TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

COMPUTATIONAL METHODS FOR STUDYING EPIGENOMIC REGULATION

Thomas Faux

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 685 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2023

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

COMPUTATIONAL METHODS FOR STUDYING EPIGENOMIC REGULATION

Thomas Faux

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 685 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2023

University of Turku

Faculty of Technology
Department of Computing
Computational Biomedicine and Bioinformatics
Doctoral Programme in Technology

Supervised by

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

Dr Asta Laiho
Turku Bioscience Centre
University of Turku and Åbo Akademi,
Turku, Finland

Reviewed by

Dr Carl Herrmann
Health Data Science Unit,
Medical Faculty Heidelberg,
Heidelberg University, Germany

Docent Minna Ollikainen,
Institute for Molecular Medicine Finland
University of Helsinki,
Helsinki, Finland

Opponent

Docent Sami Heikkinen
Institute of Biomedicine,
University of Eastern Finland
Kuopio, Finland

ISBN 978-951-29-9188-4 (PRINT)
ISBN 978-951-29-9189-1 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)
Painosalama, Turku, Finland 2023

"No one who achieves success does so without the help of others.
The wise and confident acknowledge this help with gratitude."

Alfred North Whitehead

I would like to dedicate this work to Mari and Ellen.
I count them every day among the blessings in my life.

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computational Biomedicine and Bioinformatics
THOMAS FAUX: Computational Methods for Studying Epigenomic Regulation
Doctoral Dissertation, 132 pp.
Doctoral Programme in Technology
December 2022

ABSTRACT

In the nucleus, DNA is tightly wrapped around proteins in a structure called chromatin in order to protect it from degradation. Chromatin is composed of nucleosomes which are a structure of eight histones around which the DNA is wrapped. Nucleosomes can be modified by enzymes on amino acids located on their N-terminal tails. These modifications allow the chromatin to open and close in targeted regions, providing control over gene expression.

At present, chromatin immuno-precipitation (ChIP) and assay of transposase-accessible chromatin (ATAC) combined with high-throughput sequencing (ChIP-seq and ATAC-seq) are the major high-throughput methods allowing the study of histone modifications and genome-wide chromatin openness, respectively. Typically, ChIP-seq targets one histone at a time by enriching the histone-bound regions of the genome using immuno-precipitation, while ATAC-seq uses a transposase enzyme to cut the open chromatin into fragments of DNA. The DNA fragments obtained from both techniques can be sequenced and aligned against a reference genome. Once the location of the fragments is determined, the genome is scanned for significant enrichment in a process called peak calling. Differential analysis is then used to compare local enrichment-level variations between different biological conditions. Combining ChIP-seq and ATAC-seq data with other information, such as RNA-seq–derived transcriptomics data, can further help to build a comprehensive picture of the complex underlying biology. This work therefore focuses on the development of computational tools to help with the analysis of epigenomics research data.

In this thesis, a robust workflow for the differential analysis of ChIP-seq and ATAC-seq data is developed and evaluated against existing tools using one synthetic dataset, two biological ChIP-seq datasets and two biological ATAC-seq datasets. RNA-seq data is then further correlated with the detected peaks. An efficient replicate-driven visualisation tool is also proposed to visualise coverage of DNA fragments on the genome, which is compared to two existing tools, highlighting its efficiency. Lastly, two studies are presented showcasing the usefulness of the differential analysis approaches in extracting knowledge in a real-life biological setting.

KEYWORDS: ChIP-seq, ATAC-seq, epigenomics, transcriptomics, differential analysis, visualization

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Laskennallinen lääketiede ja bioinformatiikka
THOMAS FAUX: Laskennallisia menetelmiä epigenomisen säätelyn tutkimukseen
Väitöskirja, 132 s.
Teknologian tohtoriohjelma
Tammikuu 2023

TIIVISTELMÄ

Solun tumassa DNA on sen suojelemiseksi pakattu proteiinien ympärille, joka muodostaa kromatiiniksi kutsutun rakenteen. Kromatiini koostuu nukleosomeista, jotka rakentunut kahdeksasta erilaisesta histoniproteiinista, joiden ympärille DNA on kietoutunut. Nukleosomien N-terminaalipäiden aminohapot ovat entsymaattisesti muokattavissa ja nämä entsymaattiset muokkaukset mahdollistavat kromatiinin avautumisen ja sulkeutumisen, ja siten geenien ilmentymisen.

Kromatiini-immunopresipitaatio (ChIP) ja transposaasille avoimen kromatiinin eristys yhdistettynä syväsekvensointiin (ChIP-seq ja ATAC-seq) mahdollistavat histonimodifikaatioiden ja geenien ilmentymiselle avoimien genomin avoimien kohtien tutkimuksen. ChIP-seq -kokeessa immunopresipitaatiota käytetään poimimaan DNA:sta kohdat, joihin tietty histonikompleksi on näytteessä sitoutunut. ATAC-seq -kokeessa puolestaan leikataan näytteen DNA:sta talteen transposaasientsyymin avulla geenien ilmentymiselle avointa kromatiinia sisältävät kohdat. Kummallakin tekniikalla kerätyt DNA-pätkät voidaan sekvensoida ja niiden sijainti genomissa määrittää linjaamalla sekvenssit tunnettua genomia vasten. Tämän jälkeen voidaan niin kutsutulla piikkien tunnistusvaiheella määrittää genomin alueelta kohdat, joille linjatut sekvenssit genomissa keskittyvät. Tunnistettuja piikkikohtia voidaan myös verrata näyteryhmäkohtaisten eroavaisuuksien löytämiseksi. Jotta saadaan mahdollisimman kattava kokonaiskuva koeasetelmaan liittyvästä monitahoisesta molekyylibiologiasta, voidaan ChIP-seq- ja ATAC-seq -menetelmillä tuotettuja tietoja täydentää myös esimerkiksi RNA-sekvensoinnilla tuotettavalla tiedolla geenien ilmentymisestä.

Tässä väitöstyössä keskitytään laskennallisten työkalujen kehittämiseen, joilla voidaan analysoida edellä mainituilla tavoilla tuotettuja epigenomiikan tutkimuksen data-aineistoja. Väitöstutkimuksessa on kehitetty menetelmä ChIP-seq- ja ATAC-seq -aineistojen analyysiin. Kehitettyä menetelmää verrataan muihin olemassaoleviin työkaluihin käyttäen vertailuaineistona yhtä synteettistä data-aineistoa, sekä kahta biologista ChIP-seq -aineistoa sekä kahta biologista ATAC-seq -aineistoa. Tunnistettujen piikkien validointiin käytetään myös RNA-seq -aineistoa. Väitöstyössä on kehitetty myös visualisointityökalu, jolla on mahdollista tarkastella sekvensointikokeissa tuotettuja aineistoja genomilokaatiokontekstissa. Verrattuna aiempiin työkaluihin kehitetty työkalu mahdollistaa havainnollisen visualisaation myös aineistoille, joissa on paljon näytereplikaatteja. Väitöstyön kaksi viimeistä tutkimusta tuovat esille näyteryhmien välisiä eroja tutkivien menetelmien sovellustapoja biologisten tutkimuskysymysten ratkaisemiseksi.

ASIASANAT: ChIP-sekvensointi, ATAC-sekvensointi, epigenomiikka, transkriptomiikka, näyteryhmäerot, visualisaatio

Acknowledgements

I would like to thank my supervisors, Prof. Laura Elo, Dr Asta Laiho and Dr Kalle Rytkönen, for their patience and guidance throughout the PhD process. I am truly grateful for their support, their insights and the time they have willingly shared with me to help me complete my work. I would like to extend my gratitude to the collaborators that took part in this adventure; their contribution is the *sine qua non* to the successful completion of this work.

I will hold dear the friendships that I found along the way, and I would like to thank Ye Hong, Mehrad Mahmoudian, Esko Pakarinen, Xu Qiao, Damien Kaukonen, Aidan McGlinchey, Inna Starskaya, Maria Jaakkola, Niklas Paulin, Tapio Envall, Ankitha Shetty, Ning Wang and Johannes Smolander for the scientific discussions, the emotional support and their enthusiasm.

I was lucky to be part of the ENLIGHT-ten and would like to thank the early-stage researchers (Anna NTalli, Alyssa Silva, Natalie Edner, Martina Lubrano Di Ricco, Saumya Kumar, Luís Almeida, Marisa Says, Tomás Gomes, Nigatu Ayele, Miguel Tenorio and Narendra Dhele) as well as the organisers for this wonderful experience.

I am forever grateful to my wife, Mari Päiviö, for her unconditional support throughout the years needed to complete the thesis. Thank you for giving me the courage and the strength during the toughest moments and for reminding me to take it one step at a time. I would also like to extend my gratitude to her family for the love and support they showed towards me.

Je voudrais remercier mon Grand-père Marcel Paoli qui s'est malheureusement éteint avant que je ne puisse compléter ma thèse. Merci de ne pas avoir baissé les bras. Merci à Manon, Michele et Vincent Paoli de m'avoir accueilli dans leur famille et de m'avoir offert un foyer. Merci à Bruno Duguenet, Véronique Lecrénais-Paoli et Patrick Lecrénais de m'avoir aidé à traverser ma crise d'adolescence. Je tiens à remercier ma famille d'outre-mer Antoine, Gaetan, Amandine Dominguez, Nicolas Faux, Colette et Damien Guichard pour leurs support et les encouragements à distance.

J’ai une pensée pour Damien Beggiora, Charlène Dupré, Margaux Caron, Antoine Viéville et Hélène Picard pour leur soutiens depuis la France.

Finalement, je remercie ma fille Ellen qui m’a fait redécouvrir la vie à travers ses yeux et montré que les choses importantes sont les plus simples.

Thomas Faux

Table of Contents

Acknowledgements **6**

Abbreviations **10**

List of Original Publications **12**

1 Introduction **13**
1.1 Epigenome 13
1.2 Motivations and aims of the thesis 15
1.3 Structure of the thesis 15

2 Sequencing Technology and its Applications in the Study of the Regulation of the Genome via Epigenomics **16**
2.1 DNA sequencing 16
2.1.1 Sanger sequencing technology 16
2.1.2 Illumina sequencing technology 17
2.2 Chromatin immuno-precipitation followed by sequencing 18
2.3 Assay for transposase-accessible chromatin 18
2.4 Ribonucleic acid sequencing 18

3 Computational Methods for Studying the Regulation of the Genome via Epigenomics **20**
3.1 Quality control 20
3.1.1 Read quality values and trimming 20
3.1.2 Library complexity 20
3.1.3 GC content bias 20
3.1.4 Saturation analysis 21
3.2 Read alignment 21
3.3 Strategies to handle duplicated reads 21
3.4 Peak calling 22
3.5 Read counting 24
3.6 Normalisation 25
3.7 Visual inspection of reads 26
3.8 Fraction of read in peaks 26
3.9 Strand cross-correlation 26
3.10 Differential peak calling 27

4 Datasets **31**
4.1 Datasets used in Publication I 31

4.1.1 Yellow fever ATAC-seq and RNA-seq data 31
4.1.2 Inflammation response ATAC-seq and RNA-seq data 32
4.1.3 Rheumatoid arthritis ChIP-seq and RNA-seq data..... 32
4.2 Datasets used in Publication II 33
4.2.1 Butadiene ChIP-seq and ATAC-seq data 33
4.2.2 Gastric adenocarcinoma ChIP-seq data 34
4.3 Datasets used in Publication III 34
4.4 Datasets used in Publication IV 35

5 Results 36
5.1 Differential ChIP-seq and ATAC-seq peak calling with ROTS and comparison with existing tools (Publication I)..... 36
5.1.1 Synthetic data 36
5.1.2 Number of differential peaks..... 38
5.1.3 Breadth and intensity..... 39
5.1.4 Correlation between differential peaks and RNA-seq .. 42
5.2 Replicate-oriented visualisation of peaks with RepViz (Publication II) 45
5.2.1 Visualisation in the context of replicates 46
5.3 Dynamics of broad H3K4me3 marks in hypoxia (Publication III) 49
5.3.1 H3K4me3 and hypoxia in endometrial stromal cells 49
5.3.2 Differences in breadths and heights of H3K4me3 marks 51
5.3.3 Correlation of H3K4me3 marks with transcriptional fold-changes..... 52
5.4 Application of differential gene expression analysis (Publication IV)..... 53

6 Discussion 55

7 Summary of Publications 58

List of References 60

Original Publications 71

Abbreviations

ASCII	American Standard for Code Information Interchange
ATAC	Assay for Transposase-Accessible Chromatin
ATAC-seq	ATAC sequencing
BED	Browser-Extensible Data
BLAST	Basic Local Alignment Search Tool
BWA	Burrow Wheel Aligner
cDNA	complementary DNA
ChIP	Chromatin Immuno-Precipitation
ChIP-seq	ChIP sequencing
CUT&RUN	Cleavage Under Target and Release Using Nuclease
DB	Differential Binding
DEG	Differential Expression of Genes
DER	Duke Excluded Regions
DNA	DeoxyriboNucleic Acid
DNAse	DeoxyriboNucleAse
DNAse-seq	DNAse 1 hypersensitive site sequencing
DPC	Differential Peak Calling
DSC	Decidual Stromal Cells
ENCODE	ENCyclopaedia Of DNA Elements
ESF	Endometrial Stromal Fibroblasts
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FAIRE-seq	FAIRE sequencing
FDR	False Discovery Rate
FLS	Fibroblast-Like Synoviocytes
FOXP3	Forkhead bOX protein P3
FRiP	Fraction of Reads in Peaks
HOMER	Hypergeometric Optimisation of Motif EnRichment
HMM	Hidden Markov Model
IGV	Integrative Genomic Viewer
KDM	histone lysin DeMethylase
KMT	histone lysin MethylTransferase

KO	Knock-Out
LPS	LipoPolySaccharides
mRNA	messenger RNA
MACS2	Model-based Analysis of ChIP-Seq 2
MAQ	Mapping and Assembly with Quality
mrsFAST	micro-read substitution-only Fast Alignment Search Tool
MUSIC	MUltiScale enrIchement Calling for ChIP-seq
NGS	Next-Generation Sequencing
OA	Osteoid Arthritis
PCR	Polymerase Chain Reaction
PePr	Peak calling Prioritisation pipeline
RA	Rheumatoid Arthritis
RNA	RiboNucleic Acid
RNA-seq	RNA sequencing
SBS	Sequencing By Synthesis
SGS	Second Generation Sequencing
SICER	Spatial clustering for Identification of ChIP-Enriched Regions
SNP	Single Nucleotide Polymorphism
TGS	Third Generation Sequencing
TNF	Tumour Necrosis Factor
TPM	Transcript Per Million
T reg cell	regulatory T-cell
UCSC	University of California Santa Cruz
UHS	Ultra-High Signal
WT	Wild-Type
YFV	Yellow Fever Vaccine
ZINBA	Zero-Inflated Negative Binomial Algorithm

List of Original Publications

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

I Thomas Faux, Kalle T. Rytkönen, Mehrad Mahmoudian, Niklas Paulin, Sini Junttila, Asta Laiho & Laura L. Elo. Differential ATAC-seq and ChIP-seq peak detection using ROTS. *NAR Genomics and Bioinformatics*, 2021; Volume 3, Issue 3, Article number: lqab059.

II Thomas Faux, Kalle T. Rytkönen, Asta Laiho & Laura L. Elo. RepViz: a replicate-driven R tool for visualizing genomic regions. *BMC Research Notes,* 2019; Volume 12, Article number: 441.

III Kalle T. Rytkönen, Thomas Faux, Mehrad Mahmoudian, Mauris C. Nnamani, Taija Heinosalo, Antti Perheentupa, Matti Poutanen, Laura L. Elo & Günter P. Wagner. Histone H3K4me3 breadth in hypoxia reveals endometrial core functions and stress adaptation linked to endometriosis. *iScience,* 2022; Volume 25, Article number: 5.

IV Liisa Andersen, Alexandra Franziska Gülich, Marlis Alteneder, Teresa Preglej, Maria Jonah Orola, Narendra Dhele, Valentina Stolz, Alexandra Schebesta, Patricia Hamminger, Anastasiya Hladik, Stefan Floess, Thomas Krausgruber, Thomas Faux, Syed Bilal Ahmad Andrabi, Jochen Huehn, Sylvia Knapp, Tim Sparwasser, Christoph Bock, Asta Laiho, Laura L. Elo, Omid Rasool, Riitta Lahesmaa, Shinya Sakaguchi & Wilfried Ellmeier. The transcription factor MAZR/PATZ1 regulates the development of FOXP3+ regulatory cells. *Cell Reports*, 2019; Volume 29, Issue 13, Pages: 4447–4459.

1 Introduction

1.1 Epigenome

In most living organisms, genetic information is stored in the cell as a polymer of nucleotides called deoxyribonucleic acid (DNA), which was discovered by Friedrich Miescher in 1869 (Dahm 2008). Nucleotides are the building blocks of DNA and are composed of a sugar, a nitrogenous base and a phosphate (Levene 1919). The sugar and the phosphate constitute the backbone of a DNA molecule, which, in its stable state, forms a double helix (Watson and Crick 1953). The nitrogenous bases constitute the coding part of the DNA, in which there are four bases present: adenine, thymine, cytosine and guanine. The sequence of these bases encodes the genetic information of an organism, which is inherited by each new generation, and the genome itself is composed of both coding and non-coding regions, with the coding regions containing the genes, which are sequences of DNA that are translated into RNA (ribonucleic acid), which are in turn used as the template to produce proteins.

