idea behind Galaxy is to build analysis pipelines by connecting simple premade analysis modules together. The paradigm of Galaxy, in theory, allows for building of reusable analysis pipelines, making data analysis accessible to everyone (Goecks *et al.*, 2010).

3.2.3 Example Applications

In this section we go through some of the widely-used NGS applications. The future and current applications of DNA sequencing are discussed in the review by Shendure and Lieberman Aiden (2012). Generally, one can interchange microarrays with NGS technologies with minimal changes. RNA-seq can be used for probing RNA landscapes, such as, mRNA levels. Importantly, because RNA-seq is not limited to the protein-coding genes and does not require the definition of the targeted sequences beforehand, it has proven to be useful for identifying novel RNA molecules. For instance, RNA-seq has been used in identifying lincRNAs, alternatively spliced mRNA molecules, and fusion genes yielded by scrambled genomes found in cancer cells.

Due to the lack of probes, NGS can be used for de novo construction of genomes (Li *et al.*, 2010) and transcriptomes (Guttman *et al.*, 2010; Martin and Wang, 2011; Grabherr *et al.*, 2011). Moreover, whole genome or transcriptome sequencing reads can be straightforwardly used for detecting SNPs in respect to the given reference genome (see review by Nielsen *et al.* (2011)).

Because NGS overcomes the poor resolution and need for the cumbersome probe definition step of tiling arrays, it can be easily applied to prob-

ing for various epigenomic modifications using ChIP-based approaches (Park, 2009) and chromatin accessibility using approaches relying on DNase I (deoxyribonuclease I) (Song and Crawford, 2010) or MNase (micrococcal nuclease) digestion (Valouev *et al.*, 2011). For instance, antibody-based methylation assays, such as meDIPS (Mohn *et al.*, 2009) and anti-CMS (Huang *et al.*, 2012b), are useful in identifying methylated regions and differentially methylated regions. Another interesting and important application where NGS has proven to be useful is a metagenomics (Wooley *et al.*, 2010). In a metagenomic study, the aim is to shed light on the composition of a biological sample through genomics, e.g., identification of different bacteria in a gut microbiota (Qin *et al.*, 2010).

4. Statistical Inference

Traditionally, nonparametric techniques are divided into methods which do not make assumptions about the distribution of the data, and modeling approaches in which the exact parametric structure of the relationship between variables is not fixed. The distribution-free view has been useful in deriving various nonparametric statistics, which can be utilized for traditional hypothesis testing purposes. However, in this thesis the focus is on nonparametric regression models, which belong to the latter class. Importantly, nonparametricity of a model does not imply that the model would be completely parameter-free. Instead, nonparametricity of a model states that the structure of the model is not fixed a priori; that is, the goal is to infer the model structure from data.

Several different methodologies for carrying out nonparametric regression modeling have been proposed; for instance, regression trees (Breiman *et al.*, 1984), regression splines (Friedman, 1991), wavelets (Wasserman, 2005), and various kernel-based methods (Simonoff, 1998). Often these methods rely on a set of weak assumptions about the underlying processes, such as smoothness and stationariness. Moreover, flexible nonparametric regression models are often described to be black box models because of their limited interpretability (Sjöberg *et al.*, 1995). Importantly, the interpretability usually decreases when the flexibility of the model increases. However, interpretability varies between different nonparametric modeling approaches. To give examples, neural networks are not interpretable due to the large number of neurons and possibly even multiple layers (Hassoun, 2003). Whereas, generalized additive models are more interpretable due to their linearity (Hastie and Tibshirani, 1986).

In statistical inference of a parametric regression model the goal is to identify, in some sense, the optimal values of the model parameters from

data that is subjected to random variation. Whereas, the model structure is also of interest in the inference of a nonparametric regression model. Moreover, an important subfield of statistical inference is Bayesian inference where the model parameters are assumed to be random variables. The Bayesian inference approach has been popular in inferring parametric and nonparametric models, such as nonparametric Dirichlet (Ferguson, 1973; Beal *et al.*, 2002) and Gaussian process models (Doob, 1944). In the following sections we will cover Gaussian processes (GPs) together with the parameter inference. Moreover, we will cover the model selection problem and practical issues of Bayesian inference.

4.1 Gaussian Process Prior

Gaussian processes, a family of stochastic processes, have proved useful in solving practical data-driven problems (Doob, 1944). Their usefulness in solving practical problems comes from the properties they share with the multivariate Gaussian random variables. Even though the assumption of Gaussianity might feel restrictive, GPs are fundamentally important and useful (Doob, 1944). For instance, while studying any stationary stochastic process involving only the first two moments, one can treat the variables as Gaussian (Doob, 1944). Additionally, in some cases the use of GPs in modeling can be motivated by resorting to the central limit theorem. Importantly, only a few theoretical properties are such that they hold for stationary GPs but not for stationary stochastic processes in general (Doob, 1944). More important, there are various stochastic processes motivated by physical phenomena, which are, essentially, Gaussian processes. To give examples, the Wiener process (Medhi, 1994), Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein, 1930), Brownian bridge process (Dudley, 2002), and fractional Brownian motion (Mandelbrot and Van Ness, 1968) are all Gaussian.

Recently, GPs have gained popularity in various machine learning problems. GPs have been utilized in classifying EEG (electroencephalography) signal patterns (Zhong *et al.*, 2008) and in segmenting annotation sequences (Altun *et al.*, 2004). In addition, they have been used to produce regression-based models for inferring transcription factor activities conditioned on gene expression measurements (Gao *et al.*, 2008) and for modeling biomass growth and the efficiency of nitrification (Ažman and Kocijan, 2007). Gaussian processes have been used to accelerate the Bayesian

treatment of nonlinear differential equations. To be more precise, instead of solving the dynamic system explicitly the authors use Gaussian processes for approximating the system behaviour (Calderhead *et al.*, 2009). Similarly, GPs together with variational approximation have proved useful in studying general stochastic differential equations by approximating their trajectories with a GP prior (Archambeau *et al.*, 2007).

Importantly, GPs do not scale well with the number of data points n due to the computational complexity $\mathcal{O}(n^3)$ of the conventional matrix inversion implementation. To overcome this limitation, several computationally more efficient approximative approaches have been proposed to enable Gaussian process inference to scale up (Trecate *et al.*, 1999; Csató and Opper, 2002; Smola and Bartlett, 2001). A subset of different approximative approaches are reviewed in (Quiñonero-Candela and Rasmussen, 2005; Quiñonero-Candela *et al.*, 2007). Finally, Chalupka *et al.* (2012) have defined an unbiased benchmark framework and studied the merits of different approximation techniques. In the next section we will define GPs formally.

4.1.1 Gaussian Process Definition

A stochastic process is defined as follows

Definition 1. A stochastic process f on an index set $\mathcal{X}$ is a collection of random variables $\{f_x,\, x \in \mathcal{X}\}$ (Parzen, 1987).

The Kolmogov extension theorem guarantees that a certain collection of finite-dimensional distributions will define a stochastic process

Theorem 1. *The Kolmogorov extension theorem guarantees that for a consistent family of finite-dimensional distributions* $\mathbb{P}_{f_{x_0,x_1,\ldots,x_{k-1}}}$ *for all positive* k *and* $x_i \in \mathcal{X}$, $i = 0, 1, \ldots, k-1$ *there exists a stochastic process* $\{f_x \in \mathcal{X}\}$*, which is consistent with this family (Gray and Davisson, 2005).*

The mean and covariance functions of a stochastic process are defined as follows

Definition 2. Let $\{f_x,\, x \in \mathcal{X}\}$ be a stochastic process with finite second moments, then its mean and covariance functions are $m(x) = \mathrm{E}[f_x]$ and $c(x_1, x_2) = \mathrm{E}\left[(f_{x_1} - m(x_1))\,(f_{x_2} - m(x_2))\right]$, respectively (Parzen, 1987).

A Gaussian random field is defined as follows

Definition 3. A real-valued Gaussian random field (GRF) is a random field f on an index set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}^{\mathrm{d}}$, such that the collection of random

variables $\{f_{x_1}, \ldots, f_{x_n}\}$ are multivariate Gaussian for each $1 \leq n < \infty$ and $(x_1, \ldots, x_n) \in \mathcal{X}^n$.

A real-valued Gaussian process, $f : \mathcal{X} \to \mathbb{R}$, is a special case of real-valued Gaussian random fields and it is defined as

Definition 4. A real-valued Gaussian process is a stochastic process f on an index set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, such that the collection of random variables $(f_{x_1}, \ldots, f_{x_n})$ are multivariate Gaussian for each $1 \leq n < \infty$ and $(x_1, \ldots, x_n) \in \mathcal{X}^n$.

Because in this thesis GPs are used instead of GRFs we present the required theory in the context of GPs. The following theorem guarantees that Gaussian processes are fully defined by the mean and covariance functions.

Theorem 2. *If $m : \mathcal{X} \to \mathbb{R}$ is a continuous linear functional and $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous bilinear nonnegative functional, then there is exactly one (up to equivalence) Gaussian process f whose mean function is $m_f = m$ and covariance function is $c_f = c$ (Denk* et al.*, 2003).*

To prove this one can use the Kolmogorov extension theorem after showing that all the finite-dimensional Gaussian distributions are fully parameterized by the mean vectors and covariance matrices.

The function-space view of Gaussian processes, where they can be seen as distributions over functions, is natural in the regression context. Formally, a Gaussian process, i.e., the distribution over function $f : \mathcal{X} \to \mathbb{R}$, is defined as

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right), \tag{4.1}$$

where the mean function $m(\mathbf{x})$ is used as a proxy for the unknown mean of the function

$$m(\mathbf{x}) = \mathrm{E}\left[f(\mathbf{x})\right], \quad \forall \mathbf{x} \in \mathcal{X} \tag{4.2}$$

and similarly the covariance function $k(\mathbf{x}, \mathbf{x}')$ is used to capture the covariances between the function values

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f(\mathbf{x}') - m(\mathbf{x}')\right)\right], \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{4.3}$$

The Wiener process which is a Gaussian process and the key component in many stochastic differential equation models, has the mean function $m(t) = 0$ and covariance function $k(t_1, t_2) = \min(t_1, t_2)$. In general, the exact forms of the mean and covariance functions of the studied process

are unknown (especially in the machine learning context). Therefore, the modeling includes the step of choosing the mean and covariance functions. Importantly, not all functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are valid covariance functions:

Definition 5. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a valid covariance function if it is symmetric, $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, and fulfills the nonnegative definiteness requirement

$$\int k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') \,\mathrm{d}\,\mu(\mathbf{x}) \,\mathrm{d}\,\mu(\mathbf{x}') \geq 0, \quad \forall f \in L_2(\mathcal{X}, \mu), \tag{4.4}$$

where μ is a measure on the input space $\mathcal{X}$ and L_2 is the set of square integrable functions.

The matrices generated using a nonnegative definite kernel, $[K]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, are nonnegative definite. Additionally, if the covariance function k can be written as a function of $\tau = \mathbf{x} - \mathbf{x}'$, then it is stationary. Finally, if it is a function of $|\tau|$, then it is isotropic.

4.1.2 Selection of Covariance Function

Selection of the covariance function is a fundamental step in Gaussian process modeling. This stems from the fact that the covariance function is used to assess the similarity between data points. Consequently, the covariance function encodes our assumptions about the properties of the underlying process of interest, such as stationarity, smoothness and periodicity (Rasmussen, 2004). Rasmussen (2004) discusses various covariance functions together with their properties and example applications, and presents the necessary theory for formulating new kernel functions.

In some cases, we might either know the exact covariance structure of the process or have some prior knowledge of it. More often, the covariance structure is completely unknown in practical problems. A manual inspection of the data can reveal some properties of the underlying process, and thus guide the selection of the covariance function (Shi and Choi, 2011). Unfortunately, the aforementioned approach is subjective in nature and it is limited in practice to one- and two-dimensional cases (Shi and Choi, 2011). Importantly, applicability of different covariance functions for a given modeling task can be also assessed in more quantified manner. Two commonly used approaches to do this are Bayesian model selection and cross-validation methods (Rasmussen, 2004). Different model selection techniques are discussed in more detail in Section 4.3.

4.1.3 Gaussian Processes in Regression Analysis

Next we will briefly describe how Gaussian processes can be applied for regression problems in which the aim is to estimate a mapping $f : \mathcal{X} \to \mathbb{R}$ based on the known input and output pairs. Gaussian processes provide a framework for fully probabilistic nonparametric regression, where the assumptions about the modeled process, such as smoothness, are taken into account by decoding them into the mean and covariance functions. Importantly, the Gaussian process regression framework is not limited to modeling mappings $f : \mathbb{R}^d \to \mathbb{R}$ as long as the mean and covariance functions can be defined in a meaningful way.

This paragraph follows the material presented by Rasmussen (2004). Let $\mathcal{D}$ be a list of length N containing the observed noiseless data consisting the input ($\mathbf{x}$) and output (f) pairs, i.e., $\mathcal{D} = ((\mathbf{x}_i, f_i)), i = 1, 2, \ldots, N$. Moreover, denote the list of inputs and outputs as X and $\mathbf{f}$, respectively. Then under the Gaussian process assumption, $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, the joint probability distribution is $\mathbf{f} \sim \mathcal{N}(\mathbf{m}(X), K(X, X))$, where $[\mathbf{m}(X)]_i = m(\mathbf{x}_i)$ and $[K(X, X)]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. To use the GP regression model for predicting the output $\mathbf{f}_*$ with the input X_* we first write the joint probability distribution of $\mathbf{f}$ and $\mathbf{f}_*$

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}(X) \\ \mathbf{m}(X_*) \end{pmatrix}, \begin{pmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix} \right), \tag{4.5}$$

and by conditioning on X, $\mathbf{f}$ and X_*

$$f|X, \mathbf{f}, X_* \sim \mathcal{N}(\mathbf{m}_{f|X,\mathbf{f},X_*}, K_{f|X,\mathbf{f},X_*}). \tag{4.6}$$

where $\mathbf{m}_{f|X,\mathbf{f},X_*} = \mathbf{m}(X_*) + K(X_*, X)K(X, X)^{-1}(\mathbf{f} - m(X))$ and $K_{f|X,\mathbf{f},X_*} = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)$.