Gene expression is thus a process of synthetising proteins by reading genetic information stored in the DNA and involves two main stages, transcription and translation. This chain of events, which leads to protein synthesis, is largely considered the central dogma of molecular biology, as first stated by Crick (Crick 1958). Transcription is performed by the RNA polymerase enzyme, which binds to the DNA and reads the base sequence to produce a matching RNA sequence. The genes coding for proteins produce messenger RNA (mRNA), which is read by a ribosome as a template for protein synthesis, also known as translation. The term ‘transcriptomics’ is used to describe the study of the RNA that is produced after transcription in a cell or a group of cells, and understanding the transcription is central to the ability to interpret an organism’s response to stress or disease through its gene expression. Transcriptomics provides a holistic view of gene expression, gene pathways and the interconnections and regulations involved.

The term ‘epigenetics’ refers to the study of heritable alterations of the phenotype that are not due to modifications in the DNA sequence. There are two types of epigenetic modifications, long-term and stable DNA methylation and short-term and reversible histone modifications (Handy, Castro, and Loscalzo 2011). In eukaryotes, the DNA is stored tightly packed in the nucleus to protect it from

degradation, and in order to pack the DNA, the double-stranded helix wraps around a protein complex to form a nucleosome. The nucleosome core is 146 base pairs of DNA wrapped around an octamer of proteins called histones (Kornberg 1974). When the DNA is in complex with proteins, it is called chromatin, which can be in different states of compaction, given the need for transcription—DNA is more accessible when there is a need for transcriptional activity and more compacted otherwise.

Epigenetic modifications, such as histone modifications, affect the state of chromatin compaction and regulate gene expression by allowing or denying access to the transcription machinery. This is done by recruiting histone-specific enzymes that will perform modifications on the N-terminal tails. The term 'histone code' (Strahl and Allis 2000) was coined to represent the possible modifications and their meanings, and there are more than 60 possible histone N-terminal residues that can be modified (Kouzarides 2007). While it is assumed that an individual histone tail modification leads to a biological consequence, it is a highly complex process, and the chances are high that certain effects are specific to particular combinations of histone tail modifications (Kouzarides 2007). A better understanding of the histone modifications and their combinations would enable a better understanding of many fundamental biological processes, including gene expression/repression, DNA repair and DNA replication. The complexity of the biological processes and the methods used to measure the presence of histone modifications or chromatin openness can lead to varying results; to ensure the best possible biological interpretations, there is therefore a need to constantly optimise and update the methods and tools of data analysis.

The goal of a chromatin state analysis is to identify different chromatin states under varying conditions. In comparative epigenomics, the emphasis is on the differences between the chromatin states of different samples. As an example, the chromatin state (e.g., open or closed) can differ between control group samples and groups exposed to certain conditions (e.g., hypoxia, rheumatoid arthritis), while gene expression analysis allows the monitoring of mRNA levels in sample groups based on the conditions to which they are exposed. Combined, these two methods offer a holistic view of the chromatin regulatory mechanisms and their effects on the expression of genes.

While researchers most often try to measure biological variations, the data analyst seeks to reduce technical variations introduced by the measurement tools. In high-throughput sequencing experiments, one way to reduce this variation is to use biological replicates, and over the last decade, emphasis has been placed on using the growing number of replicate samples to generate increasingly reproducible results.

1.2 Motivations and aims of the thesis

Large-scale epigenomic studies involving multiple different conditions and large numbers of replicate samples are already widely available (Akondy et al. 2017; Ai et al. 2018; S. H. Park et al. 2017), as are a range of tools covering different aspects of the relevant data analyses. In many cases, the choice of tool for a specific data analysis step heavily influences the results of the overall study. For example, there is no consensus regarding the best tool for each possible situation at the differential peak calling step (Steinhauser et al. 2016), and tools developed for differential peak calling on ChIP-seq data have not yet been extensively examined with ATAC-seq data, despite being widely used with this data type; a comparison study is therefore needed.

In a similar fashion, many established tools developed for visualising sequencing data are available, but with the recent increase in the number of replicate samples typically used in studies, the need for replicate-focused visualisation is growing.

To address these needs, the specific aims of this thesis are as follows:

1. Development of a robust and competitive workflow for differential peak calling.
2. Development of a visualisation tool for replicate-driven sequencing data visualisation.
3. Application of differential data analysis tools to biological research questions.

Publication I covers the development and evaluation of a differential peak calling tool, **Publication II** covers the development and evaluation of a visualisation tool for next-generation sequencing data and **Publications III** and **IV** showcase the use of differential analysis to extract biologically relevant information.

1.3 Structure of the thesis

A short history and description of the basics of the methods related to the production of sequencing data are explained in Chapter 2. The computational analysis of epigenomics-related data is detailed in Chapter 3, and the datasets used in the publications are described in Chapter 4. The results obtained in publications are described and discussed in Chapter 5, and the discussion and summary are available in Chapters 6 and 7. Reprints of the publications are at the end of the thesis.

2 Sequencing Technology and its Applications in the Study of the Regulation of the Genome via Epigenomics

2.1 DNA sequencing

2.1.1 Sanger sequencing technology

DNA sequencing was first introduced in 1977 (Sanger, Nicklen, and Coulson 1977) and remained the main sequencing method for the next 30 years. The method uses sequencing by synthesis (SBS); the sequencing starts with the synthesis of a primer on the DNA fragment of interest in order to provide DNA polymerase with a region of the DNA fragment upon which to bind. DNA polymerase is an enzyme that catalyses the synthesis of a new DNA base on a fragment of DNA using a nucleotide, and in the presence of the four deoxynucleotides (A, T, C and G), it elongates the DNA fragment and is an essential component of DNA replication. Di-deoxynucleotides (chain-terminating deoxynucleotides) are added in low concentrations in separated containers; the DNA polymerase then randomly stops elongating the DNA fragments, and we obtain fragments of different sizes with known terminating bases. The fragments are then deposited on one of four parallel electrophoresis lanes on a polyacrylamide gel that correspond to the A, T, C and G terminating di-deoxynucleotides; after electrophoresis, the gel contains the sequence of the DNA fragment (Figure 1).

In 1990, the Human Genome Project was launched with the goal of sequencing the whole human genome within 15 years. By the end of the project ("Initial Sequencing and Analysis of the Human Genome" 2001; "Finishing the Euchromatic Sequence of the Human Genome" 2004), the method had been developed to maturity and achieved its full capacity, making it unsuitable for the scaling-up needed for personalised genomic medicine and the required sequencing of millions of genomes (Barba, Czosnek, and Hadidi 2014).

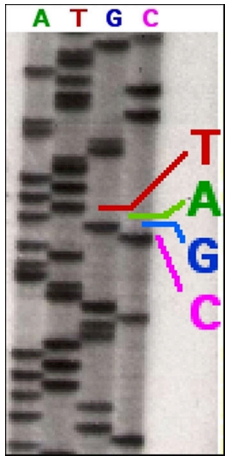

Figure 1. Piece of a radioactively labelled electrophoresis gel. Due to the randomness in elongation length, the fragments for each lane will migrate to different positions, allowing the user to read the DNA sequence of the DNA fragment. (John Schmidt: https://commons.wikimedia.org/wiki/File:Sequencing.jpg)

2.1.2 Illumina sequencing technology

The emergence of next-generation sequencing (NGS) platforms at the beginning of the 21st century opened the door to the scaling needed to develop personalised genomic medicine (Esplin, Oei, and Snyder 2014). NGS refers to the range of sequencing methods developed after the Sanger methods, rather than one particular approach; indeed, the various approaches differ in their sequencing methods, detection methods and even the lengths of the DNA strands sequenced. Here, we concentrate on the Illumina sequencing technology that was applied to produce the data sets used in this thesis.

To prepare a sample for Illumina sequencing, the DNA/RNA of interest is first extracted from the biological sample. The extracted material is then fragmented using mechanical methods (ultrasonication, nebulisation) or enzymatic methods. Next, the sequencing adapters are attached to the ends of the sequence fragments to act as primers for sequence amplification, which in turn enables better detection resolution during the sequencing. For the amplification step, the sequences are attached to a solid surface, and a polymerase enzyme is used to create local sequence clusters by cloning the input DNA fragments in a process called bridge amplification.

During the actual sequencing step, the fluorescence-labelled bases (A, C, G and T) are added, one base at a time, to the complementary template fragment, using the SBS process. The bases are then detected using an imaging approach; the sequences detected with Illumina technology are typically 50–300 base pairs.

2.2 Chromatin immuno-precipitation followed by sequencing

Chromatin immuno-precipitation sequencing (ChIP-seq) is an analysis method that offers a view of protein-binding sites on the genomic DNA. First, the method crosslinks all the proteins to the DNA *in vivo*. Next, the DNA in the cell is broken into shorter pieces by different methods, such as sonication or endonuclease, with the protein-bound DNA protected from shearing. Then, the DNA is purified by immuno-precipitation, which is aimed at the protein of interest. Finally, the crosslinking is reversed, and the DNA is ready for the sequencing process (Robertson et al. 2007).

ChIP-seq uses antibodies with varying specificities, potentially causing background signal or noise. PCR amplification is another potential source of bias because it can introduce unwanted duplication in sequencing reads (P. J. Park 2009). To address these biases, a common practice is to run control experiments (Landt et al. 2012; Flensburg et al. 2014), but computational methods during data analysis can also be used to decrease the impact of these biases.

2.3 Assay for transposase-accessible chromatin

An assay for transposase-accessible chromatin, followed by sequencing (ATAC-seq) (Buenrostro et al. 2015), is a fairly recent protocol that reveals DNA regions that are not wrapped around a nucleosome. The regions of open chromatin are associated with active gene transcription, repair, division or regulation (Tsompana and Buck 2014).

The sample preparation process starts by attaching nuclei to a surface, which is followed by incorporating a Tn5 transposase that will cut and add tags to accessible DNA. The DNA collected can then be submitted for sequencing. The particularly attractive part of this method is that it requires considerably less genetic material than other existing methods to assay open chromatin (e.g., Formaldehyde-Assisted Isolation of Regulatory Elements sequencing and DeoxyriboNuclease-seq) (Buenrostro et al. 2015).

2.4 Ribonucleic acid sequencing

Ribonucleic acid sequencing (RNA-seq) is the sequencing of the entire transcriptome, which allows a quantitative measurement of the expression of genes. This technology replaced earlier microarray technology because of its better range of detection, single base resolution and greater suitability for novel transcript discovery, as shown by multiple studies (Wang et al. 2014; X. Xu et al. 2013; Hung and Weng 2017; Rao et al. 2019).

To prepare the sample for RNA-seq, the RNA first needs to be extracted, possibly followed by a step to remove ribosomal RNA. The extracted RNA is then converted into complementary DNA (cDNA) that is suitable for the sequencing process.

3 Computational Methods for Studying the Regulation of the Genome via Epigenomics

3.1 Quality control

3.1.1 Read quality values and trimming

Problems with the quality of raw data can potentially affect data interpretation, making quality control of raw sequencing data a critical first step in its analysis. When working with public data, the starting data is most often in the form of FASTQ files, which contain the sequences of all the reads that have been sequenced in a text format, with a quality score for each base. Tools have been developed to get a quick general overview of the quality score for each base across all reads, and the issue of low-quality bases is typically addressed by trimming out the low-quality ends of the reads (Guo et al. 2014; S.-F. Yang et al. 2019). Trimming can also be used to remove the parts of the reads that match sequencing adapters, which are attached to the original sequences of interest and need to be trimmed out prior to aligning reads to the reference genome. Popular trimming tools include Ktrim (Sun 2020), SeqPurge (Sturm, Schroeder, and Bauer 2016), Trim Galore (Martin 2011) and Trimmomatic (Bolger, Lohse, and Usadel 2014), although some recent alignment tools can handle untrimmed reads (Del Fabbro et al. 2013).

3.1.2 Library complexity

Another common quality metric for ChIP-seq and ATAC-seq is library complexity (fraction of non-redundant reads). A low complexity can negatively affect peak finding and reproducibility.

3.1.3 GC content bias

The GC content is the amount in percent of Guanine-Cytosine found in a DNA or RNA molecule. In a truly random library, the presence of sequenced bases would be

expected to follow the presence of bases in the genome. For example, the GC content of the human genome is 40.9% (Piovesan et al. 2019), and GC content has been proven to be linked to fragment coverage (Benjamini and Speed 2012), with GC-rich fragments tending to have higher coverage. This GC content bias originates mostly from PCR and varies from sample to sample (Benjamini and Speed 2012), and it is commonly corrected for in the ChIP-seq and ATAC-seq data by applying a corrective model to the read counts (Benjamini and Speed 2012).

3.1.4 Saturation analysis

A saturation analysis can be done to determine whether the sequencing read coverage is sufficiently high to detect the events of interest (e.g., protein interaction). The coverage depends on the number of sequencing reads available and the size of the targeted genomic region. For example, the ENCODE consortia recommend 20 million reads for a study on mammalian transcription factors and 60 million reads for mammalian histone modifications (ENCODE 2017).

3.2 Read alignment

Read alignment (or mapping) is the step in which each read is associated with a genomic position. In the typical alignment approach, the reads are aligned to a known reference genome, allowing the determination of the precise origin location of most of the reads.

Alignment is a particular case of string matching. The features to take into account in the alignment process are seeding, base quality, existence of indels, paired-end reads and single nucleotide polymorphism (SNP) (Hatem et al. 2013). In the case of RNA-seq reads, the existence of introns spliced out of the genomic sequence also needs to be considered.

Currently, the popular aligners can be classified into two groups according to the data structure built into their seeding strategies (H. Li and Homer 2010; Ahmed, Bertels, and Al-Ars 2016): 1) hash-based (e.g., mrsFAST (Hach et al. 2010) and MAQ (Heng Li, Ruan, and Durbin 2008)) and 2) index-based (e.g., BWA (Heng Li and Durbin 2009) and Bowtie2 (Langmead and Salzberg 2012)).

3.3 Strategies to handle duplicated reads

Duplicated reads are those that map to the exact same genomic location. They may be produced by the polymerase chain reaction, or they can be naturally occurring replicates. While PCR duplicates are considered noise, because they are the same read sequenced multiple times, the naturally occurring duplicated reads are

considered a true part of the signal as they are derived from independent sequences (Bansal 2017). In RNA-seq studies, duplicated reads are not usually considered a problem, but for ChIP-seq and ATAC-seq, it is a common strategy to partially or fully deduplicate before peak calling (Bailey et al. 2013; Y. Chen et al. 2012). Instead of reducing aberrant signal by removing the duplicated reads across the genome, one can instead aim to remove the regions known to be associated with artefact signal; indeed, such regions have already been carefully curated and collected in the 'ENCODE Blacklist' (Amemiya, Kundaje, and Boyle 2019). The blacklisted regions have been shown to be enriched for reads mapping to multiple locations and duplicated reads (Carroll et al. 2014). Furthermore, filtering those regions has been shown to improve fragment length estimation and the normalisation of signal, which benefits peak calling and differential peak calling (Carroll et al. 2014).

3.4 Peak calling

Peak calling is the process of detecting genomic regions enriched in sequencing reads. A considerable number of peak calling software packages and comparison studies have been produced in recent years (Szalkowski and Schmid 2011), and the most popular software currently includes MACS2 (Zhang et al. 2008) , SICER (S. Xu et al. 2014), ZINBA (Rashid et al. 2011), F-seq (Boyle et al. 2008) and MUSIC (Harmanci, Rozowsky, and Gerstein 2014).

Ideally, ChIP-seq produces reads that are enriched in the regions in which the protein of interest binds. However, these reads only represent the 5' end of the sequenced fragments, and the density of reads around the region of interest will consequently present a bimodal distribution, with one distribution corresponding to the forward strand and the other corresponding to the reverse strand (Figures 2A and 2B). One implication of this phenomenon is that the software must adjust by shifting the reads half the estimated fragment length towards the centre of the region of interest.

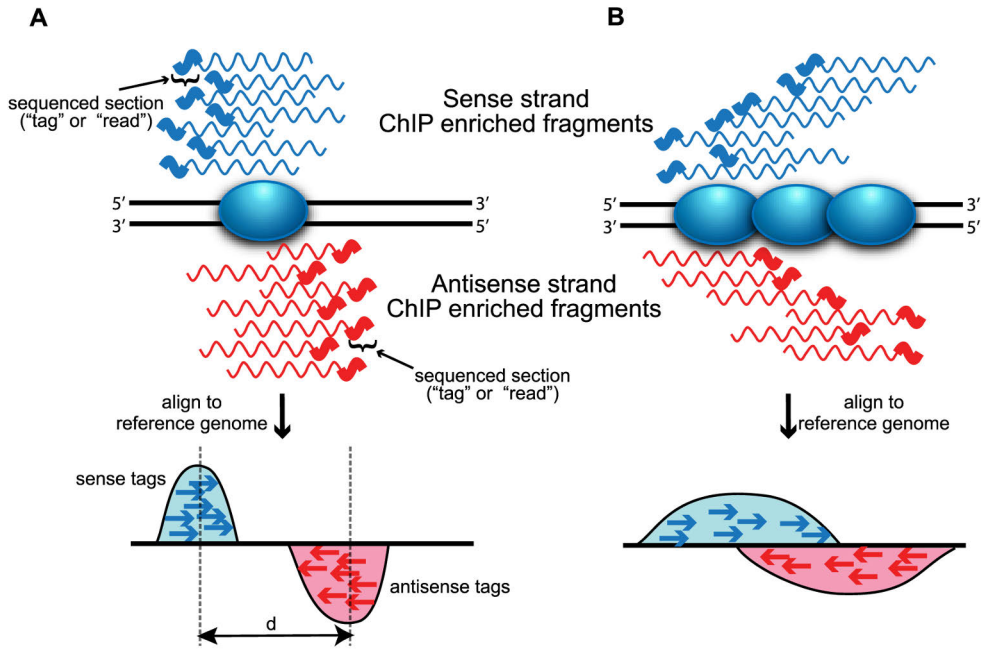

Figure 2. Visual representation of the forward (blue) and reverse (red) reads (bold) sequenced from fragments in the event of a transcription factor (A) and with a broader (such as histone) signal (B). This figure is from (Wilbanks and Facciotti 2010).

The peak calling process can be separated into two distinct steps (Wilbanks and Facciotti 2010; Thomas et al. 2016): the identification of candidate peaks and the statistical testing of those peaks. To identify candidate peaks, one must find the regions with a high density of reads, either by detecting clusters of reads or extended overlaps of reads (Fejes et al. 2008; Rozowsky et al. 2009). A popular solution is the use of a sliding window of defined size to scan the genome for the enrichment of aligned reads (Yong Zhang et al. 2008; Harmanci, Rozowsky, and Gerstein 2014). Peak calling software then tests those candidate regions for significance, which is done by using the null hypothesis that the reads are randomly distributed throughout the genome in order to model the background reads. The background reads are modelled using a Poisson model or a negative binomial model using regions of low coverage from the ChIP sample to infer the parameters of the distributions.

Over the years, there have been multiple attempts at benchmarking the various software packages developed for peak calling (Laajala et al. 2009; Wilbanks and Facciotti 2010; Rye, Sætrom, and Drabløs 2011; Micsinai et al. 2012), but the lack of a clear winner and the frequency of software releases and updates make any such benchmarks quickly outdated (Szalkowski and Schmid 2011). The latest benchmarking effort (Thomas et al. 2016) innovated by separating the peak calling problem into two sub-problems: the detection of candidate peaks and the statistical

testing of the candidates. By doing so, it identified three main features that make a good peak calling algorithm: 1) the peak detection process should not combine the reads produced by a ChIP-seq experiment (ChIP-seq sample) with the background reads obtained by running a ChIP-seq experiment without antibodies (input sample); 2) multiple sizes of windows should be used to detect enrichment; and 3) a Poisson model should be used rather than a negative binomial distribution to test for the significance of the peaks (Thomas et al. 2016).

Suitable candidate peaks are defined by a minimum of three parameters representing their position in the genome: chromosome, start coordinate and end coordinate. The peaks are also characterised by their height and breadth. A peak's height is the maximum read pileup value within the boundaries of the peak, and a peak's breadth is the distance between the boundaries of the peak. It was suggested that there are different types of peaks, based on their height and breadth (Pepke, Wold, and Mortazavi 2009; Landt et al. 2012), and that peak calling strategies should be adapted to be a function of those parameters. In general, peaks are referred to as broad or narrow, and although there is no clear boundary indicating the separation between narrow peaks and broad peaks, the term 'narrow peak' is typically used when transcription factors are studied, while the term 'broad peak' is typically used for histone modification studies. However, over the years, a strategy of narrow peak calling with specific histone modifications, such as H3K27Ac or H3K4me3, (ENCODE 2017) has been recommended; narrow peak calling for ATAC-seq data has also been recommended (ENCODE 2020).

3.5 Read counting

Both the analysis of differentially expressed genes (DEG) in the context of RNA-seq data and differential binding (DB) analysis in the context of ChIP-seq and ATAC-seq data often require the counting of reads located in specific genomic regions in order to make comparisons between conditions. Read counting in RNA-seq is generally more complex than in ChIP-seq/ATAC-seq because of issues specific to gene transcription; indeed, the genes can be spliced and some can even be overlapping. GenomicRanges (Lawrence et al. 2013), featureCounts (Liao, Smyth, and Shi 2014) and BEDTools (Quinlan and Hall 2010) are all able to handle the issues specific to RNA-seq data, while in ChIP-seq and ATAC-seq, the challenge in read counting is related to reads overlapping contiguous peaks.

Once the reads for the regions of interest have been counted (e.g., genes in DEG analysis and transcription factor binding regions or histone modification regions in DB), these read count values are saved into a count matrix with the rows representing the genomic regions of interest and the columns representing the biological replicate samples involved in the experiment.

3.6 Normalisation

The differences between the measurements taken from different samples can be explained by a combination of biological and technical factors, such as the sample preparation and handling process. Normalisation can be used to modify the read counts contained in the count matrix to decrease the effect of such systematic bias.

For the count data derived from ChIP-seq and RNA-seq studies, most of the available normalisation methods involve correcting for two main factors: sequencing depth and composition. Sequencing depth bias is shown in Figure 3A; if the sequencing depth in sample B is three times that of sample A, the regions in sample B will generally have three times as many reads than those in sample A. The composition bias is represented in Figure 3B. If a small number of regions show an extreme number of reads, normalisation can be skewed because of those extreme values.

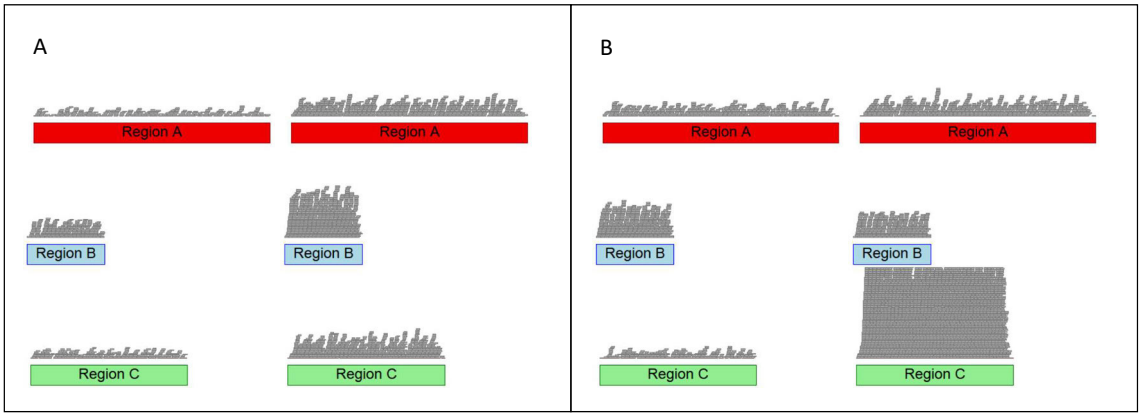

Figure 3. This figure depicts gene regions (coloured boxes) and the read coverage in grey above them. A) Representation of sequencing bias where sample B has three times the sequencing depth of sample A. B) representation of the composition bias where region C of sample B shows an extremely high read count.

A very simple library size normalisation takes the ratio of the sample with the highest sequencing depth (i.e., the number of total reads) and divides it by the number of reads in the different samples to calculate normalisation factors for each sample, which are then used to scale the values. This method then 'pulls' the samples with the lowest sequencing depths to the level of the highest.

The most popular methods used to normalise count data are the median of ratios (Anders and Huber 2010) and the trimmed means of M-values (TMM) (Mark D Robinson and Oshlack 2010). The TMM method trims the tail of the M-value (log ratio of intensities) distribution to avoid the inclusion of extreme values. The

weighted trimmed mean is then used to produce normalisation factors in a pairwise fashion between a sample selected as a reference and the other samples. The median of ratios is a simple popular method that builds a reference by taking the geometric mean for each region (genes or peaks) across the samples. All the values are then divided by the sample's reference.

3.7 Visual inspection of reads

One of the often-overlooked steps in checking the quality of ChIP-seq data is the visualisation of reads in the genomic context. Visualisation gives a global impression of the quality of the dataset, and in the case of a failed ChIP-seq experiment (e.g., too many amplification PCR cycles), abnormally high stacks of exactly identical reads can be observed. Multiple software packages have been developed for read visualisation, the most popular being UCSC genome browser (W. J. Kent et al. 2002; Fujita et al. 2011) and Integrative Genomic Viewer (IGV) (J. T. Robinson et al. 2011).

3.8 Fraction of read in peaks

One metric to assess the quality of a ChIP-seq experiment is the fraction of the read in peaks (FRiP) (Ji et al. 2008). This metric assumes that most of the reads present in a ChIP-seq experiment are part of the background reads and that only a fraction contain the true signal (i.e., the reads located in the peaks). Thus, the fraction of the reads in the peaks gives us an idea of the signal strength. The ENCODE guideline recommends a FRiP greater than 1% in order for a narrow peak experiment not to be considered a failure (Landt et al. 2012). However, this metric is quite sensitive to the size of the peaks (Qin et al. 2016), and a signal like histone mark H3K36me3, which is known to be broad and located in transcribed genes (Nelson, Santos-Rosa, and Kouzarides 2006; Bannister and Kouzarides 2011), would be expected to present a much higher FRiP than a transcription factor signal. Although the FRiP metric is less reliable for broader peaks (Qin et al. 2016), it still provides a point of comparison between replications of the same experiment.

3.9 Strand cross-correlation

Ideally, ChIP-seq will produce reads that cluster around the genomic regions bound by proteins. The enrichment of reads is bimodal around a true binding site because of the nature of sequencing platforms, which sequence both forward and reverse strands, and we can exploit this property to quantify the read clustering by calculating the Pearson correlation between the forward strand reads and the reverse

strand reads (Landt et al. 2012; Marinov et al. 2014; Carroll et al. 2014). Further, by shifting the forward reads towards 3' and calculating the correlation, we can profile the strand cross-correlation, which should reach a maximum at a shift size equal to the fragment length. When the correlation is plotted against the shift size, we typically obtain two peaks—one at read length (also called 'phantom peak'), which is an artefact caused mostly by duplicated reads in blacklisted regions, and one at fragment length, which represents the maximum overlap of the forward and reverse reads (Carroll et al. 2014) (Figure 4). A successful ChIP-seq experiment typically displays a peak at fragment length that is taller than that at read length, whereas a failed experiment displays a lower peak at fragment length than at read length.

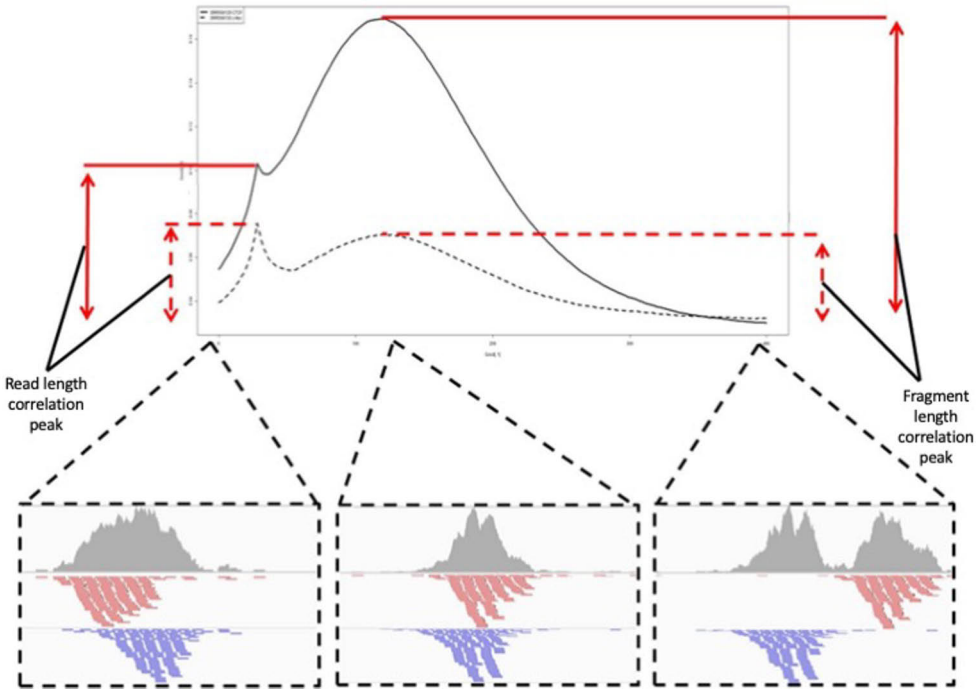

Figure 4. Representation of the strand cross-correlation plotted against shift value; the three frames on the bottom represent a visualisation of the shift values. The unbroken line is an example of a cross-correlation of a successful experiment, while the dotted line represents a cross-correlation of a failed experiment. Figure is adapted from (Carroll et al. 2014).

3.10 Differential peak calling

Differential peak calling can be described as the process of finding peaks that significantly differ between sample conditions. Differential signal evaluation is a routine step in other sequencing data types, such as RNA-seq or methylation

sequencing data, but the task is particularly challenging in ChIP-seq and ATAC-seq because of the relatively low signal-to-noise ratio (Landt et al. 2012). Further, the search space for binding events corresponds to the entire genome, and the breadth of the peak regions can vary from sample to sample (Steinhauser et al. 2016). Numerous tools for differential peak calling have been developed, and new tools are typically benchmarked by developers against some of the existing methods (Shen et al. 2013; Yanxiao Zhang et al. 2014; Allhoff et al. 2016), although a few independent method evaluation studies have also been produced (Steinhauser et al. 2016; Tu and Shao 2017).

Differential peak callers differ in their ability to handle replicate samples (Steinhauser et al. 2016). In general, biological replicates are beneficial, from a statistical point of view, as they improve peak detection accuracy and decrease the effect of background noise (Y. Yang et al. 2014). The recommendation of the ENCODE consortia, for example, is to include at least two biological replicates in the design of a ChIP-seq experiment (Landt et al. 2012). Amongst the differential peak callers that do not support replicate samples are SICER (S. Xu et al. 2014), MACS2 (Yong Zhang et al. 2008), ODIN (Allhoff et al. 2014), RSEG (Song and Smith 2011), MAnorm (Shao et al. 2012), HOMER (Heinz et al. 2010) and QChIPat (B. Liu et al. 2013). Differential peak callers that do support replicate samples include THOR (Allhoff et al. 2016), PePr (Yanxiao Zhang et al. 2014), diffReps (Shen et al. 2013), DiffBind (R. Stark 2011; Ross-Innes et al. 2012), MMDiff (Schweikert et al. 2013), Multi GPS (Mahony et al. 2014), ChIPComp (L. Chen et al. 2015), DBChip (Liang and Keleş 2012) and MAnorm2 (Tu et al. 2021).

Differential peak callers can be separated into two groups according to their need for an external candidate peak caller (Steinhauser et al. 2016; Allhoff et al. 2016; Tu and Shao 2017). Software packages that implement their own methods for candidate peak detection are called one-step methods, and those that rely on external candidate peak callers are called two-step methods.

Two-step methods were a logical first step of method development for differential peak callers, allowing the usage of existing methods for peak calling. After the peaks have been called, reads can be counted within the peaks to form an input for differential peak calling that can be carried out using existing statistical packages. The main drawback of the two-step methods is that once the regions of interest (candidate peaks) have been defined during peak calling step, they are fixed during the statistical analysis. This also allows the potential introduction of artefact regions during the peak calling step. One-step methods bring a solution to these problems by removing the separate peak calling step by separating the signal regions into inspection windows that will be tested for signal difference. In the case of one-step methods using sliding windows, windows can also be partially overlapping. The drawback with this approach is a reliance on specified preselected window size.

Setting a small window size may cause wide continuous regions with differential signal to be missed. Similarly, selecting a very large window may cause missing changes affecting narrow genomic regions. With one-step methods using hidden Markov models (HMM), statistical testing can be improved by taking advantage of signal around the windows as well. The inconvenience of such methods is that the limited number of hidden states (e.g. three hidden states in THOR) may decrease detection sensitivity (Tu and Shao 2017).

One-step methods considered in this thesis:

THOR (Allhoff et al. 2016) uses TMM as a default normalisation method, followed by a three-state hidden Markov model (HMM) (Gain in condition 1, Gain in condition 2 or Background) to call differential peaks. The probability of changing states, given prior observations, is calculated using a mixture of Poisson distributions. The statistical model is an HMM with a three-state topology which estimates the p-value using the negative binomial distribution with parameters based on states of the model. The multiple testing correction is done with a Benjamini-Hochberg correction of p-values.

PePr (Yanxiao Zhang et al. 2014) uses TMM as a default normalisation method, followed by a sliding-window approach to find enriched regions using a negative binomial distribution model and testing the enriched windows for significance with Wald's test. The multiple testing correction is done with a Benjamini-Hochberg correction of p-values.

diffReps (Shen et al. 2013) uses the geometric mean as a default normalisation, followed by a sliding-window approach to find enriched regions using a negative binomial distribution model and testing the enriched windows for significance with a test based on DESeq. The multiple testing correction is done with a Benjamini-Hochberg correction of p-values.

Two-step methods considered in this thesis:

ROTS (Suomi et al. 2017). In this thesis, the median of ratios is used to normalise ROTS input data. Statistical testing is performed by bootstrapping the data in order to optimise the two parameters, enabling ROTS to use a modified t-statistic that maximises the reproducibility of the results. The multiple testing correction is done with a Benjamini-Hochberg correction of p-values.