The analytical tractability of GPs is not limited to the theoretical noise-free setting as the following example illustrates. Let us assume Gaussian process prior for the unknown function values f, $f|X \sim \mathcal{N}(\mathbf{m}(X), K(X, X))$, and the Gaussian i.i.d. noise model $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, where σ_n^2 is the noise variance. Then the likelihood of the observed data, $y = f(x) + \epsilon$, is $\mathcal{N}(\mathbf{f}, \sigma_n^2 I)$. Under this model, the predictive distributions can be expressed analytically similarly as in Equations (4.5) and (4.6). Moreover, the marginalization over all the possible Gaussian process realizations $\mathbf{f}$ is analytically tractable

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)\,\mathrm{d}\,\mathbf{f}. \tag{4.7}$$

4.1.4 Linear Transformations

Similarly as in the case of Gaussian random variables the Gaussianity of Gaussian processes is preserved under linear transformations. This property of GPs under linear transformations is reviewed in (Murray-Smith and Pearlmutter, 2005) with applications. Moreover, this property is convenient while infering latent variables which have undergone linear transformations and linear transformations can be used for placing contraints on the unknowns.

Consequently, the derivatives and integrals of Gaussian processes are still Gaussian processes. For example, the Gaussian process $\dot{f}(t)$ of the partial derivative with respect to t of a Gaussian process $f(t)$

$$\dot{f}(t) = \frac{\partial f(t)}{\partial t}, \tag{4.8}$$

is defined by the following mean function $m_{\dot{f}}(t)$, covariance function $k_{\dot{f},\dot{f}}(t,t')$ and cross-correlation function $k_{\dot{f},f}(t,t')$

$$m_{\dot{f}}(t) = \frac{\partial m(t)}{\partial t} \tag{4.9a}$$

$$k_{\dot{f},\dot{f}}(t,t') = \frac{\partial^2 k(t,t')}{\partial t \partial t'} \tag{4.9b}$$

$$k_{\dot{f},f}(t,t') = \frac{\partial k(t,t')}{\partial t}, \tag{4.9c}$$

which are needed for stating the predictive distribution for $\dot{f}(t)$. A linear integral transformation applied to Gaussian process $f(t)$ parameterized by $m_f(t)$ and $k_f(t,t')$ has the following form

$$g(t) = \int A(t,\tau) f(\tau)\ \mathrm{d}\,\tau, \tag{4.10}$$

where A is the kernel defining the integral transformation. To state the predictive distribution of the transformed Gaussian process $g(t)$ we need the mean function $m_g(t)$, covariance function of g, $k_{g,g}(t,t')$, and cross-correlation function between f and g, $k_{g,f}(t,t')$ (Ogorodnikov and Prigarin, 1996)

$$m_g(t) = \int A(t,\tau) m_f(\tau) d\tau, \tag{4.11a}$$

$$k_{g,g}(t,t') = \int\int A(t,\tau) k_{f,f}(\tau,\tau') A(t',\tau') d\tau\, d\tau', \tag{4.11b}$$

$$k_{g,f}(t,t') = \int A(t,\tau) k_{f,f}(\tau,t') d\tau. \tag{4.11c}$$

The presented differential and integral transformations generalize to higher dimensions.

4.2 Parameter Inference

In this section we briefly introduce the concepts of point estimates and posterior estimates, and their major conceptual differences. The aim of parameter inference is to construct an estimator for estimating a value of a given parameter based on the data. Importantly, this is usually done by stating some requirements that the optimal estimator should satisfy.

4.2.1 Point Estimation

Briefly, a point estimator calculates a single value as an estimate of the parameter of interest. In this section we will describe point estimators and the main concepts related to them.

An estimator $\hat{\theta}$ is an unbiased estimator of θ if $\mathrm{E}[\hat{\theta}] = \theta$ (the expectation is taken with respect to data) for every possible value of θ (Kay, 1993). That is, if the estimator $\hat{\theta}$ is unbiased, then its probability distribution is centered around the true parameter value θ. Importantly, depending on the problem there might not be an unbiased estimator, or there could be an unique or nonunique unbiased estimator (Kay, 1993). A related important concept to the unbiasedness is the consistency of an estimator. That is, if an estimator $\hat{\theta}$ is consistent, then it will convergence in probability to the true value of the parameter θ as the number of data points increases indefinitely (Kay, 1993). Notably, consistency and unbiasedness are different concepts. There are simple examples demonstrating that an unbiased estimator is not necessarily consistent and vice versa.

The variances of the unbiased estimators can be used as a measure for their "goodness", and thus used as a criterion to choose the estimator with minimal variance. Thus, the minimum variance unbiased estimators (MVUEs) are a well-studied and important class of estimators in practical problems (Kay, 1993). The Cramér-Rao inequality gives a lower bound for the variance of any unbiased estimator; that is, the variance is at least as high as the inverse of the Fisher information (the expected value of the observed information) (Kay, 1993). Clearly, if the estimator attains the Cramér-Rao lower-bound, then it has to be an MVUE estimator. Additionally, an estimator is called efficient if it attains the Cramér-Rao lower-bound for all the parameter values (Kay, 1993). Finally, various techniques exist for finding estimators based on different assumptions and requirements. To mention a couple of the techniques: the Rao-Blackwell theorem based on the use of sufficient statistics for de-

riving estimators, and the method of moments for finding representations of various distribution characteristics (Kay, 1993).

In some practical problems the aforementioned techniques to produce estimators are not applicable; thus, more generally applicable techniques are desired. For instance, the maximum likelihood estimators (MLEs) are generally applicable with many desired properties, although, MLEs do not have any optimality properties with finite sample sizes. Suppose the data $\mathcal{D}$ is distributed according to a probability density function $p(\mathcal{D}|\theta)$, then the likelihood function is defined as $\mathcal{L}(\theta;\mathcal{D}) = p(\mathcal{D}|\theta)$. Simply, the maximum likelihood estimator of θ is defined to be the parameter value which maximizes the likelihood $\theta_{\text{MLE}} = \operatorname{argmax}_\theta \mathcal{L}(\theta;\mathcal{D})$. Importantly, the maximum likelihood estimators have many convenient properties. For example, they are usually relatively easy to calculate. If an efficient estimator exists, then it is guaranteed to be the MLE. Additionally, they are invariant under one-to-one transformations of the parameters (Kay, 1993).

Next, let us consider an example of ML estimation in the context of Gaussian processes. First, let us assume a Gaussian process prior $p(\mathbf{f}|X,\theta)$ and Gaussian likelihood $p(\mathbf{y}|\mathbf{f},X,\theta)$. Then, the marginal likelihood obtained marginalizing over all the possible Gaussian process realizations f is analytically tractable

$$p(\mathbf{y}|X,\theta) = \int p(\mathbf{y}|\mathbf{f},X,\theta)p(\mathbf{f}|X,\theta)\,\mathrm{d}\,\mathbf{f}. \tag{4.12}$$

Unfortunately, even under this simple Gaussian model, the full Bayesian inference is not straightforward because the hyperparameters θ are not analytically marginable. Therefore, a widely used approach has been the type II maximum likelihood (ML-II) principle. ML-II produces a point estimator by maximizing the marginal likelihood with respect to the hyperparameters $\theta_{\text{ML-II}} = \operatorname{argmax}_\theta p(\mathbf{y}|X,\theta)$ (Berger, 1985). Often, the ML-II principle is referred to as the empirical Bayes principle in literature.

4.2.2 Bayesian Inference

The Bayesian and frequentist inference approaches have two major differences. First, instead of assuming that an unknown model parameter has a deterministic value, as in the frequentist inference, the Bayesian analysis considers the parameter to be a random variable with a probability distribution. The second difference concerns the analysis outcome. The Bayesian inference can produce a probability distribution describing

what is known about the parameter of interest given the data; whereas, the frequentist inference produces a conclusion in the form of "yes" or "no" derived, e.g., from a significance test. Importantly, neither of the conclusions derived using the frequentist hypothesis testing have an assigned probability of being correct or false. These differences in the Bayesian and frequentist inference originate from the different interpretations of probability. That is, the frequentists interpret the probability of an event as the limit of its relative frequency in repeated experiments. Whereas, in a Bayesian setting, probability is viewed as a degree of belief and is explicitly subjective. More detailed differences between Bayesian and frequentist approaches along with practical comparisons can be found, e.g., in (Bartholomew, 1965; Bayarri and Berger, 2004)

Bayes' theorem provides a way to connect the distributions of θ before and after obtaining data $\mathcal{D}$. First, let us denote the distribution of θ before observing any data (prior distribution), as $p(\theta|\alpha)$, where α is the hyperparameter of θ. After accounting for the data $\mathcal{D}$, the posterior distribution of θ, the updated prior, is

$$p(\theta|\mathcal{D},\alpha) = \frac{p(\mathcal{D}|\theta)p(\theta|\alpha)}{\int_\theta p(\mathcal{D}|\theta)p(\theta|\alpha)\,\mathrm{d}\,\theta} = \frac{p(\mathcal{D}|\theta)p(\theta|\alpha)}{p(\mathcal{D}|\alpha)}, \tag{4.13}$$

where $p(\mathcal{D}|\theta)$ is the distribution of $\mathcal{D}$ conditioned on θ and $p(\mathcal{D}|\alpha)$ is the marginal likelihood distribution of $\mathcal{D}$. Let us assume that we observe additional data $\mathcal{D}'$ after observing the data $\mathcal{D}$. Then we can use the posterior distribution $p(\theta|\mathcal{D},\alpha)$ as the prior distribution in Equation (4.13) and update it as $p(\theta|\mathcal{D},\mathcal{D}',\alpha)$ in the light of new data $\mathcal{D}'$. This procedure is referred as sequential Bayesian updating.

Often, the Bayesian analysis of models based on practical problems leads to analytically intractable quantities, because of the marginalization of the distributions while calculating the posterior distributions. For a long time, this limited the use of Bayesian methods. Luckily, the discovery of Markov chain Monte Carlo (MCMC) methods, with increased computing resources, enabled the use of Bayesian analysis in practical problems.

4.2.3 Markov Chain Monte Carlo

MCMC methods provide a general framework for solving high-dimensional integrals and optimization problems. Briefly, their operation is based on sampling a Markov chain whose equilibrium distribution is the target distribution. The use of various MCMC methods, especially in the context of

machine learning problems, is reviewed in (Andrieu *et al.*, 2003). In addition, Andrieu *et al.* (2003) list some problems where MCMC approaches have shown to be effective. For instance, MCMC methods have been used in Bayesian inference, optimization, statistical mechanics and penalized model selection.

Next we will briefly cover the basic idea behind the MCMC techniques. In a general form, we can state the problem of calculating the value of a high-dimensional integral

$$\mathcal{I} = \mathrm{E}_p[g(\theta)] = \int g(\theta)p(\theta)\,\mathrm{d}\,\theta, \tag{4.14}$$

where $g(\theta)$ is a function of θ and $p(\theta)$ is a target distribution (Robert and Casella, 2004). For instance, the target distribution could be a posterior distribution as in the Bayesian setting. The Monte Carlo integration techniques approximate the value of Equation (4.14) by the empirical average

$$\hat{\mathcal{I}} = \bar{g}_M = \frac{1}{M}\sum_{i=1}^{M} g(\theta^{(i)}), \tag{4.15}$$

where samples $\theta^{(i)}$, $i = 1, \ldots, M$ are sampled from $p(\theta)$ (Robert and Casella, 2004). When the samples $\theta^{(i)}$, $i = 1, \ldots, M$ are independent, then the strong law of large numbers guarantees that $\hat{\mathcal{I}}_M \to \mathcal{I}$ as $M \to \infty$ (Robert and Casella, 2004). Importantly, the form of $p(\theta)$ might be known only up to an unknown normalization constant; thus, independent sampling of $p(\theta)$ might not be feasible. However, to overcome this problem one can relax the requirement of drawing independent samples. That is, to construct a Markov chain where the next sample depends only on the current sample and is drawn using a transition kernel. If the Markov chain is ergodic, i.e., aperiodic, irreducible, and positive recurrent, and $\mathrm{E}[g(\theta)] < \infty$, then based on the ergodic theorem with probability 1

$$\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} g(\theta^{(i)}) = \int g(\theta)\pi(\theta)\,\mathrm{d}\,\theta, \tag{4.16}$$

where $\pi(\theta)$ is the stationary distribution (Robert and Casella, 2004). The previous remark gives the theoretical foundation of the various MCMC methods.

The two most widely used MCMC methods are Gibbs sampler and Metropolis-Hastings algorithm, along with various modifications (Robert and Casella, 2004). Briefly, the Gibbs sampler technique updates the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_N)^{\mathrm{T}}$ element by element iteratively. To be more precise, the j^{th} element is sampled from the conditional distribution, where the

posterior distribution $p(\theta|\mathcal{D})$ is conditioned on the other $N-1$ parameters

$$p(\theta_j|\theta_1^{(i)},\ldots,\theta_{j-1}^{(i)},\theta_{j+1}^{(i-1)},\ldots,\theta_n^{(i-1)},\mathcal{D}), \tag{4.17}$$

where the superscript denotes the iteration index (Robert and Casella, 2004). The advantage of Gibbs sampling approach is the "out of the box" functionality due to the lack of free parameters. But, the drawback is the requirement for sampling directly from the conditional distributions. If sampling from the conditional distribution is not feasible, then one can use the generally applicable Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm generates random walks by utilizing user-definable proposal distributions. That is, the proposal distribution $q(\theta^*|\theta^{(i)})$ is used for drawing samples based on the current state $\theta^{(i)}$ of the chain by applying the acceptance or rejection scheme (Smith and Roberts, 1993)

$$\theta^{(i+1)} = \begin{cases} \theta^* & \text{if } \frac{p(\theta^*|\mathbf{y})q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)}|\mathbf{y})q(\theta^*|\theta^{(i)})} > u \\ \theta^{(i)} & \text{otherwise,} \end{cases} \tag{4.18}$$

where $u \sim \mathcal{U}(0,1)$ is a sample from the uniform distribution (Robert and Casella, 2004). Due to the added flexibility, the choice of proposal distribution has an effect on the convergence of the chain. Thus obtaining an efficient sampler often involves manual tuning of the proposal distribution.

As discussed in the previous section, computation of posterior distributions in Bayesian inference often leads to intractable integrals. In many cases, the Monte Carlo approach is not feasible for those problems because the direct sampling from the target distribution is not possible. However, this limitation can be bypassed by using a MCMC method for constructing a Markov chain whose stationary distribution is the target distribution. The states of a Markov chain after it has converged to its stationary distribution can be treated as samples from the target distribution. Moreover, these samples can be used for estimating the target distribution or its characteristics, e.g., the expected value with a credible interval.

4.3 Model Selection

Model selection is the problem of selecting an optimal model from a set of considered models. For instance, determining the degree of a polynomial regression model based on its estimated predictive performance, or other

measures, is a model selection problem. Clearly, the definition of the optimality is subjective; additionally, one has to define the set of considered models. Thus, the problem of model selection is fundamental in nature.

Many different criteria and procedures based on various foundations for model selection have been proposed, including Bayes factor (Lavine and Schervish, 1999), Akaike information criterion (Akaike, 1974), Bayesian information criterion (Schwarz, 1978), Deviance information criterion (Spiegelhalter *et al.*, 2002), minimum description length (Hansen and Yu, 2001), bootstrap procedure (Shao, 1996) and cross-validation (Arlot and Celisse, 2010). In this thesis the focus is on the model selection using model posterior probabilities and Bayes factors.

4.3.1 Bayesian Model Selection

The aforementioned Bayesian analysis framework provides a natural and sound way to assess the degrees of belief of alternative hypotheses after observing data. So, let there be M alternative models whose prior probabilities are $p(\mathcal{M}_i),\, i = 1, \ldots, M$. Then, the posterior probability of the model $\mathcal{M}_k$ given the data $\mathcal{D}$ is

$$p(\mathcal{M}_k|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{i=1}^{M} p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}, \tag{4.19}$$

where $p(\mathcal{D}|\mathcal{M}_k)$ is the marginal likelihood of the model $\mathcal{M}_k$

$$p(\mathcal{D}|\mathcal{M}_k) = \int p(\mathcal{D}|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)\,\mathrm{d}\,\theta_k, \tag{4.20}$$

where the model specific parameters θ_k are marginalized out. The model posterior probabilities can be used to rank the models, and quantify the degrees of belief of the considered alternative models. Especially in the context of biochemical modeling, a selected set of methods for carrying out the Bayesian model selection are reviewed in (Vyshemirsky and Girolami, 2008).

Ensemble learning is a technique where multiple models are used simultaneously for obtaining better predictive performance (Dietterich, 2000). In some applications, it is more important to produce accurate predictions than to identify a single optimal model. There are various ensemble learning methodologies, such as Bayesian model averaging (BMA) (Hoeting *et al.*, 1999; Wasserman, 2000), error-correcting output (Dietterich and Bakiri, 1995), bagging (Breiman, 1996) and boosting (Duffy and Helmbold, 2002). Actually, BMA provides a natural way to carry out ensemble learning (Hoeting *et al.*, 1999); let there be M models $\mathcal{M}_k,\, k = 1, \ldots, M$,

then the posterior distribution of future observable Δ given data $\mathcal{D}$ is

$$p(\Delta|\mathcal{D}) = \sum_{k=1}^{M} p(\Delta|\mathcal{M}_k, \mathcal{D}) p(\mathcal{M}_k|\mathcal{D}), \tag{4.21}$$

which is the weighted average of the individual models, where the weights are defined by the model posterior probabilities as calculated in Equation (4.19).

Interestingly, the organizers of the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project studied the performance of ensemble learning in a realistic setting. They carried out an analysis where they blindly combined the online contest submissions inferring gene regulatory networks (Marbach *et al.*, 2012). Their conclusion was that none of the individual methods outperformed across the different data sets. More important, the constructed ensemble model yielded a highly robust performance across the data sets, illustrating the benefits of ensemble learning (Marbach *et al.*, 2012).

5. Temporal Modeling of Gene Expression

In this chapter we will cover Publication I and Publication III.

5.1 Temporal Modeling of Microarray Data

In this section we will cover the key ideas of the LIGAP method presented in Publication I. Briefly, LIGAP is a methodology for studying kinetic behaviour of gene expression in microarray data between any number of biological conditions. Coffey and Hinde (2011) reviewed regression, differential expression, discriminant, and clustering methods for analysis of time series microarray data. Moreover, various spline-based methods, generalized F-tests and hierarchical error, and empirical Bayes models have been presented for analyzing microarray data. For instance, ANOVA (analysis of variance) based methodology, TANOVA, defines different ANOVA structures and searches for the optimal one by evaluating the effects and significances of the factors without accounting for the temporal correlation (Zhou *et al.*, 2010). Whereas, a regression spline based method, EDGE, provides comparisons between arbitrary number of conditions, but does not quantify the differential expression in a condition-specific manner (Storey *et al.*, 2005). Finally, Stegle *et al.* (2010) proposed an approach based on GPs to determine the time windows where a gene is differentially expressed. Unfortunately this method is limited to analyzing only two conditions.

In Publication I, we use LIGAP to study kinetic gene expression profiles between activated $CD4^+$ T cells (Th0) and polarized $CD4^+$ T cells subsets Th1 and Th2. We present a temporal and nonstationary model based on the existence of temporal correlation between measurements from nearby time-points. The magnitude of the temporal correlation is inferred and modeled with a temporal Gaussian process regression model (Äijö *et al.*,

2012).

5.1.1 Nonstationary Time Series

Nonstationarity of a time series is determined based on the following criterion

Definition 6. A time series $x_1, x_2, \ldots, x_N$ is said to be nonstationary if $\exists m \in \{1, 2, \ldots, N\}$ so that the joint probability distribution of $x_i, x_{i+1}, \ldots, x_{i+m-1}$ depends on the time index i, otherwise the time series is stationary.

A less formal definition could be the following: if the time series is stationary, then its statistical characteristics do not change over time.

Especially if a cell population is perturbated using an external stimulus, such as activation of $CD4^+$ T cells upon antigen exposure, the cells undergo a rapid differentiation program, followed by a transition to an equilibrium (Lund *et al.*, 2003, 2007). To cover the strongly transient changes, time-course experiments are designed to more frequently collect samples during the beginning of the differentiation program. Therefore, the dynamic model used in the analysis should, ideally, model the nonstationarity in the time series. In the Gaussian process regression framework, this is achieved using a nonstationary covariance function, e.g., the neural network covariance function (Rasmussen, 2004)

$$k_{\mathrm{NN}}(t, t') = \frac{2}{\pi} \arcsin\left(\frac{2[1\ t]\Sigma[1\ t']^{\mathrm{T}}}{\sqrt{(1 + 2[1\ t]\Sigma[1\ t]^{\mathrm{T}})(1 + 2[1\ t']\Sigma[1\ t']^{\mathrm{T}})}} \right), \qquad (5.1)$$

where $\Sigma = \mathrm{diag}(l^{-2})$, the square brackets are used to denote row vectors, and l is the length-scale parameter. Alternatively, the length-scale parameter could be a function of time instead of treating it as a scalar in the squared exponential covariance function. Figure 5.1 compares the Gaussian process regression fits obtained using stationary and nonstationary covariance functions.

5.1.2 Model Definitions for LIGAP

Let us assume that there are N different biological conditions to compare. Then in theory, a gene could have from 1 to N distinct expression patterns. Importantly, at the limit $N = 2$, the problem simplifies to the traditional detection of differential expression between two time series. Collectively, the maximum number of distinct differential expression patterns among N conditions is given by the N^{th} Bell number, which gives the number of

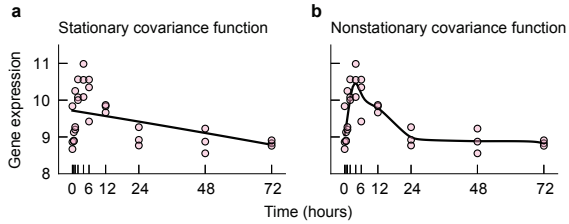

Figure 5.1. An example showing the difference between stationary and nonstationary Gaussian process regressions. (**a**) The dots correspond to the triplicate measurements taken at 0, 0.5, 1, 2, 4, 6, 12, 24, 48 and 72 hours. The black solid curve depicts the mean of the fitted Gaussian process regression model with the squared exponential covariance function. (**b**) As in (**a**) but here a nonstationary covariance function, the neural network covariance, is used in calculating the covariance matrix. In both cases the hyperparameters were selected so that the marginal likelihood was maximized.

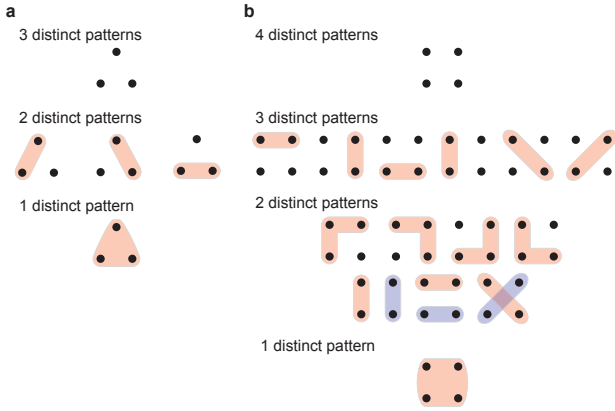

Figure 5.2. An example showing the partitions of the sets having three and four elements. (**a**) The sets with three elements have five different partitions, whereas the sets with four elements have 15 partitions as depicted in (**b**).

distinct partitions of a set (Comtet, 2010). However, in practice, genes do not express such a variety of expression patterns (Hornshøj *et al.*, 2007; Ramsköld *et al.*, 2009).

Figure 5.2 illustrates the partioning of experimental conditions in the context of differential expression analysis. In that example, all the possible distinct partitions of three or four experimental conditions are illustrated.

Let us denote the data corresponding to the condition j as $\mathcal{D}_j$ and over the conditions as $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$. Moreover, let models $\{\mathcal{M}_i\}$, $i = 1, \ldots, N_{\text{Bell}}$ correspond to the set of all the partitions of index set $\{1, \ldots, N\}$. Then, the likelihood of the data $\mathcal{D}$ given the model $\mathcal{M}_i$ and its parameters

θ_i by assuming independence is

$$p(\mathcal{D}|\mathbf{f}, \mathcal{M}_i, \theta_i) = \prod_{\mathcal{I} \in \mathcal{M}_i} p(\{\mathcal{D}_l\}_{l \in \mathcal{I}} | \mathbf{f}_{\mathcal{I}}, \theta_{\mathcal{I}}), \tag{5.2}$$

where the product is calculated over the disjoint subsets $\mathcal{I}$ defined by model $\mathcal{M}_i$, each of which have their own parameter set $\theta_{\mathcal{I}}$.

5.1.3 Model Posterior Distribution and Condition Specificities

Under the Gaussian noise model, which is widely used for log-transformed microarray intensity values, the marginalization over the function values f is analytically tractable. However, the marginalization over the hyperparameters θ is analytically intractable. Thus, we resort to the ML-II approach, where the hyperparameter values are selected to maximize the marginal likelihood in which f has been integrated out

$$\hat{\theta}^{\text{ML-II}} = \operatorname*{argmax}_{\theta} p(D|\mathcal{M}, \theta). \tag{5.3}$$

A natural approach to detect differential expression between conditions is to study the explanatory capabilities of the alternative models $\mathcal{M}_i$, $i = 1, \ldots, N_{\text{Bell}}$, which can be quantified by deriving the posterior distribution over the models. The posterior probability of the model $\mathcal{M}_i$ is

$$p(\mathcal{M}_i|\mathcal{D}, \hat{\theta}_i^{\text{ML-II}}) = \frac{p(\mathcal{M}_i) p(\mathcal{D}|\mathcal{M}_i, \hat{\theta}_i^{\text{ML-II}})}{\sum_{j=1}^{N_{\text{Bell}}} p(\mathcal{M}_j) p(D|\mathcal{M}_j, \hat{\theta}_j^{\text{ML-II}})}. \tag{5.4}$$

If no prior knowledge is available, then often the prior probabilities of the alternative models are assumed to be equal, $p(\mathcal{M}_i) = p(\mathcal{M}_j)$ for all i and j. In some applications, such as biomarker discovery, the specificity of the expression pattern is quantified to a given condition. For those purposes, we can define and calculate the following specificity score based on the model posterior probabilities

$$p(\text{“condition } j \text{ has distinct pattern”}) = \sum_{\mathcal{M}_i \in \{\mathcal{M}_k | \{j\} \in \mathcal{M}_k, k=1,\ldots,N_{\text{Bell}}\}} p(\mathcal{M}_i|\mathcal{D}, \theta_i^{\text{ML-II}}), \tag{5.5}$$

where the sum is calculated over the models in which $\mathcal{D}_j$ is modeled separately.

5.1.4 Summary of Results

The starting point in Publication I was previously published time-course measurements of Th0 (activated T helper), Th2 (T helper 2) cells (Elo *et al.*, 2010), and a previously unpublished time-course data set of Th1 (T

helper 1) cells. The human naïve CD4+ T cells were isolated from umbilical cord blood samples, activated, and polarized as described in (Elo *et al.*, 2010; Äijö *et al.*, 2012). Using the novel LIGAP methodology, we analyzed the three time-courses in parallel to gain further insight into the molecular mechanisms driving human T cell differentiation and function. The obtained gene lists contained both known and novel genes involved in T cell differentiation in humans, providing a valuable resource for biomarker purposes, as well as, further functional studies. Moreover, several novel genes were experimentally validated at the protein and gene expression levels. Interestingly, we identified a group of reciprocally regulated genes between Th1 and Th2 lineages. The group of reciprocally regulated genes included, among others, *IFNG* and *TBX21*, which are well-known Th1 signature genes known to suppress Th2 activity (Zhu *et al.*, 2010).

5.2 Temporal Modeling of Sequencing Data

In this section we cover the methodology presented in Publication III. DyNB is the first statistically sound methodology to study and detect differential temporal behavior of gene expression trajectories from RNA-seq data. In addition, it can be used to estimate differential differentiation efficiencies between conditions.

Various methodologies for identifying differentially expressed genes from RNA-seq data have been proposed. Some of the methodologies are reviewed and compared using simulated and real data in several recent review articles (Kvam *et al.*, 2012; Soneson and Delorenzi, 2013; Wesolowski *et al.*, 2013). For instance, the methodologies DESeq (Anders and Huber, 2010) and edgeR (Robinson *et al.*, 2010) have been widely used. DESeq and edgeR are based on the negative binomial distribution that has gained popularity in modeling biological variation in RNA-seq based gene expression data because of its ability to model overdispersion. Overdispersion occurs when the variance of the read count is significantly higher than the corresponding read count. DESeq and edgeR are mainly based on the same statistical assumptions, but the two methods differ in how normalization is done and overdispersion is estimated. DESeq estimates the overdispersion using a local regression approach based on a generalized linear model of the gamma family. Whereas, edgeR uses a maximum likelihood approach conditioned on the total read count per gene followed by

an empirical Bayes procedure for regularizing the dispersion estimates.

Clearly, the use of the negative binomial distribution leads to a non-Gaussian measurement model (see Publication III). Consequently, the posterior distribution of the considered Gaussian process model is analytically tractable. Therefore, the Bayesian model inference is carried out numerically using the Metropolis-Hastings algorithm. All the details are presented in Publication III. The presented methodology is used to analyze previously unpublished data sets of kinetic gene expression in activated $CD4^+$ human T and Th17 cells. The results demonstrate the effectiveness of this methodology.