DiffBind (R. Stark 2011; Ross-Innes et al. 2012) uses the median of ratios of DESeq2 (Love, Huber, and Anders 2014) as a default normalization and performs the statistical testing with DESeq2 (which uses a Wald's test) to detect differential peaks. The multiple testing correction is done with an adaptation of the Benjamini-

Hochberg correction of p-values. Features are ranked by p-value, then each ranked p-value is multiplied by the number of tests divided by the rank of feature.

MAnorm2 (Tu et al. 2021) uses a linear transformation of the log2-transformed read counts to remove the M-A trend of the common peaks (the M-value is log2 of the fold change, and the A-value is the average log2 read count). The normalisation is applied between samples within conditions and then across conditions. MAnorm2 uses the limma (Soneson and Delorenzi 2013) modelling strategy to perform the statistical testing in order to detect differential peaks. The multiple testing correction is done with a Benjamini-Hochberg correction of p-values.

4 Datasets

The datasets presented here are used in the publications included in this thesis; those used for tool development (Publications I and II) were downloaded from public data repositories, and those used in Publications III and IV were specially generated.

4.1 Datasets used in Publication I

The purpose of this study was to develop a new tool for differential ChIP-seq peak detection and to compare it to popular tools on different types of datasets. The publicly available datasets used in Publication I are composed of two ATAC-seq and RNA-seq dataset pairs and two ChIP-seq and RNA-seq dataset pairs. The ChIP-seq and ATAC-seq datasets were aligned, using Bowtie2, against the hg19 or mm10 reference genome, as appropriate. Reads of low quality (quality value < 15) or those located in the ENCODE blacklisted regions were filtered out using samtools 1.2. The candidate peaks were called using MACS2 using the narrow peak option for the ATAC-seq and H3K4me3 ChIP-seq datasets, while the broad peak option was used for the H3K36me3 ChIP-seq dataset, following the ENCODE recommendations, the threshold used was the default '-q 0.01' (p-value adjusted with Benjamini-Hochberg correction). The differential peak calling was executed with default parameters for all software (for diffReps a sliding window of 1kbp is used with a moving step size of 100bp and a p-value threshold of 1 x 10^{-4}, PePr estimates the window size and shift size from the data and uses a threshold of p-value 0.05. THOR does not need any parameters, and DiffBind was used with DEseq2 statistical analysis method). The median of ratios normalisation was used before statistical testing with ROTS.

In the case of the RNA-seq dataset, we used available normalised gene expression count data (yellow fever and inflammation response). Unnormalised gene counts were available for the rheumatoid arthritis dataset, which was then normalised using the TMM method. Differential expression testing was performed using ROTS.

4.1.1 Yellow fever ATAC-seq and RNA-seq data

The yellow fever datasets originate from a previous publication (Akondy et al. 2017) in which the authors investigated the differentiation of human memory CD8 T cells.

They used live yellow fever vaccine (YFV), which produces immunity in humans, collecting YFV-specific CD8 T cells that had proliferated due to the vaccine and compared them to naïve CD8 T cells using RNA-seq and ATAC-seq approaches. The data are composed of five replicate samples of effector YFV-specific CD8 T cells (collected in the first two weeks), three replicates of memory YFV-specific CD8 T cells (collected after three years) and eight replicates of naïve CD8 T cells. Both RNA-seq and ATAC-seq data are available for these samples.

In Publication I, the main focus was comparing naïve CD8 T cells with YFV-specific CD8 T cells. In the analysis, we combined effector and memory cell samples to obtain eight replicates of YFV-specific CD8 T cells and eight replicates of naïve CD8 T cells. The data are available in the Gene Expression Omnibus (GSE100745 for the RNA-seq and GSE101609 for the ATAC-seq).

4.1.2 Inflammation response ATAC-seq and RNA-seq data

This dataset originates from a publication (S. H. Park et al. 2017) in which the authors were investigating the regulation by cytokines of the responses of toll-like receptors, which is an essential part of host defence, toxicity avoidance and homeostasis. Indeed, tumour necrosis factor (TNF) cytokines are a strong actor in innate immunity and the inflammation defence against pathogens, and the authors stated that a 24-hour pre-treatment with TNF cytokines attenuates lipopolysaccharide (LPS)-induced epigenetic modifications. The study examined four groups of $CD14^{+}$ monocyte-derived macrophages in a 2×2 experimental design—with or without pre-treatment with TNF cytokines and with or without a challenge with LPS. The goal was to observe the variation of gene expression of a group of LPS-induced genes reacting to TNF and LPS treatments.

We selected the two groups not pre-treated with TNF to use in our study because they would show the largest differences when compared. The ATAC-seq dataset comprises five replicate samples and the RNA-seq dataset comprises three replicates. The data is available in the Gene Expression Omnibus (GSE100383).

4.1.3 Rheumatoid arthritis ChIP-seq and RNA-seq data

This dataset originates from a study (Ai et al. 2018) that aimed to create an epigenetic landscaping of fibroblast-like synoviocytes (FLS) from patients affected by rheumatoid arthritis (RA), which is an auto immune disease in which the cartilage becomes inflamed and patients progressively lose mobility, in contrast to osteoid arthritis (OA), which is believed to be caused by mechanical stress on the joints. The epigenome of the FLSs was under investigation because they move from the lining

of the synovial capsule to invade the cartilage and assume a particularly aggressive phenotype in patients with RA.

The dataset is composed of two groups of 11 patients with FLSs—one with OA and the other with RA. ChIP-seq and RNA-seq was used to examine different types of histones (H3K4me3, H3K36me3, H3K27ac, H3K27me3 and H3K9me3), and the dataset is available in the Gene Expression Omnibus (GSE112658).

4.2 Datasets used in Publication II

The emphasis of the Publication II was on the development of the visualization tool for epigenomics data. The public datasets included in the publication were used to generate the example visualizations. The datasets include two ChIP-seq datasets and an ATAC-seq dataset. The quality of the sequenced reads was checked with FastQC (Andrews S. 2010) and reads were aligned against reference genome mm10 (butadiene ChIP-seq and ATAC-seq) and hg19 (gastric adenocarcinoma ChIP-seq) using Bowtie2 (Langmead and Salzberg 2012). Duplicated reads were filtered out using samtools 1.2, and the peaks were determined with MACS2 using the options '*–broad–nomodel -q 0.05'* in order to be used with two-step methods. DiffBind for differential peak calling was used with DEseq2 statistical analysis method (Love, Huber, and Anders 2014) and an FDR < 0.05. The one-step methods were run on default parameters (diffReps a sliding window of 1kbp is used with a moving step size of 100bp and a p-value threshold of 1 x 10^{-4}, PePr estimates the window size and shift size from data and uses a threshold of p-value 0.05 and THOR does not need any parameters and returns all the peaks found regardless of the p-value. All tools used in the differential analysis steps represent robust methods widely used in the analysis of sequencing data.

4.2.1 Butadiene ChIP-seq and ATAC-seq data

A study by (Israel et al. 2018) investigated the epigenetic footprint of exposure to butadiene, which is a known carcinogenic chemical found in cigarette smoke and car exhaust; it is also known to alter chromatin structure. The authors exposed two strains of mouse (CAST/EiJ and C57BL/6J) to clean air or to 1,3-butadiene for six hours per day over a two-week period. They collected tissue from the lungs, liver and kidneys to perform H3K27ac ChIP-seq and ATAC-seq analysis.

The authors used the CAST/EiJ liver dataset, which contains five clean air–exposed replicates and five 1,3-butadiene–exposed replicates for both ChIP-seq and ATAC-seq. The dataset is available in the Gene Expression Omnibus (GSE108990).

4.2.2 Gastric adenocarcinoma ChIP-seq data

A study by (Ooi et al. 2016) used the epigenetic landscape of gastric adenocarcinoma to study regulatory enhancer elements. Gastric tissue samples were collected from the SingHealth tissue repository to create a dataset composed of five normal tissue samples (without any malignant tissue) and five tumour tissue samples that match with cancer cell lines. The dataset is available in the Gene Expression Omnibus (GSE85467).

4.3 Datasets used in Publication III

In Publication III, the emphasis was on showing that the breadth of the histone modification H3K4me3 at the promoter mark is conserved under hypoxic stress, which has the effect of retaining core and stress adaptation functions during endometriosis in endometrial stromal fibroblasts (ESF). To study the dynamics of H3K4me3 promoter marks in hypoxia, decidual stromal cells (DSC) (decidualised ESFs) and ESFs were exposed to a hypoxia condition for 16 hours. The resulting cells were used to conduct a ChIP-seq experiment and produce two biological replicates for each condition (ESF hypoxia, ESF normoxia, DSC hypoxia and DSC normoxia).

The quality of the sequenced reads was checked with FastQC (Andrews S. 2010), the sequencing reads obtained were aligned to the hg19 reference genome using Bowtie2 (Langmead and Salzberg 2012) and merged with samtools 1.2 (Heng Li et al. 2009) before calling the peaks with MACS2 using the options '--broad --nomodel --extsize 147 --broadcutoff 0.1'. Neighbouring peaks (distance < 3kb) were merged and annotated to the closest promoter using GREAT (McLean et al. 2010). The differential analysis was done with diffBind using DEseq2 and a threshold of q-value $< 1 \times 10^{-10}$ was applied for the correlations between differential peaks and differential expression. DiffBind was preferred to other differential peak calling tools because it is a two-step method which require a pre-selected list of genomic regions and we needed to detect differential peaks on selected genomic regions.

The related RNA-seq data originates from previously published data of normoxic DSCs and ESFs along with hypoxia-treated DSCs and ESFs. The dataset is available in the Gene Expression Omnibus (GSE111570 and GSE63733). The quality of the sequenced reads for this dataset was checked with FastQC (Andrews S. 2010), then aligned to the GRCh37 human reference genome using Tophat2 (Kim et al. 2013) and the gene counts were calculated with HTseq (Planet et al. 2012) according to the Ensembl gene annotation (GRCh37.69). The counts were normalized as Transcript Per Millions (TPM). The differential expression analysis was done with edgeR (M. D. Robinson, McCarthy, and Smyth 2010) using upper quartile normalization and selecting FDR < 0.01, TPM > 2 and fold change > 2 as filtering cut-offs.

4.4 Datasets used in Publication IV

In Publication IV, the focus was on proving that the transcription factor MAZR regulates the development of regulatory T cells (T reg cells) that express FOXP3. Splenic tissue was collected from mice aged between six and eight weeks—three wild-type (WT) and three MAZR-KO—and T reg cells were isolated; RNA sequencing was then performed and a DEG analysis conducted. The dataset is available in the Gene Expression Omnibus (GSE123149).

The quality of the sequenced reads was checked with the FastQC tool and the reads were aligned with STAR 2.5.2b (Dobin et al. 2013) to the mouse reference genome (mm10). The number of uniquely mapped reads associated to Ensembl annotated genes were counted using subread 1.5.1 (Liao, Smyth, and Shi 2013). The count data was normalized using the TMM normalization from the edgeR package. Lowly expressed genes (less than one count per million in at least three samples) were filtered out and an offset of one was added to the counts before log2 transformation. ROTS was used for statistical testing requiring thresholds of p-value < 0.05 and absolute fold change > 1.5 to detect differentially expressed genes. All tools used in the different analysis steps represent robust methods widely used in the analysis of sequencing data.

5 Results

5.1 Differential ChIP-seq and ATAC-seq peak calling with ROTS and comparison with existing tools (Publication I)

Together with the various differential peak callers developed over the years have come a range of strategies for assessing their performance, but comparisons between existing tools remain challenging, especially because of the lack of a reliable gold standard dataset (Steinhauser et al. 2016; Allhoff et al. 2016). Early attempts at comparing tools included overlaps between the top results lists (list of results sorted by p-value and truncated at a p-value threshold), the number of peaks detected under a certain p-value threshold and density plots of the read pileups across the detected genomic regions (Shen et al. 2013; Yanxiao Zhang et al. 2014). Publication I describes the development of a workflow for the purpose of differential peak calling based on an existing ROTS R package and a comparison of the developed workflow against popular tools (DiffBind, MAnorm2, diffReps, PePr and THOR) using synthetic ChIP-seq data, real ChIP-seq data and real ATAC-seq data. The methods were selected for their novelty and/or their popularity, as well as, their history of being benchmarked to other tools.

5.1.1 Synthetic data

Synthetic datasets for validating ChIP-seq analysis tools have been developed in at least two earlier studies (Steinhauser et al. 2016; Allhoff et al. 2016). In the first study, manually curated biological ChIP-seq data was compiled and the differential peaks then inserted into the data. In the second study, the synthetic dataset was fully generated by software that attributes reads with a negative binomial distribution to selected genomic regions.

In Steinhauser et al. (2016), the top 20000 peaks identified by a peak caller in a H3K36me3 ChIP-seq experiment were selected to create two sample conditions. For the first condition, only the reads located in the peaks, which represented the signal, were considered for the subsequent procedure, while the second condition was created by splitting the same set of peaks into two groups—one representing the non-differential peaks in which the reads are identical to the first condition, and the other

split into ten subgroups of 1000 peaks representing the differential peaks. These ten subgroups of 1000 peaks were down-sampled sequentially (10%, 20%, …, 100%) to represent the different intensities of the differential peaks. Finally, a realistic noise background was added to each condition based on a real dataset.

The Steinhauser et al. (2016) synthetic dataset has great advantages, such as a true biological read distribution and exact knowledge of differential and non-differential peaks, but it lacks the replicate samples that we were emphasising in our work. We therefore used the dataset as a base to create our own version of a reliable validation dataset in Publication I. We created five replicates for each condition by further down-sampling each original condition sample by a random percentage (10% to 30%). This validation dataset thus enabled a comparison of true and false positives found by each tool. Furthermore, the gradient of the down-sampling gave us the opportunity to examine the behaviour of each tool according to the magnitude of the difference.

We compared the top differential peaks found by each tool (false discovery rate [FDR] < 0.05) and examined the rates of true and false positives (Figure 5A). While all the tools performed similarly when the intensity differences were large (60% to 100% difference), ROTS and THOR detected larger proportions of differential peaks with lower intensity differences. Overall, ROTS, PePr and THOR reported the most true positives while diffReps was the only tool with a significant number of false positives. The tools were also benchmarked for running time and memory consumption; we found ROTS to be by far the fastest and most memory efficient, while diffReps was the slowest and DiffBind used the most memory.

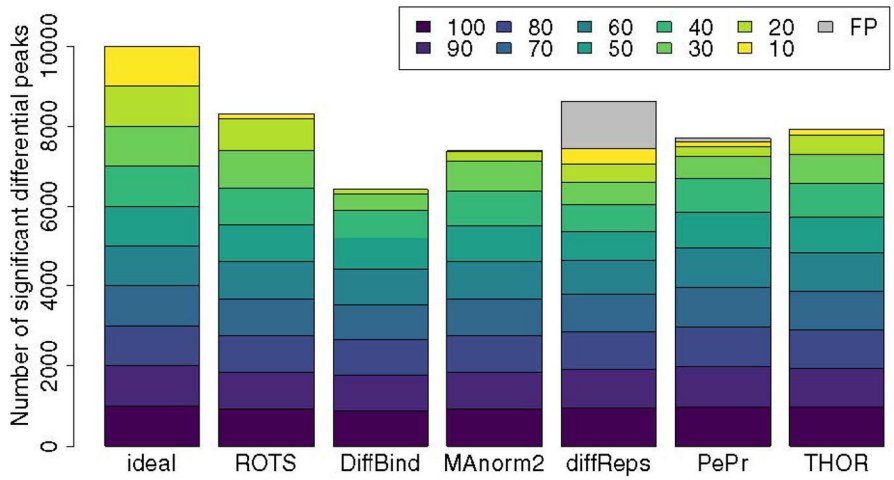

Figure 5. Detection of true positive peaks (colour) and false positive peaks (grey). The findings reported by each tool are displayed against an ideal finding of 1000 true positives per intensity category. The gradient of colour represents the percentage intensity of the difference. This figure is adapted from (Faux et al. 2021).

5.1.2 Number of differential peaks

One simple way to compare tools' performances is to examine the overlap of their results. Indeed, overlapping the regions detected by the tools shows a general agreement regarding the unique regions found, though this is in contrast to a previous study in which the general agreement between different tools was found to be quite low (Steinhauser et al. 2016).

An analysis of the number of significant peaks detected by the different tools (FDR < 0.05) shows significant differences in the numbers of detected differential peaks between the tools and across datasets (Table 1). THOR and diffReps found a significantly higher number of differential peaks than the other tools. PePr showed an inconsistent number of differential peaks, while the two-step methods (ROTS, DiffBind and MAnorm2) found similar numbers of differential peaks, except with the H3K4me3 ChIP-seq dataset, for which MAnorm2 found only ten peaks compared to the thousands found by ROTS and DiffBind. It should be noted that the two-step methods agree on the low number of differences between the conditions in the H3K36me3 ChIP-seq dataset.

We also looked at the most significant differential peaks and the overlaps between tools (Figure 6). The overlapping of the 2000 most significant differential peaks for each tool across the datasets shows a generally higher percentage of overlap for the ATAC-seq datasets than for ChIP-seq. The two-step methods (ROTS, DiffBind and MAnorm2) showed an overall high percentage of overlap (32–80% for ATAC-seq and 21–60% for ChIP-seq), while the one-step methods (diffReps, PePr and THOR) showed reasonable overlap only for the ATAC-seq datasets (26–56%), but less than 17% for ChIP-seq. The same patterns were also observed when all significant differential peaks were compared (FDR < 0.05).