5.2.1 Statistical Model of Read Counts

When RNA-seq was introduced, the count data was modeled using the Poisson distribution. But it was quickly observed that the assumption of Poissonity led to an underestimation of the intrinsic biological variance associated with high read counts between biological samples. The aforementioned intrinsic biological variance is caused by cell population hetererogeneity, intrinsic noise in transcription, and differences between the cell populations. Importantly, the negative binomial distribution allows more flexible modeling of the variance than the Poisson distribution. Therefore, it has been proposed that the negative binomial distribution is superior to the Poisson distribution for modeling sequencing read counts across biological samples (Anders and Huber, 2010; Robinson *et al.*, 2010)

$$Y \sim \mathrm{NB}(r, p), \tag{5.6}$$

where r is a predefined number of failures and $p \in (0, 1)$ is the probability of success. Moreover, the negative binomial distribution can be equivalently expressed using the mean μ and variance σ^2

$$\mu = \mathrm{E}[X] = \frac{pr}{1-p}, \quad \sigma^2 = \mathrm{Var}[X] = \frac{pr}{(1-p)^2} \tag{5.7}$$

because we can express p and r using the first two moments

$$p = \frac{\sigma^2 - \mu}{\sigma^2}, \quad r = \frac{\mu^2}{\sigma^2 - \mu}. \tag{5.8}$$

Then, let i be the gene of interest, j denote the sample of interest and ρ be a function that maps the sample index to the corresponding biological condition. Moreover, let $Y_{i,j}$ be the random variable representing the read count of the gene i in the sample j

$$Y_{i,j} \sim \mathrm{NB}\left(\mu_{i,j}, \sigma^2_{i,j}\right), \tag{5.9}$$

where we have used the mean and variance parameterization as shown in Equation (5.8). In practice, different samples do not have equal sequencing depths, and thus the raw read counts are not directly comparable. The samples, i.e., the read counts, are made comparable by using the sample specific size factors s_j. These size factors are estimated using the median of the ratios of the observed counts as presented by Anders and Huber (2010). Instead of scaling the raw observed read counts the scaling is applied to the condition and gene-specific parameter under estimation which is denoted as $q_{i,\rho(j)}$, that is

$$\mu_{i,j} = q_{i,\rho(j)} s_j, \tag{5.10}$$

where $q_{i,\rho(j)}$ represents the condition specific mean parameter, which is proportional to the unknown sample and gene-specific concentration. The estimation of the variances $\sigma^2_{i,j}$ directly per gene, and independently from the mean $\mu_{i,j}$, could be unstable because of the large biological variance and small number of replicates. Therefore, DESeq uses the following extended Poisson variance formulation

$$\sigma^2_{i,j} = \mu_{i,j} + s^2_j v(q_{i,\rho(j)}), \tag{5.11}$$

where $v : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a smooth function of $q_{i,\rho(j)}$. This function is estimated using a robust regression approach by assuming that the genes yielding similar read counts should have similar variances. More details can be found from (Anders and Huber, 2010).

5.2.2 Temporal Extension to Read Count Data Model

The current methodologies for studying differential gene expression in RNA-seq data do not take into account the temporal dimension. Therefore, in Publication III we presented a novel temporal methodology based on the statistical read count model presented by Anders and Huber (2010). The temporal dimension is incorporated into the analysis by assuming that the counts for a given gene are temporally correlated across the time series. In practice, this is achieved by generalizing the scalar parameter $q_{i,\rho(j)}$ to be a function of time, $q_{i,\rho(j)}(t)$. The function $q_{i,\rho(j)}(t)$ represents the kinetics of gene expression, i.e.,

$$\mu_{i,j}(t) = q_{i,\rho(j)}(t) s_j. \tag{5.12}$$

Consequently, the variance defined in Equation (5.11) is a function of time

$$\sigma^2_{i,j}(t) = \mu_{i,j}(t) + s^2_j v(q_{i,\rho(j)}(t)). \tag{5.13}$$

To estimate the smooth function $q_{i,\rho(j)}(t)$ we set a zero-mean Gaussian process prior with the squared exponential covariance function on $q_{i,\rho(j)}(t)$. That is, $q_{i,\rho(j)}(t) \sim \mathcal{GP}(0, k(t,t'))$. We denote $\mathbf{q}_{i,\rho(j)} = \{q_{i,\rho(j)}(t)\}_t$. First, the likelihood of the observed time series data $\mathbf{y} = \{y_{i,j}(t)\}$, where t, i and j denotes the time-point, gene and replicate, respectively, under the negative binomial likelihood assumption is

$$p(\mathbf{y}|q_{i,\rho(j)}, X, \theta) = \prod_{t\in X, j\in\{1,\ldots,M\}} \frac{\Gamma\left(y_{i,j}(t) + \xi(q_{i,\rho(j)}(t))\right)}{y_{i,j}(t)!\,\Gamma\left(\xi(q_{i,\rho(j)}(t))\right)} \times \left(1-\zeta(q_{i,\rho(j)}(t))\right)^{\xi(q_{i,\rho(j)}(t))} \zeta(q_{i,\rho(j)}(t))^{y_{i,j}(t)}, \tag{5.14}$$

where X is the set of time-points, θ is the set of hyperparameters (see Publication III) and

$$\xi(q_{i,\rho(j)}(t), j) = \frac{(q_{i,\rho(j)}(t)s_j)^2}{\sigma_{i,j}(t)^2 - q_{i,\rho(j)}(t)s_j}, \quad \zeta(q_{i,\rho(j)}(t), j) = \frac{\sigma_{i,j}(t)^2 - q_{i,\rho(j)}(t)s_j}{\sigma_{i,j}(t)^2}. \tag{5.15}$$

Importantly, under the negative binomial likelihood model the marginalization over the function values $\mathbf{q}_{i,\rho(j)}$ or over the hyperparameters θ is not analytically tractable. Therefore, a MCMC approach was used in the model inference as described later.

5.2.3 Inference of Differential Differentiation Efficiency

Observed variance between the samples within a condition $\rho(j)$ can be due to intrinsic and extrinsic stochasticity in gene expression. Alternatively, the observed variance could simply be due to a slight change in the experimental setting. The latter type of variance may be easier to detect and study, especially, if the effect is systematic and observable across the time series. Due to the possibility of having differences in experimental settings, we consider the possibility that the gene expression trajectories are accelerated or decelerated. Here the acceleration or deceleration could depend on biological variability or the properties of the given stimulus, such as its strength. This type of transformation can be expressed as t/k_j, $j = 1,\ldots,M$. Notably, one of the scaling factors k_j is set to 1 to make the model identifiable. An example of the time scaling is depicted in Figure 5.3.

In the model, the time scaling is taken into account through the Gaussian process $\mathbf{q}_{i,\rho(j)}(t)$ by introducing the replicate-specific time scaling parameter k_j. These time scaling parameters are used to scale the temporal dimension, i.e., $\mathbf{q}_{i,\rho(j)}(t/k_j)$. The scale parameters k_i are in practice un-

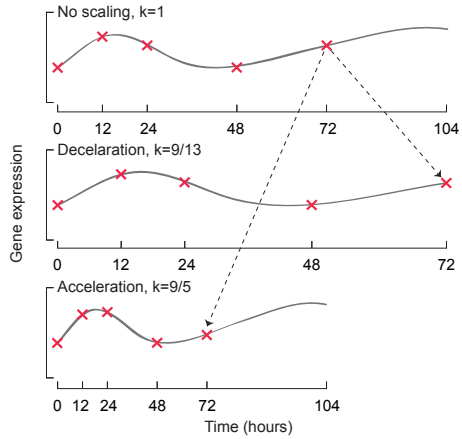

Figure 5.3. An example showing the effects of deceleration and acceleration. The gray lines show the unknown continuous gene expression trajectories and the red crosses depict the sampling times 0, 12, 24, 48 and 72 hours. In the case of $k = 9/13$ the 72 hour time-point corresponds roughly to the 50 hour time-point in the unscaled time axis (middle panel). Similarly, the 72 hour time-point in the unscaled time space is mapped to the 40 hour time-point with the scaling $k = 9/5$ (bottom panel).

known; thus, we first assign prior on them and then marginalize over them.

5.2.4 Posterior Inference of Temporal Dynamics

Like previously stated, under the negative binomial likelihood $p(\mathbf{y}|\mathbf{q}_{i,\rho(j)}, X, \theta)$ marginalization over the function values $\mathbf{q}_{i,\rho(j)}$ is analytically intractable. Thus, the use of the ML-II approch to optimize the hyperparameter is not as straightforward or attractive as in the case of Publication I. Consequently, the full Bayesian inference is used on the model. That is, we marginalize over the function values $\mathbf{q}_{i,\rho(j)}$, hyperparameters θ and time-scaling parameters k

$$p(\mathbf{y}|X) = \int p\left(\mathbf{y}|\mathbf{q}_{i,\rho(j)}, X\right) \left(\int \left(\int p\left(\mathbf{q}_{i,\rho(j)}|X, \theta, \mathbf{k}\right) p\left(\theta\right) \mathrm{d}\,\theta \right) p(\mathbf{k}) \,\mathrm{d}\,\mathbf{k} \right) \mathrm{d}\,\mathbf{q}_{i,\rho(j)}. \tag{5.16}$$

The traditional numerical quadrature approaches are not applicable due to multidimensionality of the integrals in Equation (5.16). Thus we resort to the Metropolis-Hastings algorithm (see Publication III). As an output we get the posterior distribution of the hyperparameters θ and time-scaling parameters k, and can estimate the marginal likelihood us-

ing, e.g., the harmonic mean estimator

$$p(\mathbf{y}|X) \approx \left(\frac{1}{m} \sum_{l=1}^{m} p(\mathbf{y}|X, \mathbf{q}^{(l)}, \theta^{(l)}, \mathbf{k}^{(l)})^{-1} \right)^{-1}, \quad \text{where } \mathbf{q}^{(l)}, \theta^{(l)}, \mathbf{k}^{(l)} \sim p(\mathbf{q}, \theta, \mathbf{k}|\mathbf{y}). \tag{5.17}$$

5.2.5 Quantification of Differential Dynamics

Let us assume that there are two conditions being compared, but the analysis could be generalized in the same way as presented in Publication I. The problem of detecting differential expression can be approached by formulating models $\mathcal{M}_0$ and $\mathcal{M}_1$ for nondifferential and differential expression, respectively. That is, $\mathcal{M}_0$ deals with the pooled data together with shared parameters, whereas $\mathcal{M}_1$ models the two data sets independently with independent parameters. After estimating the marginal likelihoods of the models and assuming equal priors we can calculate the Bayes factor

$$\mathrm{BF} = \frac{p(\mathbf{y}|X, \mathcal{M}_1)}{p(\mathbf{y}|X, \mathcal{M}_0)}, \tag{5.18}$$

which quantifies the relative evidence for differential expression.

5.2.6 Summary of Results

The DyNB methodology was originally developed for analyzing previously unpublished RNA-seq time-course data sets measured in human activated T cells and Th17 (T helper 17) cells. The initial DyNB analysis led to a conclusion that the signature genes, *IL17A* and *IL17F*, behaved systematically differently in one of the biological replicates. Using a version of the DyNB model with the addition of temporal scaling factors, we were able to identify widespread differences between that replicate and the others. The dynamic behaviour of the signature genes, as well as many others, were decelerated in this replicate. Moreover, computationally detected deceleration was experimentally verified by measuring selected signature Th17 genes in the same cells using a qRT-PCR assay. We speculated that the genes showing deceleration similar to the signature genes have a role in the human Th17 differentiation pathway. To assess the ability of DyNB to detect differential expression we compared it with DESeq which analyzes time-points independently (Anders and Huber, 2010). As expected, the gene lists overlapped significantly. The difference observed could result from the fact that DESeq does not

model the correlation between time-points or the differential differentiation efficiency and could be more sensitive to changes observed only at one time-point; hence, the differences observed in the gene lists. Together these results suggest that the DyNB methodology is able to detect differential expression between time series. More important, DyNB was able to estimate systematic differences between replicates, such as differential differentiation efficiency.

6. Modeling of Signal Transduction

In this chapter we cover the methodology presented in Publication II. Various machine learning techniques have been proposed for modeling signal transduction, such as Bayesian networks (Sachs *et al.*, 2005) and fuzzy logic (Aldridge *et al.*, 2009). Although these models are easy to handle, interpret and analyze computationally, they are hard to justify from a biophysical and mechanistic point of view. Consequently, models based on ordinary differential equations (ODEs) have been proposed because they allow definition of dynamic, continuous and mechanistic models (reviewed in (Aldridge *et al.*, 2006; Chakraborty and Das, 2010)). However, there are drawbacks with the ODE approaches: ODE models require a parametric formulation of the model, which can be tedious, difficult, and requires existing knowledge on the studied phenomenon. Nonetheless, ODEs have been applied successfully in various modeling applications, e.g., receptor-dependent mitogen activated protein kinase signaling (Chaudhri *et al.*, 2010) and the epidermal growth factor driven activation of the extracellular signal–regulated kinase signaling pathway (Xu *et al.*, 2010). Estimation of parameters of a nonlinear dynamic system, expressed using the differential equation system, is studied from the Bayesian point of view in (Girolami, 2008).

We propose a flexible nonparametric ODE model, which combines the interpretability of parametric models and the ease of nonparametric models. This is achieved by using a semiparametric ODE model with a nonparametric GP component representing the driving functions. Importantly, the regulatory functions do not have any explicit free parameters in this formulation. Consequently, we only have to estimate the parameters related to basal protein level and protein degradation, and the hyperparameters of the covariance function. The performance of the proposed algorithm is studied in identifying signaling pathways and learning their dynamics

from phosphoprotein data.

6.1 Dynamical Model of Signal Transduction

We base our model on a widely used ordinary differential equation model consisting of basal, regulatory and degradation processes (Barenco *et al.*, 2006)

$$\frac{\mathrm{d}\, x_i(t)}{\mathrm{d}\, t} = \alpha_i + f_i(t) - \lambda_i x_i(t), \tag{6.1}$$

where α_i and λ_i are the basal and degradation rates, respectively, and $f_i(t)$ is a driving function. The differential equation in Equation (6.1) can be solved to yield

$$x_i(t) = \frac{\alpha_i}{\lambda_i} + c_i e^{-\lambda_i t} + \int_0^t f_i(\tau) e^{-\lambda_i (t-\tau)}\, \mathrm{d}\, \tau, \tag{6.2}$$

where c_i depends on the initial state $x_i(0)$.

6.2 Nonparametric Extension

In practice the function f_i in Equation (6.2) is often unknown because the biological mechanisms of interest are not well understood. To overcome this we resort to the Gaussian process regression technique. This approach allows us to use data to carry out probabilistic and nonparametric inference on the regulatory function. Thus, we assign a zero-mean Gaussian process prior with the squared exponential covariance function on the regulatory function $f_i(t)$. Moreover, we know that $x_i(t)$ is also a Gaussian process because only a linear integral transformation is applied to $f_i(x)$ in Equation (6.2). Consequently, by assuming the squared exponential covariance function we can derive the analytic forms of the mean, covariance and cross-correlation functions required for the Gaussian process representation of $x_i(t) \sim \mathcal{GP}(m_{x_i}(t), k_{x_i}(t, t'))$ (see Publication II).

In Publication II, inference of the model was a two-step process as depicted in Figure 6.1. First, the Gaussian process model $x_i(t)$ is fitted to the experimental data, i.e., the condition dependent Gaussian processes $f_{i,j}(t)$ are estimated and the values α_i and λ_i are assumed to be constant across all conditions. Then, we approximate each inferred regulatory functions $f_{i,j}(t)$, $j = 1, \ldots, M$ using a single Gaussian process $g_i(\hat{\mathbf{x}}_i(t))$, where $\hat{\mathbf{x}}_i(t)$ is a vector consisting of the phosphoprotein activities of the putative regulators. This enabled us to model and consider many possible regulatory

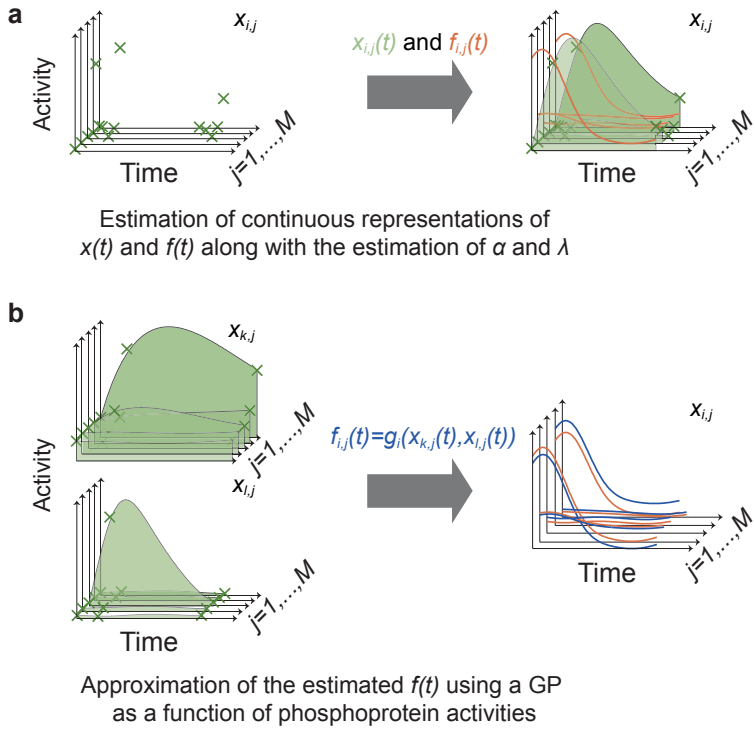

Figure 6.1. A schematic illustration of the two-step inference approach used in the Sorad methodology. (**a**) Based on the measurements (green crosses) continuous Gaussian process representation of $f_i(t)$ (orange lines) and $x_i(t)$ (green shaded area) are inferred for each of the conditions separately. (**b**) The inferred time-dependent driving functions f_i (orange lines) are approximated by a regulatory function $g_i(x_k, x_l)$ (blue lines), which is a function of the regulatory phosphoprotein activities, to enable the modeling of regulatory interactions.

interactions between phosphoproteins. The regulatory interactions, i.e., how $\hat{\mathbf{x}}_i(t)$ is defined, could be defined using existing biological knowledge. Alternatively, $\hat{\mathbf{x}}_i(t)$ could be defined using data-driven approaches, such as, cross-validation and Bayesian inference (see Publication II).

6.3 Solving Systems Trajectory

An attractive property of *in silico* models of biological systems is the ease and cost-effectiveness of simulating systems' behaviour under various conditions. For example, one could study how the system will behave over time under various knockouts or overexpression conditions. Let us assume that using a set of experimental data we have inferred the values of the parameters α_i and λ_i and functions $g_i(\hat{\mathbf{x}}_i(t))$. After defining the initial

state of the system, $\mathbf{x}(0)$, and the perturbations in a meaningful manner, one can simulate the system's behaviour using any numerical differential equation solver. The choice of the solver depends on the properties of the system, such as, variability in the time scales and stiffness. In our application, we found the Euler method for solving the systems trajectory was sufficient.

6.4 Inference for Interventions

One of the long-term goals of building models of various biological systems is to enable initial hypothesis testing in *in silico*. For example, an *in silico* model could be used to identify optimal interventions without the danger of severe harmful effects. Importantly, the Sorad framework depicted in Figure 6.1 allows intervention inference. Suppose a system has been learnt, i.e., the nonparametric functions and parameters g, α and λ have been inferred. In addition, let us classify the variables of the system to free and fixed variables. The free parameters $\mathbf{x}_{\text{free}}$ are the ones whose trajectories are initially undefined. Whereas, for the fixed parameters $\mathbf{x}_{\text{fixed}}$ we have a predefined target trajectory which we aim to achieve by controlling the free parameters. In other words, trajectories of the free parameters $\mathbf{x}_{\text{free}}$, or a subset of them, are estimated so that the system produces the desired behaviour specified by the fixed parameters $\mathbf{x}_{\text{fixed}}(t)$. An an example, one could use the following cost function

$$\mathcal{C} = ||x_{\text{fixed}}(t) - \hat{x}_{\text{fixed}}(t)||_2 + \beta||\operatorname{Var}[\hat{x}_{\text{free}}(t)]||_1, \tag{6.3}$$

where we minimize the Euclidean distance between the trajectories together with the weighted amount of uncertainty included in the estimated $\hat{x}_{\text{free}}(t)$. Note that the fitting of the fixed parameters $\mathbf{x}_{\text{fixed}}(t)$ is done by varying the free parameters $\mathbf{x}_{\text{free}}(t)$. Finally, by investigating the estimated behaviours $\hat{x}_{\text{free}}(t)$ one can see how the free parameters should be perturbated (i.e., intervened). Further details can be found from Publication II.

6.4.1 Summary of Results

In Publication II, we analyzed a part of a publicly available phosphoprotein data set (Alexopoulos *et al.*, 2010). The analyzed data subset was published as a challenge in DREAM (The New York Academy of Sciences, 2009). The data consisted of phosphoprotein levels in a human hepato-

cellular carcinoma cell line (HepG2) subjected to various perturbations; for instance, stimulation of cell surface receptors and inhibition of messenger molecules. The DREAM organizers provided a performance metric for comparing results, which takes into account the prediction in unseen conditions and the sparseness of the inferred network topology (The New York Academy of Sciences, 2009).

The Sorad methodology outperformed the methods originally utilized in the challenge (The New York Academy of Sciences, 2009) by improving prediction accuracy while maintaining a sparse network topology. Moreover, using an independent test data set we demonstrated the applicability of the proposed intervention prediction procedure. In addition, a careful inspection of the inferred model highlighted a putative role for IKK in activating AKT in TGFα stimulated cells. However, further experimental validation is required to verify this regulatory interaction and to explore its potential clinical implications. In conclusion, Sorad is applicable for learning structures for various biochemical systems. Importantly, Sorad can be used to predict the required perburbations to the system in order to modulate the output in a desired manner.

7. Studies on Transcriptional Regulation

In this chapter we will cover the main results of Publication IV.

7.1 High-resolution Mapping of Nucleosomes

In Publication IV we studied nucleosome positioning and its regulation in murine $CD8^+$ cytolytic lymphocytes (CTL) and $CD4^+$ subsets Th1 and Th2. Previous studies of nucleosome positioning in mouse and human cells have been based on nucleosome maps obtained with rather low coverage (approximately 10–100X) (Valouev *et al.*, 2011). Moreover, the studies in mammals have focused on average behaviour of nucleosomes at annotated genomic regions, such as promoters, instead of focusing on individual nucleosomes (Valouev *et al.*, 2011). Importantly, it has been proposed that reliable detection of changes in nucleosome occupancy and positioning could require at least 200-fold nucleosome core coverage (Chen *et al.*, 2013). Clearly, genome-wide mapping of nuclesomes with that coverage is not feasible at the moment. To overcome this limition, we utilized the recently published method, BEM-seq (Bacterial artificial chromosomes Enriched Mono-nucleosomal DNA Sequencing) (Yigit *et al.*, 2013). This targeted-sequencing approach enabled us to obtain high-resolution maps of nucleosomes and to subsequently identify differentially remodelled nucleosomes.

7.1.1 Experimental Approach

In our study, we decided to focus on nine loci spanning nine genes known to play an important role in T cell differentiation, *Prf1*, *Il4*, *Ifng*, *Eomes*, *Cd4*, *Cd8a*, *Tbx21*, *Il2ra*, and *Il2rb* (Pipkin and Rao, 2009). Each locus is approximately 200 kilobases in length, so we covered approximately 1.9 megabases of the mouse genome in total. The sequencing libraries

were generated from the BAC enriched mononucleosomal DNA treated with MNase (micrococcal nuclease) (Yigit *et al.*, 2013). The generated sequencing libraries were sequenced on the Applied Biosystems SOLiD 4 instrument, yielding approximately 9, 13, and 13 million properly paired reads in the Th1, Th2, and CTL cells, respectively. Each nucleosome in Th1, Th2, and CTL cells was sequenced an average of 1010, 1000 and 662 times, respectively. Finally, this led to 10–100 times greater nucleosome core coverage than in previously published studies on nucleosome positioning in mammals (Schones *et al.*, 2008; Valouev *et al.*, 2011; Teif *et al.*, 2012). The obtained high nucleosome core coverage enabled us to estimate the level of variation in nucleosome positions between cells, which allowed us to investigate the mechanisms behind nucleosome organization.

7.1.2 Identification of Differentially Remodelled Nucleosomes (DRNs)

First we calculated center-weighted nucleosome occupancy signals as described previously (Yigit *et al.*, 2013). Briefly, a Gaussian kernel is positioned on the centers of the mapped paired-end reads for calculating a kernel density estimation of the nucleosome center positions. To quantify the differences between cell-type specific nucleosome maps, we utilized a sliding-window based approach in which the detection limit for a significant difference was calibrated by comparing two technical replicates (see Publication IV). Using this approach, we next quantified DRNs between the cell types in a pair-wise manner. Surprisingly, only 556 DRNs were detected between the three cell types studied. Only 6% of the nucleosomes were detected to be differentially remodelled, which suggests that nucleosome organization was highly conserved between the three cell types studied.

Next, we investigated the biological relevance of the detected DRNs by an integrative analysis of the data. We made several findings which suggest a high biological significance of the DRNs. First, the distribution of DRNs between cell types and loci correlated with developmental biology and gene expression patterns. Second, DRNs were enriched at promoter and intergenic regions, thus linking DRNs to transcriptional regulation via promoters and enhancers. Third, DRNs correlated in a cell type specific manner with chromatin accessibility measured using DNase I activity assays (Agarwal and Rao, 1998; Pipkin *et al.*, 2007; Balasubramani *et al.*, 2010). Fourth, lineage-specific transcription factors showed a strong

enrichment at lineage-specific DRNs, which we will explain in more detail below.

7.1.3 Transcription Factor Binding Coincides with Nucleosome Depletion

First, the clustering of the exhibited nucleosome occupancies identified groups of DRNs showing similar nucleosome occupancy patterns over cell types. A motif enrichment analysis was performed to further investigate the biological role of the DRNs. That is, we computationally checked if certain transcription factor binding motifs were enriched within the identified groups of DRNs or not. The hypothesis is that DRNs are a result of competition between transcription factors and nucleosomes for DNA occupancy. The Runx and Gata motifs were enriched among the DRNs showing nucleosome depletion in CTL and Th2 cells, respectively (see Publication IV). Interestingly, the previously identified lineage-specifying factors GATA3 (Th2) and RUNX3 (CTL) recognize Gata and Runx motifs, respectively. Additionally, the enrichments of Gata and Runx motifs were mutually exclusively, which correlates with the expression of *Gata3* and *Runx3* in Th2 and CTL lineages, respectively (see Publication IV).

To validate the functionality of the identified Gata and Runx motifs at DRNs, we checked if those are occupied by GATA3 and RUNX3. To do this, we analyzed publicly available GATA3 and RUNX3 ChIP-seq data from the same cell types (Wei *et al.*, 2011; Lotem *et al.*, 2013). The overlay of DRNs and binding maps revealed a striking overlap between nucleosome depletion and transcription factor binding; GATA3 and RUNX3 binding coincided with nucleosome depletion in Th2 and CTL, respectively (see Publication IV). Furthermore, the binding sites not overlapping DRNs were mostly in regions which were nucleosome-free in all three cell types (see Publication IV).

Further experiments, e.g., titration of RUNX3 and GATA3, are necessary to determine whether the nucleosome depletion is the consequence of competition between transcription factors and nucleosomes for DNA occupancy or if additional mechanisms are involved.

8. Discussion

For a long time, most computational studies on biology focused solely on sequence and structure analysis. But molecular biology research has been revolutionized over the past twenty years by the development of various genome-wide assays. As a result, the exponential growth of data has attracted mathematicians and computer scientists to study biological questions. The past decade has demonstrated significant growth in data analysis and the application of computation to biology; for instance, it has become common for molecular laboratories to employ bioinformaticians. However, many of our expectations about the genome-wide techniques for improving our understanding the cell have not been satisfied. While it is impossible to judge whether or not those expectations were realistic, we have all been surprised by the complexity of the interactions that are involved in molecular biology.

Altogether three introductory chapters were presented for demonstrating the breadth of the field of computational biology research. First, an introduction to the selected concepts in functional genomics and epigenomics was presented to place this work into a broader context. We aimed to contextualize molecular biology relevant to our work and, at the same time, to provide a knowledge base for readers without a biology background. To connect experiments with data analysis, we presented a brief introduction to the first steps in the analysis of microarray and next-generation sequencing data. Specifically, we explained the main steps of quality control, alignment and normalization, that are required before any subsequent downstream analysis. In addition to making recommendations on how best to preprocess the data, we discussed some of the inherent caveats or biases that must be taken into account. We went on to introduce and discuss the statistical concepts that appear in the publications featured in this thesis. First, we discussed nonparametric methods

generally, as well as, formally defined Gaussian processes. Next, we elucidated specific techniques for parameter inference and model selection while keeping the focus on techniques that emphasize Bayesian concepts.