Finally, the number of differential peaks detected and the overlaps between tools outline an overall higher agreement amongst the two-step methods (ROTS, DiffBind and MAnorm2) than the one-step methods (diffReps, PePr and THOR) (Figure 6).

Table 1: Number of differential peaks reported for each tool and dataset. Table from (Faux et al. 2021)

	Two-step			One-step		
	ROTS	DiffBind	MAnorm2	diffReps	PePr	THOR
YF ATAC-seq	2017	8736	3816	9168	1955	36009
IFN ATAC-seq	37630	40001	32362	57209	44143	91118
RA H3K4me3	1913	3111	10	21443	1072	17343
RA H3k36me3	11	25	0	27549	1077	17483

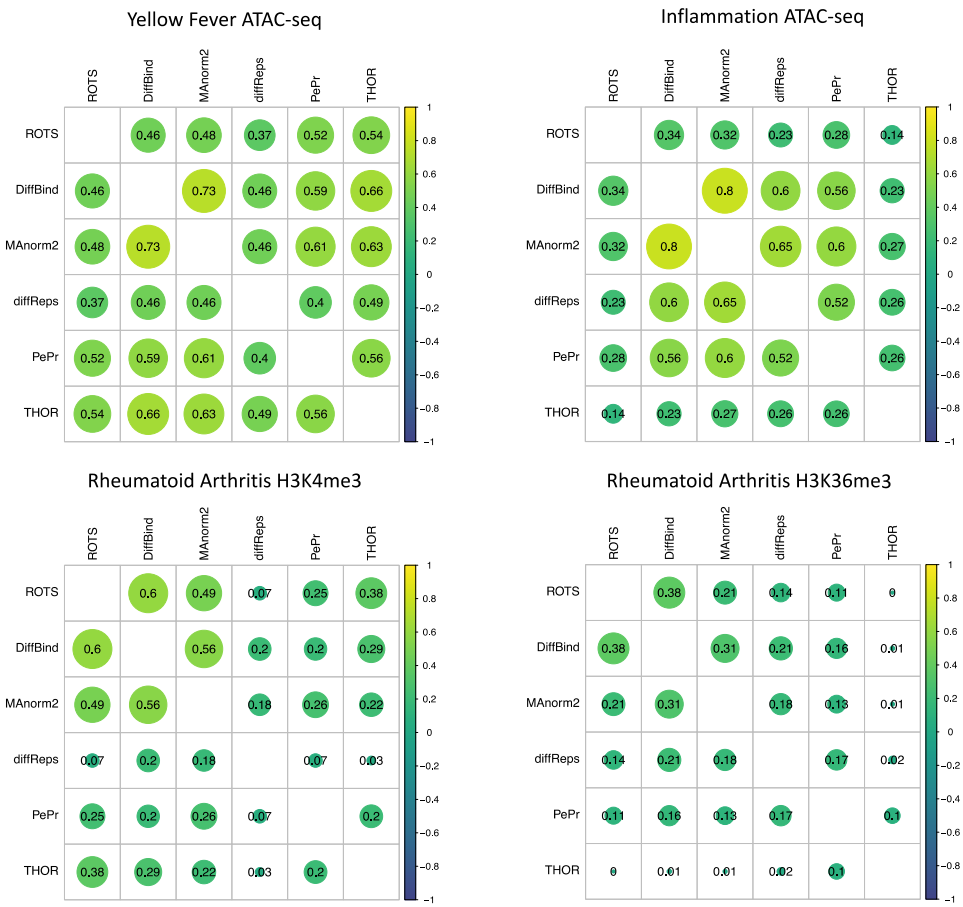

Figure 6. Percentage overlap between tools for the 2000 most significant differential peaks across the datasets for A) yellow fever ATAC-seq, B) inflammatory response ATAC-seq, C) rheumatoid arthritis H3K4me3 and D) rheumatoid arthritis H3K36me3. Figure adapted from (Faux et al. 2021).

5.1.3 Breadth and intensity

The shapes of peaks can be used to confirm results; different histone modifications, for example, have their own typical profiles of breadth and height (Karmodiya et al. 2015; Smolle and Workman 2013; Barth and Imhof 2010), and these profiles represent the fine regulation of chromatin openness in which biological processes can take place. Thus, a comparison of the shape and intensity of differential peaks allows a deeper understanding of the different tools' capabilities and a better overall comprehension of the underlying algorithms used.

An analysis related to the breadth, intensity and shape of the peaks was performed in three steps. First was calculating the average difference in read counts

for the first 2000 most significant differential peaks, which allows a global view of the differences of breadth and intensity. Second was creating a heatmap of the 2000 most significant differential peaks, displaying their intensity and their intensity fold change. Third, a few selected examples were visualised to highlight interesting differences in the behaviour of the differential peak callers.

In general, there were clear differences in signal patterns in most of the datasets (Figures 7A, 7B and 7C), except for RA H3K36me3 (Figure 7D). A tendency for narrower peaks in the ATAC-seq data was observed across the tools, and broader peaks were reported by the one-step methods (diffReps, PePr and THOR). DiffReps and THOR showed a pattern of detecting peaks with high intensity in the middle of the peaks but almost no differences on the sides, particularly with the ATAC-seq data. Some of the differential peaks detected by diffReps and THOR also exhibited a pattern in which the high intensity of difference splits and surrounds a peak centre, indicating the detection of a composite peak. Additionally, diffReps and PePr detected differential peaks that, in most cases, increased in signal, whereas few or no detections were reported with decreasing signal. The behaviour of the two-step methods (ROTS, DiffBind and MAnorm2) was very similar across the datasets.

A visual inspection of selected regions confirms these findings. Indeed, we can note identical regions for ROTS, DiffBind and MAnorm2 that precisely detect a peak (Figure 8), while THOR and diffReps frequently included lower intensity regions surrounding the peaks as well as regions that included two peaks, which were typically detected separately by two-step methods (ROTS and DiffBind). THOR exhibited a wide range of different behaviours, including correctly splitting a peak according to a fold change (Figure 8C) and detecting the surrounding of a peak without detecting the peak itself (supplement to Publication I).

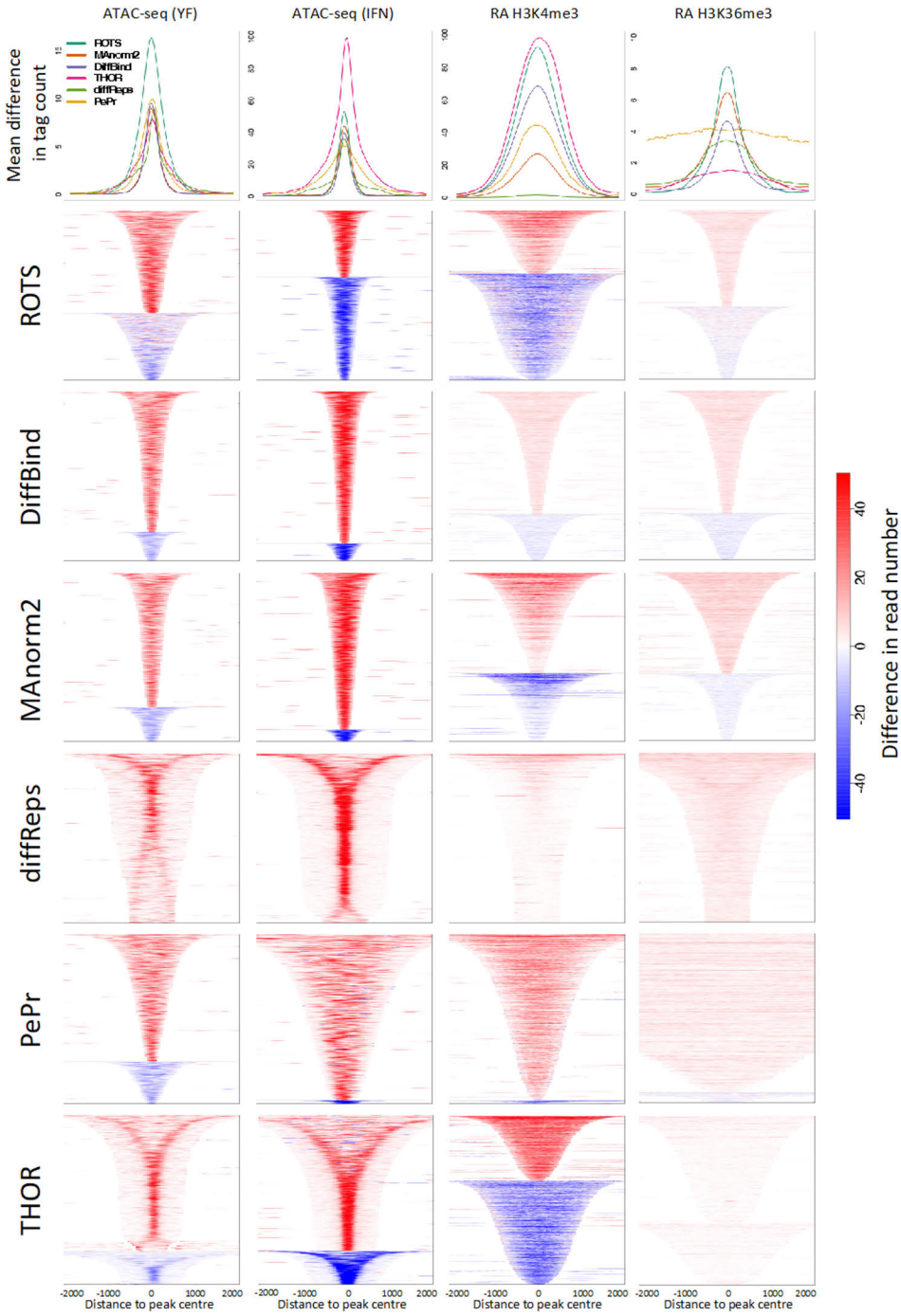

Figure 7. Percentage overlap between tools for the 2000 most significant differential peaks across the datasets. Figure is adapted from (Faux et al. 2021).

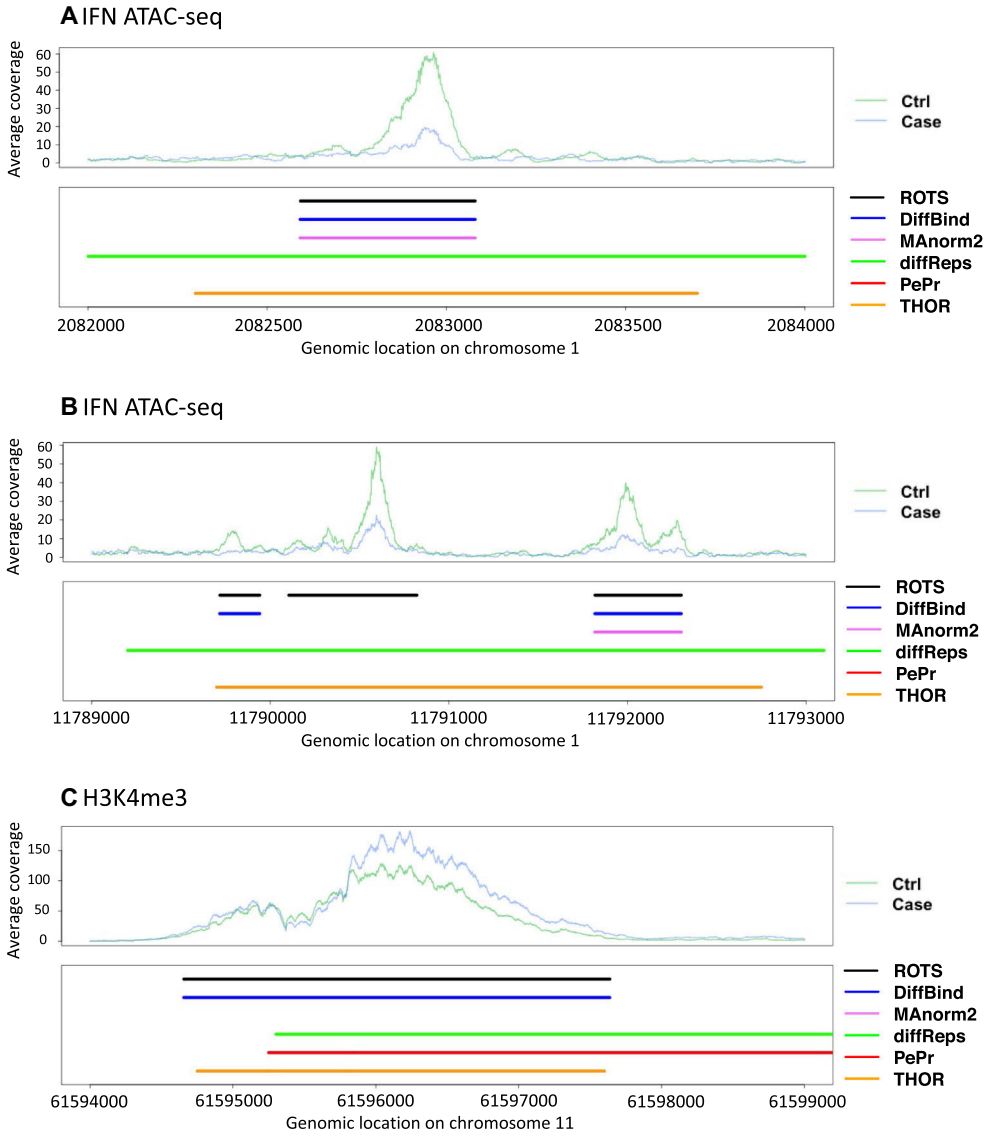

Figure 8. Percentage overlap between tools for the 2000 most significant differential peaks across the datasets. Figure is adapted from (Faux et al. 2021).

5.1.4 Correlation between differential peaks and RNA-seq

Evaluating the tools' performances on biological data can be challenging due to the absence of an absolute truth to which we can refer. Despite the fact that there is no gold standard biological dataset, we can nevertheless approximate the performance of a tool by associating different parts of the experiment to create a criterion for comparison. This can be based on the finding that levels of histone modifications and chromatin openness correlate well with associated gene expression levels (Gates,

Foulds, and O'Malley 2017; Karlić et al. 2010; Starks et al. 2019). We used this knowledge to correlate differential peaks resulting of ChIP-seq and ATAC-seq (providing information on the presence of histone modifications and open chromatin regions respectively) with differential expression levels of the matching genomic regions. This was done in order to measure the relationship between the fold difference in the differential peaks detected by the different tools and the fold difference in the expression of related genes.

The evaluation of the tools' performances was carried out with selected datasets containing differential conditions in epigenetics (ChIP-seq and ATAC-seq) and in gene expression (RNA-seq). The most significant findings of the differential peak detection tools were compared using the first 2000 differential peaks of the ranked results. The intensity fold change of the detected regions was then compared to the fold change of the expression of the closest gene.

Plotting the fold change values from a defined list of best results against the fold change of the expression of the closest gene gives a snapshot of the correlation for the different tools (Figure 9). This correlation can be used as a criterion of performance and reveals that, for the H3K36me3 ChIP-seq, ROTS, MAnorm2 and THOR are the best performers when considering the top 500 differential peaks. The fold change values of the differential peaks reported by diffReps were also within a change range of −1-fold to +1-fold for the epigenetic data and for the differential expression data, indicating little changes between conditions within the differential peaks reported by diffReps.

The strategy used to calculate the correlation was based on an increasing window size in order to obtain a correlation curve representing the correlation variation as we include a larger list of significant results in every iteration. A higher curve represents a better relationship between the fold change of differential peaks and matching fold change of gene expression. A curve that stays horizontal represents consistent results. In Figure 10, the Pearson correlation of the fold change values against the fold change of the expression of the closest gene is shown for a top list of lengths ranging from 100 to 2000 in increments of 100.

Of the tools tested, ROTS showed the best performance on the ATAC-seq datasets; for the ChIP-seq dataset, ROTS and THOR showed the strongest performances for H3K4me3, while MAnorm2 showed the best performance for H3K36me3. Additionally, there was a consistently lower correlation for the one-step sliding-window methods (diffReps and PePr), while amongst the two-step methods, DiffBind and MAnorm2 showed correlations consistently lower than ROTS, except for the H3K36me3 dataset.

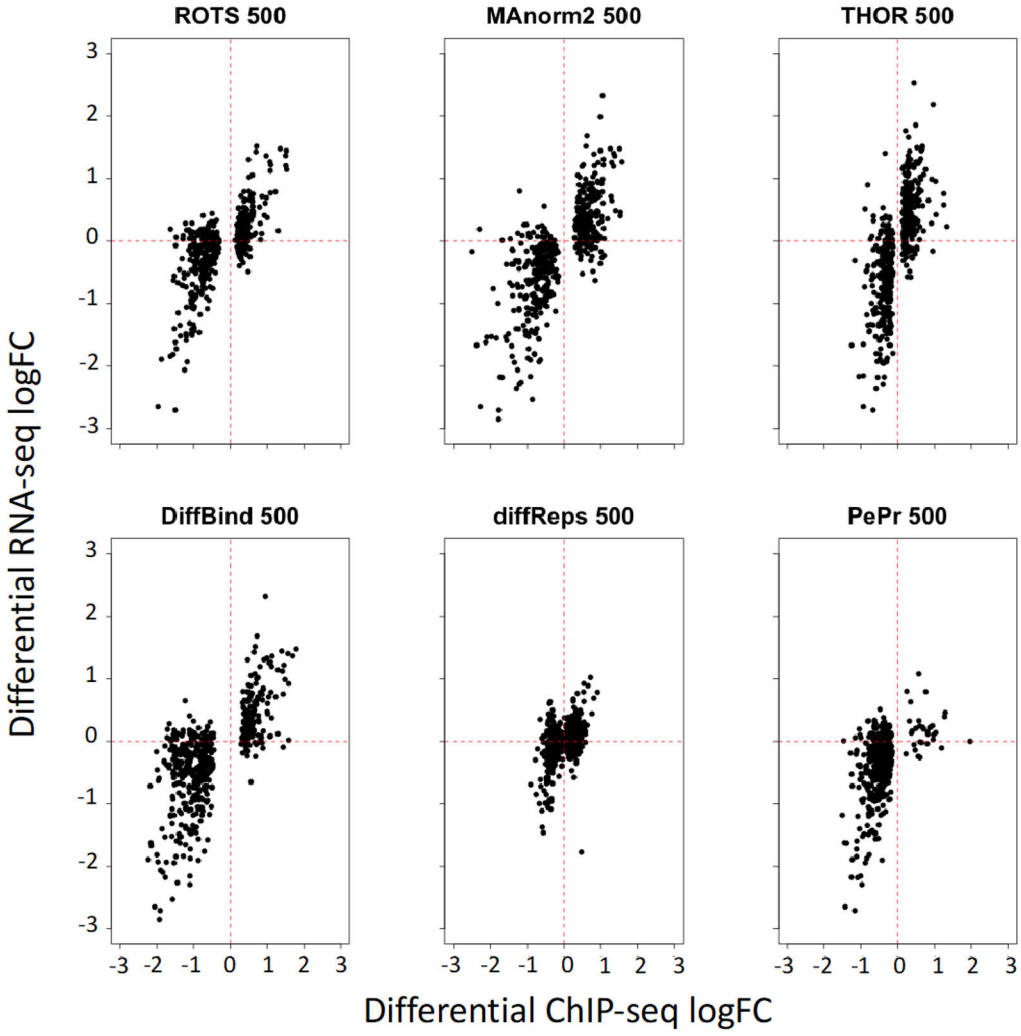

Figure 9. Scatterplot of the fold change values of the 500 top-ranked differential peaks from each tool against the fold change expression of the closest gene in the RA H3K36me3 dataset. ROTS, MAnorm2, diffReps, PePr, DiffBind and THOR.

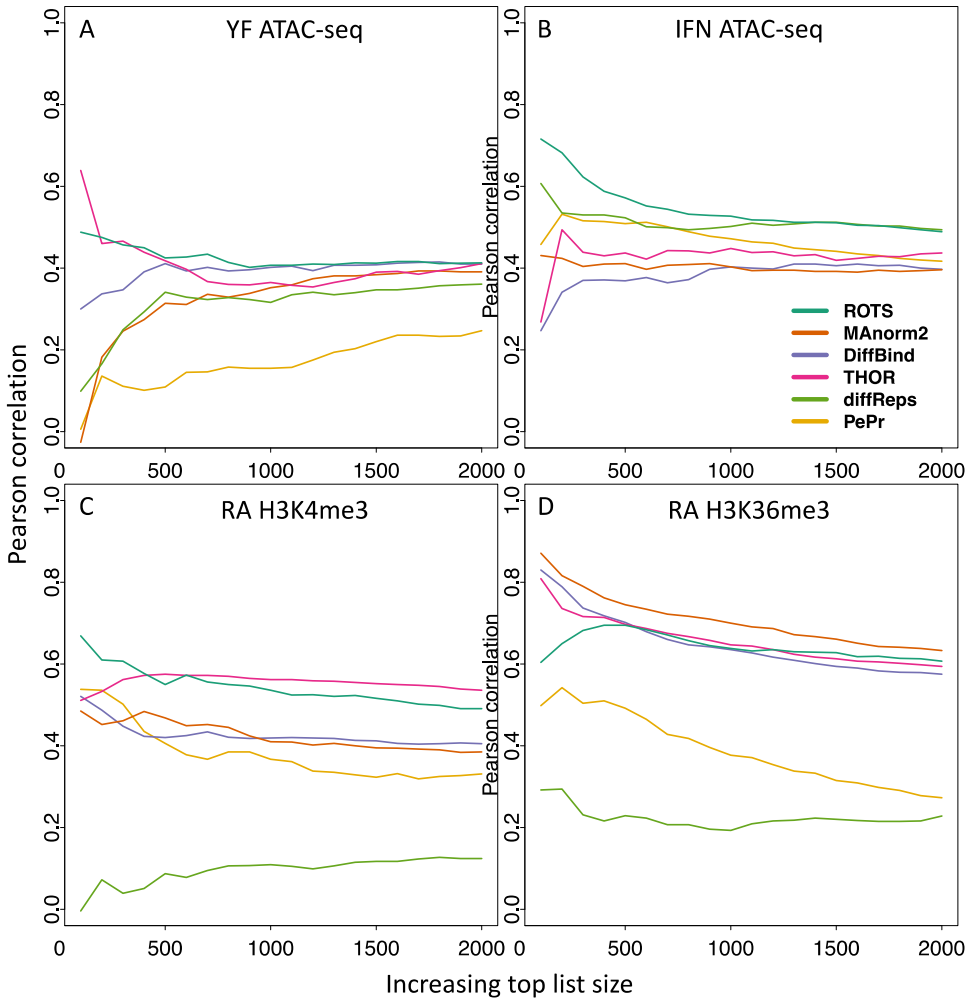

Figure 10. Pearson correlations between the fold changes of differential peaks and the fold changes of associated differentially expressed genes. The plots represent the behaviour of the correlation values from 100 peaks to 2000 peaks at intervals of 100 peaks. A) Yellow fever ATAC-seq, B) interferon response ATAC-seq, C) rheumatoid arthritis H3K4me3 and D) rheumatoid arthritis H3K36me3. Figure is adapted from (Faux et al. 2021).

5.2 Replicate-oriented visualisation of peaks with RepViz (Publication II)

The visualisation of data is often overlooked, despite it being a crucial step in data analysis. Indeed, visualisation by an experienced individual can yield information on library complexity, signal quality, analysis design and results interpretation, and despite a variety of existing tools, there is still a need for an efficient tool to visualise

replicated datasets, as current packages usually stack the replicates, making data interpretation difficult. The purpose of Publication II was the development of an efficient visualisation tool useful for differential ChIP-seq and ATAC-seq analysis and other sequencing data types.

5.2.1 Visualisation in the context of replicates

Visualisation is a powerful tool used to check the quality of the steps in differential peak calling and to support tool parameter adjustment and hypothesis validation. It is critical for the user to be able to efficiently visualise read coverage data in the genomic context, and several tools for visualising sequencing data exist. The most popular are the UCSC genome browser (W James Kent et al. 2002) and the IGV (J. T. Robinson et al. 2011), but other tools include bamView (Carver et al. 2013), ggbio (Yin, Cook, and Lawrence 2012), GenVisR (Skidmore et al. 2016), Gviz (Hahne and Ivanek 2016), rbamtools (Kaisers, Schaal, and Schwender 2015) and sushi (Phanstiel et al. 2014). These packages incorporate independent visualisation tools and R packages but lack the visual efficiency and simplicity of RepViz, which was developed as part of Publication II. Note that, while Gviz can produce unstacked replicate views similar to those available with RepViz, it can be a daunting task for people who are not experts in R programming.

The output of RepViz is a plot of a genomic region containing three different panels (Figure 11). The first is related to the visualisation of the read alignment BAM files, which contain the read positions; the second is the visualisation of the browser-extensible data (BED) files, which contain regions of interest; and the third is the genomic track that allows the previous tracks to be put into genomic context. In the example in Figure 11, the first four panels show five replicate samples of two conditions in the ATAC-seq and ChIP-seq datasets. The fifth panel shows the average read coverage of each sample condition group. The penultimate panel shows the BED files regions, and the last panel shows the genomic track.

In Publication II, RepViz is compared to two popular visualisation tools—IGV and Gviz (Figure 12). IGV displays the replicates from different condition groups on top of each other, which makes intra-condition and inter-condition comparisons difficult. While Gviz can produce a similar visualisation to RepViz, it does not accept different conditions having different numbers of replicates. Furthermore, while it is possible to see all conditions and replicates with Gviz, it is impossible to visualise the average coverage in the same plot. In a nutshell, RepViz can pack more information in a single visualisation than previous tools while keeping the visualisation easy to interpret.

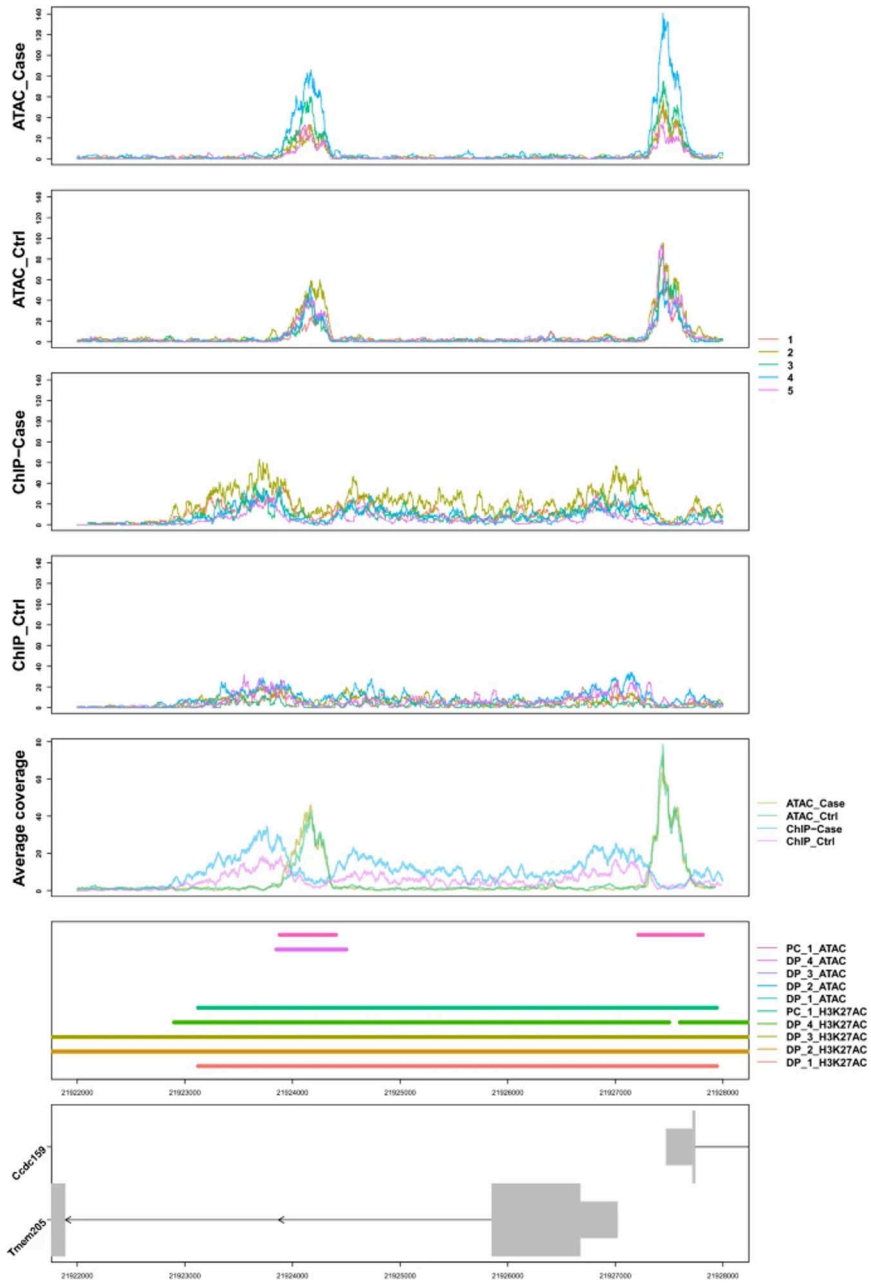

Figure 11. Visualisation of a promoter region for H3K27ac ChIP-seq and ATAC-seq in five replicates using RepViz. The first four panels are a visualisation of the reads for each of the five replicates present in the case and control conditions for both the ChIP-seq and ATAC-seq data. The fifth panel represents the average read coverage for each condition. The penultimate panel displays the regions detected by peak calling (PC) and differential peak calling (DP), and the last panel displays the genomic region. Figure is reproduced from (Faux et al. 2019).

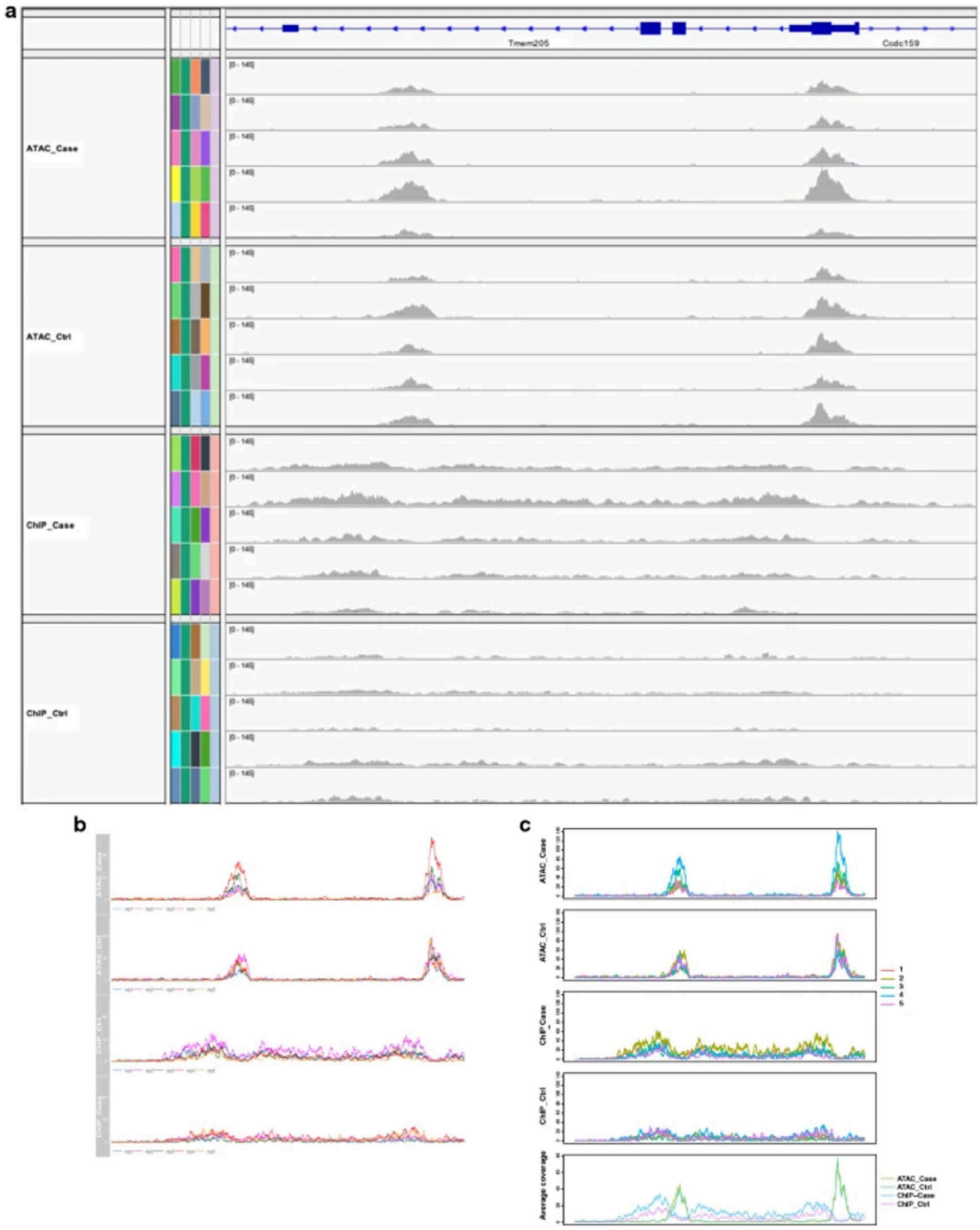

Figure 12. Visualisation of H3K27ac ChIP-seq and ATAC-seq in five replicates using (a) IGV, (b) GViz and (c) RepViz. Figure is reproduced from (Faux et al. 2019).

5.3 Dynamics of broad H3K4me3 marks in hypoxia (Publication III)

Tri-methylation of the lysin 4 in histone 3 (H3K4me3) is a marker of an active promoter (Pekowska et al. 2011; spicuglia and Vanhille 2012), and such changes have been shown to correlate with transcriptional changes (Okitsu, Hsieh, and Hsieh 2010). H3K4me3 is generally accepted as a narrow mark located on the transcription start site of active genes, but recent studies have shown that it is a bivalent mark that exhibits both broad and narrow binding behaviours (X. Liu et al. 2016; S. Park et al. 2020). Furthermore, broad H3K4me3 marks have been shown to be associated with core functions that define cell types and with transcriptional consistency by providing a buffer for stress (Benayoun et al. 2014).