The transcriptional program of a cell is dynamic in nature and it largely determines its function, fate, and response to a stimulus. Moreover, dysregulation of these transcriptional programs can cause various diseases. This thesis was composed of three methodological publications and a cross-disciplinary publication in which we approached the transcription process and its regulation from different point of views. The LIGAP and DyNB methodologies improve the analysis of time series gene expression data and can be used to detect differential gene expression from microarray and RNA-seq data. However, these methodologies alone are unable to provide information about mechanisms of transcription regulation, or the consequences of changes in gene expression. Unfortunately, these functional and mechanistic questions are the type of questions we should focus on answering if the goal is to understand the role of genetics and epigenetics in disease. For that reason, the focus of future genome-wide studies should be on functional studies instead of screening gene expression landscapes of various cell types. Related to this aspect, we also studied chromating remodeling and its influence on transcription by mapping nucleosomes and correlating that information with transcription and transcription factor binding. Strikingly, only 6% of the nucleosomes were repositioned between Th1, Th2, and CTL cells within the nine loci studied, but, importantly, they correlated with the known transcription signatures, transcription factor binding, and chromatin accessibility. This data suggests that the chromatin accessibility is finely tuned between immune cell types. Moreover, this might be due to a competition between nucleosomes and transcription factors for DNA occupancy.

The presented dynamical models of gene expression could be extended in various ways; for example, they could account for alternative splicing and regulation of recruitment and elongation of RNAPII by promoter and enhancer activities. Hopefully, in the near future there will measurement technologies available for high-throughput screening of proteins levels. Based upon this information, it would be intriguing to model the interplay between transcription and translation, which could dramatically improve our understanding of translation and its regulation. Intriguingly, the presented novel time-scaling of DyNB revealed a widespread delay of Th17 differentiation, which could be used as an novel starting point to study

the different pathways involved in Th17 differentiation.

An obvious extension to the Sorad methodology would be the introduction of a link from signalling pathways to to the transcriptional level, and thus enabling the study of downstream effects of intracellular signaling. We briefly demonstrated how to use Sorad for predicting interventions for obtaining desired signaling pathway response. It would be interesting to further investigate the possible limitations of the presented approach and ways to improve it. Additionally, there are various possible clinical and experimental applications for the prediction of modulation strategies.

The study described in Publication IV can be followed up in various ways. For instance, one could study whether RUNX3 and GATA3 need additional cofactors to displace nucleosomes, why nucleosomes are dispaced only at a subset of binding motifs and does the concentration of RUNX3 and GATA3 correlate inversely with nucleosome occupancy at differentially repositioned nucleosomes. The same analysis could also be done for other transcription factors.

Undoubtedly it will still be beneficial to develop statistically sound methodologies for the analysis of individual data types. But as soon as possible more and more research work should aim at developing methodologies for integrating data types together. Of course this extremely important task is nontrivial, but it is essential for taking the knowledge about the cell to the next level. Additionally, this is also prerequisite for taking full advantage of the systems biology paradigm. We believe this is a problem in which bioinformaticians and computational biologists have much to contribute.

Despite many advances in experimental techniques and data analysis methods that have greatly increased knowledge in the cell biology field, the complexity of understanding cellular processes remains daunting. For instance, the role of most cellular proteins is still unknown. So far, our understanding of the complexity of the cell has gone hand-in-hand with advances in measurement technologies. It will be intriguing to see when the continued development of experimental techniques will change this trend. Undoubtedly, we need to also change our conceptual thinking of biology in order to overcome both present and future challenges. Additionally, this will also require great advancements in bioinformatics and computational biology. However daunting this task may be, we are excited about the potential for advancement both in the bioinformatic and computational biology fields. The near future is going to be an intriguing

time of advancement in the molecular biology field and due to the abundance and complexity of open questions in biology, bioinformaticians have become a part of molecular biology research.

9. Conclusion

The main contribution of this thesis has been the development of methodologies for time series analysis of various biological data. Importantly, implementations of all the developed methodologies have been made freely available for everyone to use.

The first methodology, LIGAP, provides a flexible framework for differential expression analysis between an arbitrary number of time series microarray data sets. To validate LIGAP it was compared with existing methods using previously published Th0 and Th2, as well as, unpublished Th1 genome-wide kinetic gene expression data sets. LIGAP's performance was proved when the simultaneous analysis of three T cell lineages identified genes known to be reciprocally regulated during T cell differentiation. Beyond previously validated genes, LIGAP was able to identify novel candidate genes that are differentially or even reciprocally regulated in these cell populations. These novel candidates, which we experimentally validated, could serve as useful biomarkers; but at the very least, the study results are a valuable transcriptome resource for future studies of early human T cell differentiation.

The second method, DyNB, is a statistically sound analysis framework for temporal RNA-seq data. This method can be viewed as a generalization of the Gaussian-Cox process, where the negative binomial distribution is utilized instead of the Poisson distribution. In the model, the Gaussian process component is used for modeling the mean of the read counts over time. The negative binomial distribution is utilized for modeling the distribution of RNA-seq read counts as previously proposed. This generalization enables a temporal analysis similar to LIGAP, but applicable to RNA-seq data. Importantly, DyNB is novel in its ability to quantitatively study differential differentiation efficiencies between biological replicates. This feature is especially useful for analysis of primary human

samples, because it accounts for intrinsic differences between samples or treatments. We used DyNB to analyze previously unpublished RNA-seq data from human Th0 and Th17 cells. To validate DyNB we compared it with existing methods. Importantly, the analysis revealed that Th17 differentation in one of the cultures was systematically delayed, which was independently validated using qRT-PCR. This systematic delay among a subset of genes could be valuable in studying different pathways involved in Th17 differentiation.

The third publication describes the Sorad methodology for analyzing measurements of phosphoprotein activity levels. Sorad uses a parametric ordinary differential equation model with a nonparametric Gaussian process component. We proposed a novel and efficient two-step scheme for model inference; first the system is analytically solved solely as a function of time, then the estimated regulatory function is approximated using a Gaussian process driven by regulatory factors. Additionally, we described and demonstrated how Sorad can be applied to predict modulation strategies, i.e., how the system should be perturbed in order to achieve the desired system behaviour. We benchmarked the performance of Sorad against other methods in predicting behaviour of the network under unseen conditions. The results from the comparison suggested that Sorad produced the most accurate results. Finally, our analysis with Sorad pinpointed a putative novel role for IKK in activating AKT in TGFα stimulated cells.

In the last publication we studied the role of chromatin structure on transcription. Specifically, we mapped nucleosome positions using BEM-seq in Th1, Th2, CTL cells across nine important genomic loci. The use of BEM-seq allowed us to map the nucleosomes with high-resolution and, consequently to identify computationally the differentially repositioned nucleosomes. Strikingly, only 6% of the nucleosomes showed differential occupancy or positioning between the cell types studied. the remodeled nucleosome map correlated with the previously available data describing More importantly, the remodelled nucleosomes correlated with the previously available data describing differentiation program, chromatin accessibility, and gene expression. An unbiased binding motif analysis suggested putative binding of key lineage-specifying factors, GATA3 and RUNX3, at differentially remodelled nucleosomes exhibiting nucleosome depletion in Th2 and CTL cells, respectively. The functional binding of GATA3 and RUNX3 at these motif positions was validated using ChIP-

seq data. Finally, our data supports the hypothesis that transcription factors and nucleosomes compete for DNA occupancy.

Bibliography

Agarwal, S. and Rao, A. (1998). Modulation of chromatin structure regulates cytokine gene expression during t cell differentiation. *Immunity*, **9**(6), 765–775.

Äijö, T., Edelman, S. M., Lönnberg, T., Larjo, A., Kallionpää, H., Tuomela, S., Engström, E., Lahesmaa, R., and Lähdesmäki, H. (2012). An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human t helper cell differentiation. *BMC Genomics*, **13**, 572.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**(6), 716–723.

Alberts, B. (2007). *Molecular Biology of the Cell*. Other, 5 edition.

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., and Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nat Cell Biol*, **8**(11), 1195–1203.

Aldridge, B. B., Saez-Rodriguez, J., Muhlich, J. L., Sorger, P. K., and Lauffenburger, D. A. (2009). Fuzzy logic analysis of kinase pathway crosstalk in tnf/egf/insulin-induced signaling. *PLoS Comput Biol*, **5**(4), e1000340.

Alexopoulos, L. G., Saez-Rodriguez, J., Cosgrove, B. D., Lauffenburger, D. A., and Sorger, P. K. (2010). Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Mol Cell Proteomics*, **9**(9), 1849–1865.

Altun, Y., Hofmann, T., and Smola, A. J. (2004). Gaussian process classification for segmenting and annotating sequences. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA. ACM.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106.

Andrews, A. J. and Luger, K. (2011). Nucleosome structure(s) and stability: variations on a theme. *Annu Rev Biophys*, **40**, 99–117.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, **50**(1-2), 5–43.

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. *Journal of machine learning research*, **1**, 1–16.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**(0), 40–79.

Ažman, K. and Kocijan, J. (2007). Application of gaussian processes for black-box modelling of biosystems. *ISA transactions*, **46**(4), 443–457.

Balasubramani, A., Shibata, Y., Crawford, G. E., Baldwin, A. S., Hatton, R. D., and Weaver, C. T. (2010). Modular utilization of distal cis-regulatory elements controls ifng gene expression in t cells activated by distinct stimuli. *Immunity*, **33**(1), 35–47.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-dna binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 28–37. ACM.

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol*, **7**(3), R25.

Bartholomew, D. J. (1965). A comparison of some bayesian and frequentist inferences. *Biometrika*, **52**(1-2), 19–35.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, **19**(1), 58–80.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden markov model. In *Advances in Neural Information Processing Systems 14*. MIT Press.

Beisel, C. and Paro, R. (2011). Silencing chromatin: comparing modes and mechanisms. *Nat Rev Genet*, **12**(2), 123–135.

Belmont, A. S., Dietzel, S., Nye, A. C., Strukov, Y. G., and Tumbar, T. (1999). Large-scale chromatin structure and function. *Curr Opin Cell Biol*, **11**(3), 307–311.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition.

Bergman, Y. and Cedar, H. (2013). Dna methylation dynamics in health and disease. *Nat Struct Mol Biol*, **20**(3), 274–281.

Blanchette, M. and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, **12**(5), 739–748.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., and Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**(5611), 1391–1394.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**(2), 123–140.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, 1 edition.

Calderhead, B., Girolami, M., and Lawrence, N. D. (2009). Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 217–224. MIT Press.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, **116**(4), 499–509.

Chakraborty, A. K. and Das, J. (2010). Pairing computation with experimentation: a powerful coupling for understanding t cell signalling. *Nat Rev Immunol*, **10**(1), 59–71.

Chalupka, K., Williams, C. K. I., and Murray, I. (2012). A framework for evaluating approximation methods for gaussian process regression.

Chaudhri, V. K., Kumar, D., Misra, M., Dua, R., and Rao, K. V. S. (2010). Integration of a phosphatase cascade with the mitogen-activated protein kinase pathway provides for a novel signal processing function. *J Biol Chem*, **285**(2), 1296–1310.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). Danpos: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res*, **23**(2), 341–351.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., and Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mrna sequencing. *Nat Methods*, **5**(7), 613–619.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res*, **38**(6), 1767–1771.

Coffey, N. and Hinde, J. (2011). Analyzing time-course microarray data using functional data analysis-a review. *Statistical Applications in Genetics and Molecular Biology*, **10**(1).

Comtet, L. (2010). *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Springer, softcover reprint of hardcover 1st ed. 1974 edition.

Cosgrove, M. S., Boeke, J. D., and Wolberger, C. (2004). Regulated nucleosome mobility and the histone code. *Nat Struct Mol Biol*, **11**(11), 1037–1043.

Cramer, P., Armache, K.-J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., Damsma, G. E., Dengl, S., Geiger, S. R., Jasiak, A. J., Jawhari, A., Jennebach, S., Kamenski, T., Kettenberger, H., Kuhn, C.-D., Lehmann, E., Leike, K., Sydow, J. F., and Vannini, A. (2008). Structure of eukaryotic rna polymerases. *Annu Rev Biophys*, **37**, 337–352.

Crawford, G. E., Davis, S., Scacheri, P. C., Renaud, G., Halawi, M. J., Erdos, M. R., Green, R., Meltzer, P. S., Wolfsberg, T. G., and Collins, F. S. (2006). Dnase-chip: a high-resolution method to identify dnase i hypersensitive sites using tiled microarrays. *Nat Methods*, **3**(7), 503–509.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, **107**(50), 21931–21936.

Csató, L. and Opper, M. (2002). Sparse on-line gaussian processes. *Neural Computation*, **14**(3), 641–668.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*, **7**(12), e1002384.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with glimmer. *Nucleic Acids Res*, **27**(23), 4636–4641.

Denk, G., Meintrup, D., and Schäffler, S. (2003). Transient noise simulation: Modeling and simulation of 1/f-noise. In *Modeling, simulation, and optimization of integrated circuits*, pages 251–267. Springer.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, **1857**, 1–15.

Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via Error-Correcting output codes. *Journal of Artificial Intelligence Research*, **2**, 263–286.

Doob, J. L. (1944). The elementary gaussian processes. *The Annals of Mathematical Statistics*, **15**(3), 229–282.

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. (2006). Reliability and reproducibility issues in dna microarray measurements. *Trends Genet*, **22**(2), 101–109.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, 2nd edition.

Duffy, N. and Helmbold, D. (2002). Boosting methods for regression. *Machine Learning*, **47**(2), 153–200.

Elo, L. L., Järvenpää, H., Tuomela, S., Raghav, S., Ahlfors, H., Laurila, K., Gupta, B., Lund, R. J., Tahvanainen, J., Hawkins, R. D., Oresic, M., Lähdesmäki, H., Rasool, O., Rao, K. V., Aittokallio, T., and Lahesmaa, R. (2010). Genome-wide profiling of interleukin-4 and stat6 transcription factor regulation of human th2 cell programming. *Immunity*, **32**(6), 852–862.

Emilien, G., Ponchon, M., Caldas, C., Isacson, O., and Maloteaux, J. M. (2000). Impact of genomics on drug discovery and clinical medicine. *QJM*, **93**(7), 391–423.

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nat Methods*, **9**(3), 215–216.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345), 43–49.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Res*, **8**(3), 175–185.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**(2), 209–230.

Fernández, M. and Miranda-Saavedra, D. (2012). Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res*, **40**(10), e77.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. J. (2013). Ensembl 2013. *Nucleic Acids Res*, **41**(Database issue), D48–D55.

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). mirdeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades. *Nucleic Acids Res*, **40**(1), 37–52.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, **19**(1), 1–67.

Gao, P., Honkela, A., Rattray, M., and Lawrence, N. D. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**(16), i70–i75.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using rna-seq. *Nat Methods*, **8**(6), 469–477.

Gaspar-Maia, A., Alajem, A., Meshorer, E., and Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nat Rev Mol Cell Biol*, **12**(1), 36–47.

Ghosh, D. and Qin, Z. S. (2010). Statistical issues in the analysis of ChIP-seq and RNA-seq data. *Genes*, **1**(2), 317–334.

Giorgi, F. M., Del Fabbro, C., and Licausi, F. (2013). Comparative study of rna-seq- and microarray-derived coexpression networks in arabidopsis thaliana. *Bioinformatics*, **29**(6), 717–724.

Girolami, M. (2008). Bayesian inference for differential equations. *Theoretical Computer Science*, **408**(1), 4–16.

Goecks, J., Nekrutenko, A., Taylor, J., and , G. T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, **11**(8), R86.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotechnol*, **29**(7), 644–652.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of encode. *Genome Biol Evol*, **5**(3), 578–590.

Gray, R. M. and Davisson, L. D. (2005). *An Introduction to Statistical Signal Processing*. Cambridge University Press.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C.,

de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, **328**(5979), 710–722.

Greer, E. L. and Shi, Y. (2012). Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet*, **13**(5), 343–357.

Grewal, S. I. S. and Jia, S. (2007). Heterochromatin revisited. *Nat Rev Genet*, **8**(1), 35–46.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, **458**(7235), 223–227.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, **28**(5), 503–510.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J. L., Root, D. E., and Lander, E. S. (2011). lincrnas act in the circuitry controlling pluripotency and differentiation. *Nature*, **477**(7364), 295–300.

Haab, B. B. (2001). Advances in protein microarray technology for protein expression and interaction profiling. *Curr Opin Drug Discov Devel*, **4**(1), 116–123.

Hall, J., Dennler, P., Haller, S., Pratsinis, A., Säuberli, K., Towbin, H., Walther, K., Walthe, K., and Woytschak, J. (2010). Genomics drugs in clinical trials. *Nat Rev Drug Discov*, **9**(12), 988.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, **96**(454), 746–774.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., and Kjems, J. (2013). Natural rna circles function as efficient microrna sponges. *Nature*.

Hardiman, G. (2004). Microarray platforms-comparisons and contrasts. *Pharmacogenomics*, **5**(5), 487–502.

Hassoun, M. (2003). *Fundamentals of Artificial Neural Networks*. A Bradford Book.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, **1**(3), 297–318.

He, L. and Hannon, G. J. (2004). Micrornas: small rnas with a big role in gene regulation. *Nat Rev Genet*, **5**(7), 522–531.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, **39**(3), 311–318.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243), 108–112.

Heller, M. J. (2002). Dna microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, **4**, 129–153.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**(4), 382–401.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, **7**(3), 200–210.

Hood, L. and Friend, S. H. (2011). Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nat Rev Clin Oncol*, **8**(3), 184–187.

Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform*, **11**(2), 181–197.

Hornshøj, H., Conley, L. N., Hedegaard, J., Sørensen, P., Panitz, F., and Bendixen, C. (2007). Microarray expression profiles of 20.000 genes across 23 healthy porcine tissues. *PLoS One*, **2**(11), e1203.

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., Guo, Y., Lu, Y., Zhou, C., Fan, D., Weng, Q., Zhu, C., Huang, T., Zhang, L., Wang, Y., Feng, L., Furuumi, H., Kubo, T., Miyabayashi, T., Yuan, X., Xu, Q., Dong, G., Zhan, Q., Li, C., Fujiyama, A., Toyoda, A., Lu, T., Feng, Q., Qian, Q., Li, J., and Han, B. (2012a). A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**(7421), 497–501.

Huang, Y., Pastor, W. A., Zepeda-Martínez, J. A., and Rao, A. (2012b). The anti-cms technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat Protoc*, **7**(10), 1897–1908.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–214.

Ibarra-Laclette, E., Lyons, E., Hernandez-Guzman, G., Perez-Torres, C. A., Carretero-Paulet, L., Chang, T. H., Lan, T., Welch, A. J., Juarez, M. J., Simpson, J., Fernandez-Cortes, A., Arteaga-Vazquez, M., Gongora-Castillo, E., Acevedo-Hernandez, G., Schuster, S. C., Himmelbauer, H., Minoche, A. E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Perez, S. A., de Jesus Ortega-Estrada, M., Cervantes-Luevano, J. I., Michael, T. P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V. A., and Herrera-Estrella, L. (2013). Architecture and evolution of a minute plant genome. *Nature*.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–945.

Ip, J. Y., Schmidt, D., Pan, Q., Ramani, A. K., Fraser, A. G., Odom, D. T., and Blencowe, B. J. (2011). Global impact of rna polymerase ii elongation inhibition on alternative splicing regulation. *Genome Res*, **21**(3), 390–401.

Jain, J., McCaffrey, P. G., Valge-Archer, V. E., and Rao, A. (1992). Nuclear factor of activated t cells contains fos and jun. *Nature*, **356**(6372), 801–804.

Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science*, **293**(5532), 1074–1080.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, **8**(1), 118–127.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). Dna-binding specificities of human transcription factors. *Cell*, **152**(1-2), 327–339.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**(7314), 430–435.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**(5830), 1484–1488.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory (v. 1)*. Prentice Hall, 1 edition.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome Res*, **12**(6), 996–1006.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). Bigwig and bigbed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**(17), 2204–2207.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small rnas in animals. *Nat Rev Mol Cell Biol*, **10**(2), 126–139.

Kitano, H. (2002a). Computational systems biology. *Nature*, **420**(6912), 206–210.

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, **295**(5560), 1662–1664.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res*, **40**(9), e69.

Klee, C. B., Crouch, T. H., and Krinks, M. H. (1979). Calcineurin: a calcium- and calmodulin-binding protein of the nervous system. *Proc Natl Acad Sci U S A*, **76**(12), 6270–6273.

Kogenaru, S., Qing, Y., Guo, Y., and Wang, N. (2012). Rna-seq and microarray complement each other in transcriptome profiling. *BMC Genomics*, **13**, 629.

Kohoutek, J. (2009). P-tefb- the final frontier. *Cell Div*, **4**, 19.

Kooistra, S. M. and Helin, K. (2012). Molecular mechanisms and potential functions of histone demethylases. *Nat Rev Mol Cell Biol*, **13**(5), 297–311.

Kramer, R. and Cohen, D. (2004). Functional genomics to new drug targets. *Nat Rev Drug Discov*, **3**(11), 965–972.

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**(11), 1571–1572.

Kvam, V. M., Liu, P., and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. *Am J Bot*, **99**(2), 248–256.

Lam, H. Y. K., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., Gerstein, M. B., Butte, A. J., Ji, H. P., and Snyder, M. (2012). Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol*, **30**(1), 78–82.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert,

J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and , I. H. G. S. C. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Methods*, **9**(4), 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, **10**(3), R25.

Latchman, D. S. (1997). Transcription factors: an overview. *Int J Biochem Cell Biol*, **29**(12), 1305–1312.

Lavine, M. and Schervish, M. J. (1999). Bayes factors: what they are and what they are not. *The American Statistician*, **53**(2), 119–122.

Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, **34**, 77–137.

Lewis, J. D. and Izaurralde, E. (1997). The role of the cap structure in rna processing and nuclear export. *Eur J Biochem*, **247**(2), 461–469.

Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, **128**(4), 707–719.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, **11**(5), 473–483.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, **18**(11), 1851–1858.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and , . G. P. D. P. S. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G. K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010). The sequence and de novo assembly of the giant panda genome. *Nature*, **463**(7279), 311–317.

Liou, J., Kim, M. L., Heo, W. D., Jones, J. T., Myers, J. W., Ferrell, Jr, J. E., and Meyer, T. (2005). Stim is a ca2+ sensor essential for ca2+-store-depletion-triggered ca2+ influx. *Curr Biol*, **15**(13), 1235–1241.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, **2012**, 251364.

Lotem, J., Levanon, D., Negreanu, V., Leshkowitz, D., Friedlander, G., and Groner, Y. (2013). Runx3-mediated transcriptional program in cytotoxic lymphocytes. *PloS one*, **8**(11), e80467.

Lund, R. J., Ylikoski, E. K., Aittokallio, T., Nevalainen, O., and Lahesmaa, R. (2003). Kinetics and stat4- or stat6-mediated regulation of genes involved in lymphocyte polarization to th1 and th2 cells. *Eur J Immunol*, **33**(4), 1105–1116.

Lund, R. J., Löytömäki, M., Naumanen, T., Dixon, C., Chen, Z., Ahlfors, H., Tuomela, S., Tahvanainen, J., Scheinin, J., Henttinen, T., Rasool, O., and Lahesmaa, R. (2007). Genome-wide identification of novel genes involved in early th1 and th2 cell differentiation. *J Immunol*, **178**(6), 3648–3660.

Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*, **9**, 34.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM review*, **10**(4), 422–437.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., , D. R. E. A. M. C., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nat Methods*, **9**(8), 796–804.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*.

Margueron, R. and Reinberg, D. (2011). The polycomb complex prc2 and its mark in life. *Nature*, **469**(7330), 343–349.

Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet*, **12**(10), 671–682.

Martincic, K., Alkan, S. A., Cheatle, A., Borghesi, L., and Milcarek, C. (2009). Transcription elongation factor ell2 directs immunoglobulin secretion in plasma cells by stimulating altered rna processing. *Nat Immunol*, **10**(10), 1102–1109.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, **7**, 29–59.

Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, **6**(5), 386–398.

McManus, C. J. and Graveley, B. R. (2011). Rna structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev*, **21**(4), 373–379.

Medhi, J. (1994). *Stochastic processes*. New Age International.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., le Noble, F., and Rajewsky, N. (2013). Circular rnas are a large class of animal rnas with regulatory potency. *Nature*.

Merkenschlager, M. and Odom, D. T. (2013). Ctcf and cohesin: Linking gene regulatory elements with their targets. *Cell*, **152**(6), 1285–1297.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, **11**(1), 31–46.

Mohn, F., Weber, M., Schübeler, D., and Roloff, T.-C. (2009). Methylated dna immunoprecipitation (medip). *Methods Mol Biol*, **507**, 55–64.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, **5**(7), 621–628.

Murray-Smith, R. and Pearlmutter, B. (2005). Transformations of gaussian process priors. In J. Winkler, M. Niranjan, and N. Lawrence, editors, *Deterministic and Statistical Methods in Machine Learning*, volume 3635 of *Lecture Notes in Computer Science*, pages 110–123. Springer Berlin Heidelberg.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and snp calling from next-generation sequencing data. *Nat Rev Genet*, **12**(6), 443–451.

Nikolov, D. B. and Burley, S. K. (1997). Rna polymerase ii transcription initiation: a structural view. *Proc Natl Acad Sci U S A*, **94**(1), 15–22.

Noble, D. (2002). The rise of computational biology. *Nat Rev Mol Cell Biol*, **3**(6), 459–463.

Noller, H. F. (1991). Ribosomal rna and translation. *Annu Rev Biochem*, **60**, 191–227.

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., and Nielsen, J. (2012). A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. *Nucleic Acids Res*, **40**(20), 10084–10097.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, A., Rilakovic, N., Ritland, C., Rosselló, J. A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Thompson, S. L., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P. K., Lundeberg, J., and Jansson, S. (2013). The norway spruce genome sequence and conifer genome evolution. *Nature*, **497**(7451), 579–584.

Ogorodnikov, V. A. and Prigarin, S. M. (1996). *Numerical modelling of random processes and fields: algorithms and applications*. VSP, Utrecht, The Netherlands.

Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H., and Nakatani, Y. (1996). The transcriptional coactivators p300 and cbp are histone acetyltransferases. *Cell*, **87**(5), 953–959.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven symposia in biology*, **23**, 366–370.

Okamura, H., Aramburu, J., García-Rodríguez, C., Viola, J. P., Raghavan, A., Tahiliani, M., Zhang, X., Qin, J., Hogan, P. G., and Rao, A. (2000). Concerted dephosphorylation of the transcription factor nfat1 induces a conformational switch that regulates transcriptional activity. *Mol Cell*, **6**(3), 539–550.

Okoniewski, M. J. and Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, **7**, 276.

Olins, D. E. and Olins, A. L. (2003). Chromatin history: our view from the bridge. *Nat Rev Mol Cell Biol*, **4**(10), 809–814.

Ong, C.-T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet*, **12**(4), 283–293.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From rna-seq reads to differential expression results. *Genome Biol*, **11**(12), 220.

Ozsolak, F. and Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**(2), 87–98.

Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: structural families and principles of dna recognition. *Annu Rev Biochem*, **61**, 1053–1095.

Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**(10), 669–680.

Parzen, E. (1987). *Stochastic processes*, volume 24. Society for Industrial and Applied Mathematics.

Patel, R. K. and Jain, M. (2012). Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**(2), e30619.

Pearson, H. (2006). Genetics: what is a gene? *Nature*, **441**(7092), 398–401.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for chip-seq and rna-seq studies. *Nat Methods*, **6**(11 Suppl), S22–S32.

Pipkin, M. E. and Rao, A. (2009). Snapshot: effector and memory t cell differentiation. *Cell*, **138**(3), 606.e1–606.e2.

Pipkin, M. E., Ljutic, B., Cruz-Guilloty, F., Nouzova, M., Rao, A., Zúñiga-Pflücker, J. C., and Lichtenheld, M. G. (2007). Chromosome transfer activates and delineates a locus control region for perforin. *Immunity*, **26**(1), 29–41.

Probst, A. V., Dunleavy, E., and Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol*, **10**(3), 192–206.

Proudfoot, N. J. (2011). Ending the message: poly(a) signals then and now. *Genes Dev*, **25**(17), 1770–1782.

Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009). Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37**(Database issue), D32–D36.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., , M. I. T. C., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**(7285), 59–65.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet*, **32 Suppl**, 496–501.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, **6**, 1939–1959.

Quiñonero-Candela, J., Rasmussen, C. E., and Williams, C. K. (2007). Approximation methods for gaussian process regression. *Large-scale kernel machines*, pages 203–224.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B. (2013). Rfecs: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol*, **9**(3), e1002968.

Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, **5**(12), e1000598.