Hypoxia is a fundamental stress known to induce cell adaptation mechanisms (Lee, Chandel, and Simon 2020). Endometrial stromal cells (ESC) of the uterus are regularly exposed to periods of hypoxia during menstruation and placentation (Maybin et al. 2018). During each menstrual cycle, ESCs differentiate from ESFs to DSCs to prepare the uterus for the implantation of a potential foetus. Consequently, repeated menstrual cycles can be seen as a constant cycle of stress, which makes ESCs a relevant model for studying the epigenetics mechanisms of cyclic stress and hypoxia.

This study provides a view of the dynamics of H3K4me3 in stress by studying the behaviour of H3K4me3 marks in ESCs when exposed to hypoxic periods.

5.3.1 H3K4me3 and hypoxia in endometrial stromal cells

Histone lysin demethylase (KDM) and histone lysin methyltransferase (KMT) are enzymes known to regulate the state of histones and can be grouped by the lysin they target (H3K4, H3K9, H3K27 or H3K36). Several KDMs and KMTs target H3K4 and, of these, there are groups with high transcription levels in normoxia and reduced transcription levels in hypoxia due to down-regulation which is consistent with previous findings (Batie et al. 2019; Chakraborty et al. 2019). Thus, the down-regulation of KDM and KMT that target H3K4me3 suggests the importance of this histone mark in the stress response of ESCs.

First, the ChIP-seq peaks obtained by peak calling on DSCs and ESFs in a hypoxic state and at a normal oxygen level were compared. A heatmap of peaks (Figure 13A) provides a holistic view of the peaks' breadths; the wide range of breadths is notable, ranging from a couple of hundred to a couple of thousand base pairs. After annotation of the peaks to the closest promoter, no significant differences in the number of promoter-associated peaks between hypoxia and normoxia was found. Furthermore, no differences appeared when filtering associated peaks using a low expression cut-off (expression level measured in transcripts per million

mapped reads (TPM > 2) or a high expression cut-off (TPM > 50), suggesting that the hypoxia-triggered changes are modifications made to pre-existing H3K4me3 marks (Figure 13B). The TPM > 2 criterion for expressed genes is inspired by (Wagner, Kin, and Lynch 2013) while the TPM >50 criterion is more arbitrary and represents an estimation of high expression of genes in the context of this publication.

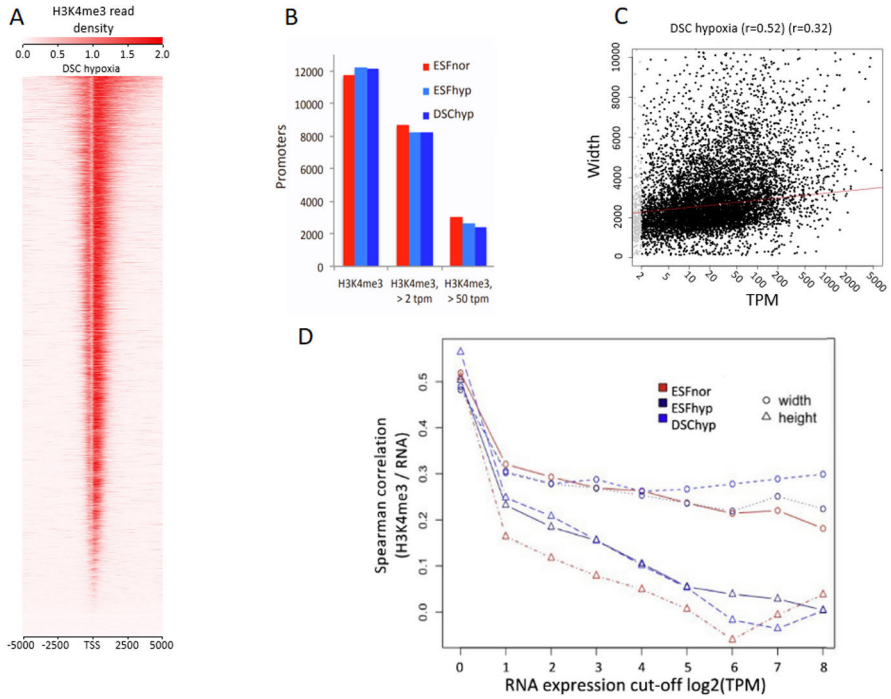

Figure 13. Broad H3K4me3 domains correlate with high intensity transcription. A) Density of the signal around the transcription start site in DSC hypoxia. B) Number of promoters associated with significant H3K4me3 peaks in the three conditions; all promoters are shown on the left, then promoters with transcribed genes (TPM expression > 2) and then promoters with high levels of transcription on the right (TPM > 50). C) Scatter plots representing the relationship between breadth and intensity of transcription. Correlation of peak width with all transcription levels result in a Pearson correlation of 0.52 while a correlation with transcripts levels with TPM >1 result in a Pearson correlation of 0.32. D) Behaviour of the Pearson correlation between transcripts levels and height or breadth of peaks according to transcription cut-offs.

The absence of a gain or loss of a clear subset of H3K4me3-marked regions motivated the study of broad H3K4me3 marks (broad regions with consistent and diffused ChIP-seq signals) as opposed to tall H3K4me3 marks (narrower regions with a spike in ChIP-seq signals). Consequently, the correlations of the heights and

breadths of the ChIP-seq peaks with the RNA-seq transcription levels of the promoter-associated genes were studied. The Spearman correlations of the different cell types (ESF and DSC) and conditions (hypoxia and normoxia) show a relationship between transcription levels and both height and breadth of the ChIP-seq signal (Figure 13C), and the behaviour of this correlation when the intensity of the transcription is considered (TPM > 2 to TPM > 1024) (Figure 13D) reveals a difference between height and breadth of H3K4me3 marks in their relationship with transcription levels. Indeed, the first level of filtration (TPM > 2) exhibits a decrease in correlations for all cell types and conditions due to the removal of lowly expressed genes. Additionally, a weak relationship is conserved by the correlation of transcription levels with the breadth of the peaks compared to the correlation with the height of the peaks. This suggests that the height of the peaks correlates with the associated gene being transcribed, while the breadth partially captures the intensity of the transcription.

5.3.2 Differences in breadths and heights of H3K4me3 marks

The analysis of the broad and tall H3K4me3 marks was extended by taking a subset of the 500 broadest and 500 tallest; the broadest were associated with a significantly higher transcriptional output than the expressed genes subset (TPM > 2) and all the H3K4me3-marked genes (Figure 14A). Furthermore, gene set enrichment analysis (GSEA) was done using the GSEA tool (www.gsea-msigdb.org/gsea) and a user defined list of the top 500 broadest peaks. The GSEA p-value is calculated by randomly permuting gene labels N times and taking the number of instances where the random permutation is better than the actual result obtained with the input list divided by N. The GSEA results of the broad H3K4me3 marks relative to the up- and down-regulated genes in hypoxia showed an enrichment of hypoxia up-regulated genes in broad H3K4me3 marks ($p < 0.001$) but not of hypoxia down-regulated genes ($p = 0.27$) (Figure 14B). Additionally, the majority of the genes related to the 500 broadest marks were maintained in hypoxia, whereas the portion of genes conserved in hypoxia for the 500 tallest peaks was small (Figure 14C). Thus, the results indicate that broad H3K4me3 promoters are an important group related to the regulation of genes during stress by at least partly maintaining their transcriptional output.

The gene promoters associated with the 500 broadest H3K4me3 marks were enriched with known functional categories relevant to endometrial physiology ($P < 1 \times 10^{-20}$), such as blood vessel development or embryonic morphogenesis. In contrast, the gene promoters associated with the 500 tallest genes showed an enrichment in the mRNA processing function known to be related to housekeeping

(Figure 14D). Further, progesterone receptor targets that allow decidualisation of the ESFs to DSCs (Mazur et al. 2015) when overlapped with the hypoxia-maintained broad H3K4me3 showed up-regulation of HOX10A and HAND2, which are core decidualisation genes. Overall, broad H3K4me3 promoters are linked to transcriptional up-regulation in hypoxia, allowing the conservation of core cell-type functions during stress, which supports previous findings (Benayoun et al. 2014).

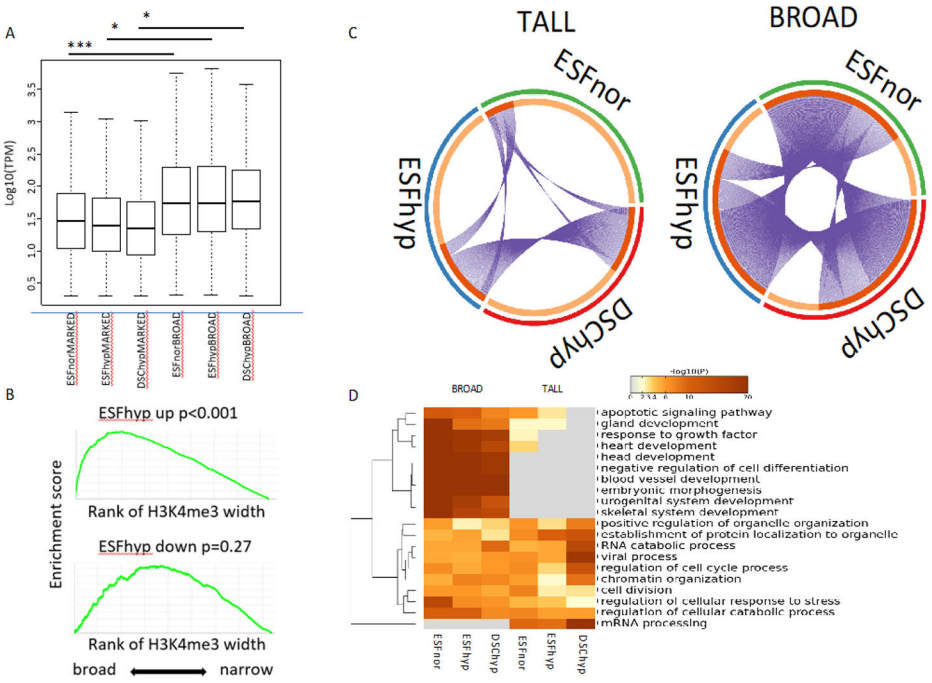

Figure 14. Conserved transcriptional activity in hypoxia for genes associated with broad H3K4me3 is associated with metabolic repression; A) boxplot of average transcriptional intensity in the top 500 broad H3K4me3 versus all expressed and H3K4me3-marked genes; B) gene set enrichment analysis (GSEA) of genes transcriptionally up- or down-regulated associated with the 500 broadest H3K4me3 peaks; C) proportion of peaks shared between the three conditions in the top 500 broadest and in the top 500 tallest H3K4me3 peaks; and D) hierarchical clustering (p-value) of the enriched GO terms of the gene promoters associated with the 500 broadest H3K4me3 peaks.

5.3.3 Correlation of H3K4me3 marks with transcriptional fold-changes

Further investigations of variations in peak height were performed using differential peak calling with the H3K4me3 genomic regions proposed by MACS2. Differential peak calling with DiffBind revealed a decrease in the H3K4me3 signal of peaks in hypoxia compared to the normal state in 1700 promoter regions associated with

down-regulated genes (qval < 10^{-10}). Further, the genes associated with these 1700 regions were highly enriched with RNA processing, RNA metabolism and cell cycle functions, as observed in the subset of the 500 tallest H3K4me3 marks. Thus, a substantial part of the regions containing decreased signal in hypoxic conditions is associated with metabolic repression, which is a core hypoxia response across all cell types (Semenza 2012).

Extension of the breadth of H3K4me3 domain in hypoxia revealed hypoxia adaptation genes relevant for endometrial functions. To investigate the changes in H3K4me3 peak height and breadth, we correlated these variables with changes in transcription. Peaks detected by DiffBind had a weak positive correlation with transcriptional changes (Pearson r = 0.25) while breadth changes displayed a marked positive correlation (Pearson r = 0.49). The 112 promoters detected with hypoxia-extended H3K4me3 peaks had robust enrichment for gene sets related to hypoxia and when these extended peaks were intersected with relevant endometriosis regulated datasets we discovered several endometriosis genes among the hypoxia-extended H3K4me3 peaks.

5.4 Application of differential gene expression analysis (Publication IV)

A multicellular organism relies on the generation of the proper amount and diversity of cell types. The fate of a cell is determined by its interactions with external stimuli, which lead a cell from a multipotent precursor towards a highly specialised cell. Adaptative immunity depends on differentiation decisions to drive early lymphoid progenitors to become $CD8^+$ and $CD4^+$ T cells, which will eventually become B cells and T cells (Germain 2002). T reg cells are T cells that differentiate in the thymus from a $FOXP3^-$ $CD4^+$ precursor (Burchill et al. 2008; Lio and Hsieh 2008) in order to regulate or suppress other cells of the immune system. T reg cells regulate the immune response against foreign particles but also against particles innate to the organism, thus helping to prevent autoimmune disease. Forkhead box protein P3 (FOXP3) expression is a defining factor in T reg cell function, triggering their differentiation (Khattri et al. 2003; Hori, Nomura, and Sakaguchi 2017; Fontenot, Gavin, and Rudensky 2003), but it also needs to be continually expressed for the T reg cells to keep their functional and transcriptional signatures (Josefowicz, Lu, and Rudensky 2012).

In Publication IV, the focus is on the transcription factor MAZR, coded by the gene Patz1, which contains the BTB (broad-complex, tramtrack and bric-a-brac) domain, which is common in transcription factors that assist with the down-regulation of gene expression (Bilic et al. 2006). It has previously been proven, using MAZR-deficient mice, that MAZR is a negative regulator of $CD8^+$ differentiation

from the precursor as well as differentiation from $CD8^+$ to $CD4^+$ (Bilic et al. 2006; Sakaguchi et al. 2010). While we know that MAZR regulates the CD8 lineage differentiation, the importance of MAZR in the regulation of CD4 lineage differentiation is still unclear. The focus of this study is to investigate the role of MAZR in the production of $FOXP3^+$ T reg cells.

A T-cell–specific deletion of MAZR in mice (MAZR-cKO) was used to study the production of $FOXP3^+$ T reg cells, which were found in greater numbers in various tissues related to immune system cell-line differentiation (spleen, thymus and lymphatic nodes) in the MAZR-cKO mice than in WT mice. This was confirmed by the production of *in vitro* T reg cells from WT and MAZR-cKO–naïve $CD4^+$ T cells and an observed increase in $FOXP3^+$ T reg cells.

Fluorescence-activated Cell sorting (FACS) was used in the study to separate the population of FOXP3+ Treg cells from other types of cells. FACS is a special type of flow cytometry used to separate heterogeneous cell mixtures based on fluorophores attached to the cells of interest with an antibody. Flow cytometry is the use of a flow to measure individual cells of interest with the help of a light focused at the point of measurement. (McKinnon 2018).

In the first steps of the Publication IV the dynamics of MAZR and FOXP3 are studied. It is observed that a decrease in MAZR expression levels upon FOXP3 introduction suggests down-regulation of MAZR in T reg cell differentiation, and an overexpression of MAZR in T reg cell precursors led to a decrease in the number of T reg cells in the MAZR-enforced population compared to the WT population. This suggests that expression levels of MAZR are critical for T reg cell differentiation.

With MAZR established as regulating the differentiation of T reg cells, it was then important to determine whether MAZR also controls the transcriptional functions of T reg cells. First, the expression of several characteristic T reg cell surface markers (CD25, Nrp1, CD62L, CD44, CD69, CTLA-4, GITR and KLRG1) was compared between MAZR-cKO $FOXP3^+$ and WT $FOXP3^+$. The fact that the pattern of expression of these surface markers did not change between the conditions indicates that there were no major alterations of the transcriptional functions of T reg cells in the absence of MAZR. When performing the differential expression analysis of MAZR-deficient mice against MAZR+ mice, there were only 33 genes up-regulated and 14 genes down-regulated in MAZR-deficient mice compared to the WT ($p < 0.05$; fold change > 1.5). Combined with the previous results, this indicates that, while MAZR is essential for T reg cell development, it only plays a minor role in the establishment of core functions once the fate of the cell has been decided.

6 Discussion

In this thesis, I have addressed different aspects of the analysis of epigenomic regulation data and given examples of how data analysis is helping to develop biological insights. I developed a data analysis workflow by applying existing ROTS software to differential peak calling and proposed an efficient visualisation solution to ease the quality control steps of the sequencing data analysis.

Differential peak calling is the process of highlighting differences in binding signal intensities between two conditions and is most often used to validate assumptions or explore differences between a disease state and a healthy control. In Publication I, I proposed a workflow based on the R package ROTS for differential peak calling and compared it to existing tools, showing the overall good performance of the workflow, particularly for ATAC-seq data.

The available tools for differential peak calling in ChIP-seq have recently been shown to produce highly variable results (Steinhauser et al. 2016; Tu and Shao 2017), and indeed the data analysis methodology of peak calling, normalisation and statistical testing varies greatly from tool to tool. While Publication I described a lack of overall consistency across tools for the ChIP-seq datasets, the ATAC-seq datasets showed more agreement, and the two-step methods (those using external peak callers) also displayed an overall higher percentage of overlap in results across all biological datasets. This higher agreement is likely due to the use of the same candidate peaks (MACS2 was used to call the peaks for both two-step methods), making the regions tested for differences highly similar. Similarly, the lower agreement amongst the one-step methods (diffReps, PePr and THOR) could be partially due to the different peak calling methodologies used, with THOR using an HMM model to select candidate regions and PePr and diffReps using a sliding-window method. It would be interesting to further compare the results of the two-step methods using multiple peak calling methods for initial candidate peak calling.