Rasmussen, C. (2004). *Gaussian Processes in Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Redon, C., Pilch, D., Rogakou, E., Sedelnikova, O., Newrock, K., and Bonner, W. (2002). Histone h2a variants h2ax and h2az. *Curr Opin Genet Dev*, **12**(2), 162–169.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of dna binding proteins. *Science*, **290**(5500), 2306–2309.

Rhodes, D. R. and Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nat Genet*, **37 Suppl**, S31–S37.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). Gc-content normalization for rna-seq data. *BMC Bioinformatics*, **12**, 480.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edition.

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

Rosenfeld, J. A., Wang, Z., Schones, D. E., Zhao, K., DeSalle, R., and Zhang, M. Q. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, **10**, 143.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol*, **14**(5), R51.

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**(1), 55–72.

Rühlmann, A. and Nordheim, A. (1997). Effects of the immunosuppressive drugs csa and fk506 on intracellular signalling and gene regulation. *Immunobiology*, **198**(1-3), 192–206.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**(5721), 523–529.

Saunders, A., Core, L. J., and Lis, J. T. (2006). Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol*, **7**(8), 557–567.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**(5235), 467–470.

Scherf, M., Epple, A., and Werner, T. (2005). The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform*, **6**(3), 287–297.

Schilsky, R. L. (2010). Personalized medicine in oncology: the future is now. *Nat Rev Drug Discov*, **9**(5), 363–366.

Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*, **9**(3), 179–191.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**(5), 887–898.

Schwartz, Y. B. and Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet*, **8**(1), 9–22.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Shandilya, J. and Roberts, S. G. E. (2012). The transcription cycle in eukaryotes: from productive initiation to rna polymerase ii recycling. *Biochim Biophys Acta*, **1819**(5), 391–400.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, **91**(434).

Sharma, A., Singh, K., and Almasan, A. (2012). Histone h2ax phosphorylation: a marker for dna damage. *Methods Mol Biol*, **920**, 613–626.

Sharma, S., Ding, F., and Dokholyan, N. V. (2008). ifoldrna: three-dimensional rna structure prediction and folding. *Bioinformatics*, **24**(17), 1951–1952.

Sharma, S., Findlay, G. M., Bandukwala, H. S., Oberdoerffer, S., Baust, B., Li, Z., Schmidt, V., Hogan, P. G., Sacks, D. B., and Rao, A. (2011). Dephosphorylation of the nuclear factor of activated t cells (nfat) transcription factor is regulated by an rna-protein scaffold complex. *Proc Natl Acad Sci U S A*, **108**(28), 11381–11386.

Sharov, A. A., Piao, Y., Matoba, R., Dudekula, D. B., Qian, Y., VanBuren, V., Falco, G., Martin, P. R., Stagg, C. A., Bassey, U. C., Wang, Y., Carter, M. G., Hamatani, T., Aiba, K., Akutsu, H., Sharova, L., Tanaka, T. S., Kimber, W. L., Yoshikawa, T., Jaradat, S. A., Pantano, S., Nagaraja, R., Boheler, K. R., Taub, D., Hodes, R. J., Longo, D. L., Schlessinger, D., Keller, J., Klotz, E., Kelsoe, G., Umezawa, A., Vescovi, A. L., Rossant, J., Kunath, T., Hogan, B. L. M., Curci, A., D'Urso, M., Kelso, J., Hide, W., and Ko, M. S. H. (2003). Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol*, **1**(3), E74.

Shaw, J. P., Utz, P. J., Durand, D. B., Toole, J. J., Emmel, E. A., and Crabtree, G. R. (1988). Identification of a putative regulator of early T cell activation genes. *Science*, **241**(4862), 202–205.

Shendure, J. and Ji, H. (2008). Next-generation dna sequencing. *Nat Biotechnol*, **26**(10), 1135–1145.

Shendure, J. and Lieberman Aiden, E. (2012). The expanding scope of dna sequencing. *Nat Biotechnol*, **30**(11), 1084–1094.

Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data*. Chapman and Hall/CRC, 1 edition.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat Genet*, **38**(11), 1348–1354.

Simonoff, J. S. (1998). *Smoothing Methods in Statistics (Springer Series in Statistics)*. Springer.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). Abyss: a parallel assembler for short read sequence data. *Genome Res*, **19**(6), 1117–1123.

Sims, 3rd, R. J., Belotserkovskaya, R., and Reinberg, D. (2004). Elongation by rna polymerase ii: the short and long of it. *Genes Dev*, **18**(20), 2437–2468.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., Hjalmarsson, H., and Juditsky, A. (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, **31**(12), 1691–1724.

Smith, A. F. and Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**(1), 195–197.

Smith-Garvin, J. E., Koretzky, G. A., and Jordan, M. S. (2009). T cell activation. *Annu Rev Immunol*, **27**, 591–619.

Smola, A. J. and Bartlett, P. (2001). Sparse greedy gaussian process regression. In *Advances in Neural Information Processing Systems 13*, pages 619–625. MIT Press.

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

Smyth, G. K., Michaud, J., and Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**(9), 2067–2075.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, **14**, 91.

Song, J. S., Liu, X., Liu, X. S., and He, X. (2008). A high-resolution map of nucleosome positioning on a fission yeast centromere. *Genome Res*, **18**(7), 1064–1072.

Song, L. and Crawford, G. E. (2010). Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, **2010**(2), pdb.prot5384.

Spiegelhalter, S. D., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, **13**(9), 613–626.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J Comput Biol*, **17**(3), 355–367.

Steinbach, W. J., Reedy, J. L., Cramer, Jr, R. A., Perfect, J. R., and Heitman, J. (2007). Harnessing calcineurin as a novel anti-infective agent against invasive fungal infections. *Nat Rev Microbiol*, **5**(6), 418–430.

Stoltenburg, R., Reinemann, C., and Strehlitz, B. (2007). Selex–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng*, **24**(4), 381–403.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A*, **102**(36), 12837–12842.

Stormo, G. D. (2000). Dna binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, **403**(6765), 41–45.

Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev*, **12**(5), 599–606.

Struhl, K. and Segal, E. (2013). Determinants of nucleosome positioning. *Nat Struct Mol Biol*, **20**(3), 267–273.

Suganuma, T. and Workman, J. L. (2011). Signals and combinatorial functions of histone modifications. *Annu Rev Biochem*, **80**, 473–499.

Teif, V. B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.-P., Marth, C., Höfer, T., and Rippe, K. (2012). Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol*, **19**(11), 1185–1192.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

The New York Academy of Sciences (2009). In recomb regulatory genomics/systems biology/dream conference 2009. http://www.nyas.org/publications/ebriefings/Detail.aspx?cid=40d18ef4-6939-4deb-accc-b5e4516d78a0. Accessed Oct 2, 2013.

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

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**(5), 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat Protoc*, **7**(3), 562–578.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, **13**(1), 36–46.

Trecate, G. F., Williams, C. K., and Opper, M. (1999). Finite-dimensional approximation of gaussian processes. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 218–224. MIT Press.

Tremethick, D. J. (2007). Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*, **128**(4), 651–654.

Tsukada, Y.-i., Fang, J., Erdjument-Bromage, H., Warren, M. E., Borchers, C. H., Tempst, P., and Zhang, Y. (2006). Histone demethylation by a family of jmjc domain-containing proteins. *Nature*, **439**(7078), 811–816.

Turner, B. M. (2005). Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*, **12**(2), 110–112.

Tursz, T., Andre, F., Lazar, V., Lacroix, L., and Soria, J.-C. (2011). Implications of personalized medicine–perspective from a cancer center. *Nat Rev Clin Oncol*, **8**(3), 177–183.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, **36**(5), 823.

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature*, **474**(7352), 516–520.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V.,

Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Vignali, M., Hassan, A. H., Neely, K. E., and Workman, J. L. (2000). Atp-dependent chromatin-remodeling complexes. *Mol Cell Biol*, **20**(6), 1899–1910.

Vyshemirsky, V. and Girolami, M. A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, **24**(6), 833–839.

Wasserman, L. (2000). Bayesian model selection and model averaging. *J Math Psychol*, **44**(1), 92–107.

Wasserman, L. (2005). *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer.

Weake, V. M. and Workman, J. L. (2010). Inducible gene expression: diverse regulatory mechanisms. *Nat Rev Genet*, **11**(6), 426–437.

Wei, G., Abraham, B. J., Yagi, R., Jothi, R., Cui, K., Sharma, S., Narlikar, L., Northrup, D. L., Tang, Q., Paul, W. E., *et al.* (2011). Genome-wide analyses of transcription factor gata3-mediated gene regulation in distinct t cell types. *Immunity*, **35**(2), 299–311.

Wesolowski, S., Birtwistle, M. R., and Rempala, G. A. (2013). A comparison of methods for rna-seq differential expression analysis and a new empirical bayes approach. *Biosensors*, **3**(3), 238–258.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**(7199), 1239–1243.

Willingham, A. T., Orth, A. P., Batalov, S., Peters, E. C., Wen, B. G., Aza-Blanc, P., Hogenesch, J. B., and Schultz, P. G. (2005). A strategy for probing the function of noncoding rnas finds a repressor of nfat. *Science*, **309**(5740), 1570–1573.

Wilson, R. C. and Doudna, J. A. (2013). Molecular mechanisms of rna interference. *Annu Rev Biophys*, **42**, 217–239.

Wood, A. and Shilatifard, A. (2004). Posttranslational modifications of histones by methylation. *Adv Protein Chem*, **67**, 201–222.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol*, **6**(2), e1000667.

Woollard, P. M., Mehta, N. A. L., Vamathevan, J. J., Van Horn, S., Bonde, B. K., and Dow, D. J. (2011). The application of next-generation sequencing technologies to drug discovery and development. *Drug Discov Today*, **16**(11-12), 512–519.

Workman, J. L. and Kingston, R. E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem*, **67**, 545–579.

Wu, J. Q., Habegger, L., Noisa, P., Szekely, A., Qiu, C., Hutchison, S., Raha, D., Egholm, M., Lin, H., Weissman, S., Cui, W., Gerstein, M., and Snyder, M. (2010). Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc Natl Acad Sci U S A*, **107**(11), 5254–5259.

Xiang, Y., Zhu, Z., Han, G., Lin, H., Xu, L., and Chen, C. D. (2007). Jmjd3 is a histone h3k27 demethylase. *Cell Res*, **17**(10), 850–857.

Xu, T.-R., Vyshemirsky, V., Gormand, A., von Kriegsheim, A., Girolami, M., Baillie, G. S., Ketley, D., Dunlop, A. J., Milligan, G., Houslay, M. D., and Kolch, W. (2010). Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci Signal*, **3**(113), ra20.

Yankulov, K., Yamashita, K., Roy, R., Egly, J. M., and Bentley, D. L. (1995). The transcriptional elongation inhibitor 5,6-dichloro-1-beta-d-ribofuranosylbenzimidazole inhibits transcription factor iih-associated protein kinase. *J Biol Chem*, **270**(41), 23922–23925.

Yigit, E., Zhang, Q., Xi, L., Grilley, D., Widom, J., Wang, J.-P., Rao, A., and Pipkin, M. E. (2013). High-resolution nucleosome mapping of targeted regions using bac-based enrichment. *Nucleic Acids Res*, **41**(7), e87.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification chip-seq data. *Bioinformatics*, **25**(15), 1952–1958.

Zentner, G. E. and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol*, **20**(3), 259–266.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, **18**(5), 821–829.

Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*, **3**(9), 698–709.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biol*, **9**(9), R137.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006). Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, **38**(11), 1341–1347.

Zhong, M., Lotte, F., Girolami, M., and Lecuyer, A. (2008). Classifying EEG for brain computer interfaces using gaussian processes. *Pattern Recognition Letters*, **29**(3), 354–359.

Zhou, B., Xu, W., Herndon, D., Tompkins, R., Davis, R., Xiao, W., Wong, W. H., , I., to Injury Program, H. R., Toner, M., Warren, H. S., Schoenfeld, D. A., Rahme, L., McDonald-Smith, G. P., Hayden, D., Mason, P., Fagan, S., Yu, Y.-M., Cobb, J. P., Remick, D. G., Mannick, J. A., Lederer, J. A., Gamelli, R. L., Silver, G. M., West, M. A., Shapiro, M. B., Smith, R., Camp, 2nd, D. G., Qian, W., Storey, J., Mindrinos, M., Tibshirani, R., Lowry, S., Calvano, S., Chaudry, I., West, M. A., Cohen, M., Moore, E. E., Johnson, J., Moldawer, L. L., Baker, H. V., Efron, P. A., Balis, U. G. J., Billiar, T. R., Ochoa, J. B., Sperry, J. L., Miller-Graziano, C. L., De, A. K., Bankey, P. E., Finnerty, C. C., Jeschke, M. G., Minei, J. P., Arnoldo, B. D., Hunt, J. L., Horton, J., Cobb, J. P., Brownstein, B., Freeman, B., Maier, R. V., Nathens, A. B., Cuschieri, J., Gibran, N., Klein, M., and O'Keefe, G. (2010). Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proc Natl Acad Sci U S A*, **107**(22), 9923–9928.

Zhu, J., Yamane, H., and Paul, W. E. (2010). Differentiation of effector cd4 t cell populations (*). *Annu Rev Immunol*, **28**, 445–489.

Zlatanova, J. and Thakar, A. (2008). H2a.z: view from the top. *Structure*, **16**(2), 166–179.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD21/2014 Cho, Kyunghyun
Foundations and Advances in Deep Learning. 2014.

Aalto-DD49/2014 Lindh-Knuutila, Tiina
Computational Modeling and Simulation of Language and Meaning: Similarity-Based Approaches. 2014.

Aalto-DD80/2014 Toivola, Janne
Advances in Wireless Damage Detection for Structural Health Monitoring. 2014.

Aalto-DD105/2014 Parkkinen, Juuso
Probabilistic components of molecular interactions and drug responses. 2014.

Aalto-DD108/2014 Faisal, Ali
Retrieval of Gene Expression Measurements with Probabilistic Models. 2014.

Aalto-DD110/2014 Virtanen, Seppo
Bayesian latent variable models for learning dependencies between multiple data sources. 2014.

Aalto-DD120/2014 Bergström-Lehtovirta, Joanna
The Effects of Mobility on Mobile Input. 2014.

Aalto-DD127/2014 Zhang, He
Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis. 2014.

Aalto-DD138/2014 Sovilj, Dušan
Learning Methods for Variable Selection and Time Series Prediction. 2014.

Aalto-DD144/2014 Eirola, Emil
Machine learning methods for incomplete data and variable selection. 2014.

Aalto-DD 149/2014

ISBN 978-952-60-5884-9
ISBN 978-952-60-5885-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

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

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**