The fast innovation pace of the field of sequencing methods can not be understated and new sequencing methods were introduced during the time of working on the studies of this thesis. In Publication I both ChIP-seq and ATAC-seq are used in an effort to benchmark differential peak callers for both of these data types. However, with the recent emergence of ChIP-exo (Rhee and Pugh 2012) and

Cut & RUN (Skene and Henikoff 2017) technology, which are combining ChIP with a nuclease followed by high-throughput sequencing enabling better resolution and lower background than ChIP-seq, a new differential peak detection benchmark study including these technologies would now be a valuable update.

An existing synthetic dataset was used in Publication I as one of the validation datasets, and the publication in which that dataset was created (Steinhauser et al. 2016) reported overall poor performance of the compared differential peak detection tools; this is likely due to the use of a single replicate per condition because, in our comparison using a greater number of biological replicates per condition, all the tools performed relatively well. This highlights the importance of replicates in statistical testing in general. The changes in performance with greater numbers of replicates would be interesting to study further, and it is notable that a manually curated gold standard dataset with a reasonable number of replicates for the evaluation of differential peak callers is still lacking.

The majority of the normalization methods used for ChIP-seq and ATAC-seq data are based on methods originally developed for RNA-seq data. However, assuming a constant signal-to-noise ratio across conditions in the way these methods do may lead to erroneous biological conclusions. Recently, normalization approaches have started to emerge specifically for ChIP-seq and ATAC-seq data (Polit et al. 2021; Allhoff et al. 2016).Thus, a systematic benchmark study of ChIP-seq and ATAC-seq normalization methods including these novel approaches would be beneficial to the scientific community.

Visual inspection of data can improve the complex and iterative process of data analysis by helping the analysis design. The first occasion to take advantage of visualisation is in early quality control to ensure that the sequencing library is complex enough. Later in the workflow, visualising can help in the proper parametrisation of software, such as peak callers and differential peak callers—the choice between narrow and broad peak detection modes can have a drastic influence on the outcome of the differential peak callers, as discussed in Publication I. The results of a study can also often be confirmed with the visualisation of key genomic regions. For example, in Publication I, we visualised genomic regions using RepViz (presented in Publication II) to confirm the tendency of sliding-window differential peak callers to call peaks that are broader than necessary and the accuracy of MACS2 peak calls in general. The main limitation of RepViz is that it summarises a relatively small genomic region. The method could be further improved to allow better visualisation over larger genomic regions.

Strategies for efficient and clear visualization of ChIP-seq and ATAC-seq data have not been much addressed in the literature, apart from the choice of the colour palette to be used (Yin, Cook, and Lawrence 2012). In practise, simple visualisations

of genomic regions and peak callers' outputs are typically used in publications of the field in order to optimize the interpretability of figures.

The number of replicate samples used in epigenomics is continually increasing, making some of the earlier visualisation tools suboptimal; for example, some tools stack the coverage visualisations on top of each other without the ability to fully merge them by conditions. To address the situation, we implemented RepViz in Publication II, which is an efficient replicate-oriented visualisation tool for sequencing data. RepViz is published as a Bioconductor R package and has been actively maintained since publication. However, incorporating features facilitating the inclusion of visualisation in workflows or scripts would still greatly benefit experienced R users.

During my thesis work, I have analysed a large number of ChIP-seq and ATAC-seq datasets (some of which did not end up being included in the final publications), and carefully considered the details of the the different data analysis steps of these data types. Based on my experiences I would recommend the following general workflow for ChIP-seq and ATAC-seq data analysis. I recommend starting by checking the raw sequencing data quality (e.g. using FastQC), followed with read alignment (e.g. using Bowtie2 or STAR). The aligned reads should be visualized at this stage as a sanity check of the signal quality using for example RepViz. Low quality reads and reads located in blacklisted regions should be removed (e.g. using samtools) in order to remove potential artefacts that could lead to false positive peaks in the peak calling step. Appropriate parameters need to be selected for the peak calling step according to the type of protein targeted by the ChIP and broad and narrow peak detection options of the peak caller should be considered according to ENCODE recommendation (Landt et al. 2012). Further visualization of the signal in the genomic context (e.g. RepViz) across the detected peak regions enables a sanity check of the accuracy of the peak calling. The differential peak calling involves the choice of differential peak caller according to the needs of the study, such as number of replicates, broad or narrow peaks and presence of a predefined set of target regions (e.g. ROTS, THOR and MAnorm2 which performed best in Publication I). Last, the visualisation of the differential peaks in a genomic context allows a final check on the quality of the overall study (e.g using RepViz).

7 Summary of Publications

In Publication I, the successful adaptation of the ROTS method to differential peak calling is presented, and its integration into a full analysis workflow and a comparison to popular differential peak callers is described. The comparison with other methods is performed using both synthetic data, adapted from an existing synthetic dataset to increase the number of available replicates, and four biological datasets. The comparison with synthetic data exhibits the robust performance of the ROTS workflow, with particularly good results compared to other tools when the differences in signal are small. The comparison using real data was separated into three main categories related to the number of findings, the intensity and breadth of the peaks found and correlational analyses of matching genomic and epigenomic data. The study of the number of findings revealed a strong consistency amongst the results of the two-step methods, and the study of the peaks' breadth and intensity showed clear differences in behaviour between the one-sep and two-step tools. Finally, the correlational analyses demonstrated the robust and competitive results of the ROTS workflow. While the study underlines the general success of the ROTS workflow as compared to the other tools, it also discusses the underlying methodologies of the different tools and the recommended situations in which to use them thus providing a useful guideline for the sequencing data analysis community.

Publication II reported the development of RepViz, an R tool for visualising the coverage of sequencing data across replicated experiments. The package provides visualisation of coverage, average coverage, peaks, differential peaks and genomic tracks in one efficient and comprehensive picture, regardless of the number of replicates and conditions in the experiment. RepViz was compared to existing tools, which partially or completely lack its efficiency in visualising the coverage of a large number of replicates. The importance of sequencing data visualisation and its benefits for quality control, software parametrisation and hypothesis confirmation are also discussed. Publication II reports the publication of RepViz, a visualization tool that enables the replicate driven visualisation of genomic regions. RepViz can clearly and concisely picture information while also being easy to use for the neophyte.

In Publications III and IV, two studies are showcased in which bioinformatics tools are used to analyse and visualise data to extract useful knowledge and draw conclusions at the biological level. Publication III describes the dynamics of H3K4me3 in ESCs exposed to hypoxia. Correlation of the H3K4me3 promoter marks' height and breadth with the transcript levels indicates that tall marks are associated with genes being active or not, while broad marks capture the intensity of the transcription. Additionally, genes up-regulated in hypoxia contain a significant part of the genes associated with the 500 broadest promoter marks, indicating a role of the broadest marks in keeping core functions active during stress. Interestingly, the genes associated with the 500 tallest promoter marks are enriched with housekeeping functions. Lastly, the differential ChIP-seq analysis of down-regulated genes in hypoxia shows a general decrease in signal associated with functions related to metabolism, indicating metabolic repression as a response to hypoxia. In conclusion, Publication III supports the notion that H3K4Me3 promoter modifications are safe-guards of the cell identity by highlighting that H3K4me3 broad promoter domains are maintained during hypoxia and are associated with cell type-specific regulation in endometrial stromal cells.

Publication IV describes the role of the MAZR transcription factor in T reg cell development. The study demonstrates an increase in T reg cells *in vivo* and *in vitro* in the absence of MAZR. The enforced expression of MAZR is found to produce a decrease in T reg cell counts, and together, these results show the central role of the MAZR transcription factor in T reg cell development.

List of References

Ahmed, Nauman, Koen Bertels, and Zaid Al-Ars. 2016. “A Comparison of Seed-and-Extend Techniques in Modern DNA Read Alignment Algorithms.” In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1421–28. IEEE. https://doi.org/10.1109/BIBM.2016.7822731.

Ai, Rizi, Teresina Laragione, Deepa Hammaker, David L. Boyle, Andre Wildberg, Keisuke Maeshima, Emanuele Palescandolo, et al. 2018. “Comprehensive Epigenetic Landscape of Rheumatoid Arthritis Fibroblast-like Synoviocytes.” *Nature Communications* 9 (1): 1921. https://doi.org/10.1038/s41467-018-04310-9.

Akondy, Rama S., Mark Fitch, Srilatha Edupuganti, Shu Yang, Haydn T. Kissick, Kelvin W. Li, Ben A. Youngblood, et al. 2017. “Origin and Differentiation of Human Memory CD8 T Cells after Vaccination.” *Nature* 552 (7685): 362–67. https://doi.org/10.1038/nature24633.

Allhoff, Manuel, Kristin Seré, Heike Chauvistré, Qiong Lin, Martin Zenke, and Ivan G. Costa. 2014. “Detecting Differential Peaks in ChIP-Seq Signals with ODIN.” *Bioinformatics* 30 (24): 3467–75. https://doi.org/10.1093/bioinformatics/btu722.

Allhoff, Manuel, Kristin Seré, Juliana F. Pires, Martin Zenke, and Ivan G. Costa. 2016. “Differential Peak Calling of ChIP-Seq Signals with Replicates with THOR.” *Nucleic Acids Research* 44 (20): gkw680. https://doi.org/10.1093/nar/gkw680.

Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle. 2019. “The ENCODE Blacklist: Identification of Problematic Regions of the Genome.” *Scientific Reports* 9 (1): 9354. https://doi.org/10.1038/s41598-019-45839-z.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Andrews S. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data. Available Online at: Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc.” 2010.

Bailey, Timothy, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. 2013. “Practical Guidelines for the Comprehensive Analysis of ChIP-Seq Data.” *PLoS Computational Biology* 9 (11): e1003326. https://doi.org/10.1371/journal.pcbi.1003326.

Bannister, Andrew J., and Tony Kouzarides. 2011. “Regulation of Chromatin by Histone Modifications.” *Cell Research*. Nature Publishing Group. https://doi.org/10.1038/cr.2011.22.

Bansal, Vikas. 2017. “A Computational Method for Estimating the PCR Duplication Rate in DNA and RNA-Seq Experiments.” *BMC Bioinformatics* 18 (S3): 43. https://doi.org/10.1186/s12859-017-1471-9.

Barba, Marina, Henryk Czosnek, and Ahmed Hadidi. 2014. “Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology.” *Viruses* 6 (1): 106. https://doi.org/10.3390/V6010106.

Barth, Teresa K., and Axel Imhof. 2010. “Fast Signals and Slow Marks: The Dynamics of Histone Modifications.” *Trends in Biochemical Sciences* 35 (11): 618–26. https://doi.org/10.1016/J.TIBS.2010.05.006.

Batie, Michael, Julianty Frost, Mark Frost, James W. Wilson, Pieta Schofield, and Sonia Rocha. 2019. "Hypoxia Induces Rapid Changes to Histone Methylation and Reprograms Chromatin." *Science* 363 (6432): 1222–26. https://doi.org/10.1126/SCIENCE.AAU5870.

Benayoun, Bérénice A., Elizabeth A. Pollina, Duygu Ucar, Salah Mahmoudi, Kalpana Karra, Edith D. Wong, Keerthana Devarajan, et al. 2014. "H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency." *Cell* 158 (3): 673–88. https://doi.org/10.1016/j.cell.2014.06.027.

Benjamini, Yuval, and Terence P. Speed. 2012. "Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing." *Nucleic Acids Research* 40 (10): e72. https://doi.org/10.1093/NAR/GKS001.

Bilic, Ivan, Christina Koesters, Bernd Unger, Masayuki Sekimata, Arnulf Hertweck, Romana Maschek, Christopher B Wilson, and Wilfried Ellmeier. 2006. "Negative Regulation of CD8 Expression via Cd8 Enhancer–Mediated Recruitment of the Zinc Finger Protein MAZR." *Nature Immunology 2006 7:4* 7 (4): 392–400. https://doi.org/10.1038/ni1311.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Boyle, Alan P, Justin Guinney, Gregory E Crawford, and Terrence S Furey. 2008. "F-Seq: A Feature Density Estimator for High-Throughput Sequence Tags." *Bioinformatics (Oxford, England)* 24 (21): 2537–38. https://doi.org/10.1093/bioinformatics/btn480.

Buenrostro, Jason D, Beijing Wu, Howard Y Chang, and William J Greenleaf. 2015. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Current Protocols in Molecular Biology* 109 (January): 21.29.1-9. https://doi.org/10.1002/0471142727.mb2129s109.

Burchill, Matthew A., Jianying Yang, Kieng B. Vang, James J. Moon, H. Hamlet Chu, Chan-Wang J. Lio, Amanda L. Vegoe, Chyi-Song Hsieh, Marc K. Jenkins, and Michael A. Farrar. 2008. "Linked T Cell Receptor and Cytokine Signaling Govern the Development of the Regulatory T Cell Repertoire." *Immunity* 28 (1): 112–21. https://doi.org/10.1016/J.IMMUNI.2007.11.022.

Carroll, Thomas S., Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. 2014. "Impact of Artifact Removal on ChIP Quality Metrics in ChIP-Seq and ChIP-Exo Data." *Frontiers in Genetics* 5: 75. https://doi.org/10.3389/FGENE.2014.00075.

Carver, Tim, Simon R Harris, Thomas D Otto, Matthew Berriman, Julian Parkhill, and Jacqueline A McQuillan. 2013. "BamView: Visualizing and Interpretation of next-Generation Sequencing Read Alignments." *Briefings in Bioinformatics* 14 (2): 203–12. https://doi.org/10.1093/bib/bbr073.

Chakraborty, Abhishek A., Tuomas Laukka, Matti Myllykoski, Alison E. Ringel, Matthew A. Booker, Michael Y. Tolstorukov, Yuzhong Jeff Meng, et al. 2019. "Histone Demethylase KDM6A Directly Senses Oxygen to Control Chromatin and Cell Fate." *Science* 363 (6432): 1217–22. https://doi.org/10.1126/SCIENCE.AAW1026.

Chen, Li, Chi Wang, Zhaohui S. Qin, and Hao Wu. 2015. "A Novel Statistical Method for Quantitative Comparison of Multiple ChIP-Seq Datasets." *Bioinformatics* 31 (12): 1889–96. https://doi.org/10.1093/bioinformatics/btv094.

Chen, Yiwen, Nicolas Negre, Qunhua Li, Joanna O Mieczkowska, Matthew Slattery, Tao Liu, Yong Zhang, et al. 2012. "Systematic Evaluation of Factors Influencing ChIP-Seq Fidelity." *Nature Methods* 9 (6): 609–14. https://doi.org/10.1038/nmeth.1985.

Crick, F. H.C. 1958. "On Protein Synthesis - PubMed." *Symposia of the Society for Experimental Biology* Symposia o (12): 138–63.

Dahm, Ralf. 2008. "Discovering DNA: Friedrich Miescher and the Early Years of Nucleic Acid Research." *Human Genetics*. Springer. https://doi.org/10.1007/s00439-007-0433-0.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15. https://doi.org/10.1093/BIOINFORMATICS/BTS635.

ENCODE. 2017. "Histone ChIP-Seq Data Standards and Processing Pipeline."

———. 2020. “ATAC-Seq Data Standards and Processing Pipeline.” 2020. https://www.encodeproject.org/atac-seq/.

Esplin, Edward D, Ling Oei, and Michael P Snyder. 2014. “Personalized Sequencing and the Future of Medicine: Discovery, Diagnosis and Defeat of Disease.” *Pharmacogenomics* 15 (14): 1771. https://doi.org/10.2217/PGS.14.117.

Fabbro, Cristian Del, Simone Scalabrin, Michele Morgante, and Federico M Giorgi. 2013. “An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis.” *PloS One* 8 (12): e85024. https://doi.org/10.1371/journal.pone.0085024.

Faux, Thomas, Kalle T. Rytkönen, Asta Laiho, and Laura L. Elo. 2019. “RepViz: A Replicate-Driven R Tool for Visualizing Genomic Regions.” *BMC Research Notes* 12 (1): 441. https://doi.org/10.1186/s13104-019-4473-z.

Faux, Thomas, Kalle T Rytkönen, Mehrad Mahmoudian, Niklas Paulin, Sini Junttila, Asta Laiho, and Laura L Elo. 2021. “Differential ATAC-Seq and ChIP-Seq Peak Detection Using ROTS.” *NAR Genomics and Bioinformatics* 3 (3). https://doi.org/10.1093/NARGAB/LQAB059.

Fejes, Anthony P., Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven J. M. Jones. 2008. “FindPeaks 3.1: A Tool for Identifying Areas of Enrichment from Massively Parallel Short-Read Sequencing Technology.” *Bioinformatics* 24 (15): 1729. https://doi.org/10.1093/BIOINFORMATICS/BTN305.

“Finishing the Euchromatic Sequence of the Human Genome.” 2004. *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.

Flensburg, Christoffer, Sarah A. Kinkel, Andrew Keniry, Marnie E. Blewitt, and Alicia Oshlack. 2014. “A Comparison of Control Samples for ChIP-Seq of Histone Modifications.” *Frontiers in Genetics* 0 (SEP): 329. https://doi.org/10.3389/FGENE.2014.00329.

Fontenot, Jason D., Marc A. Gavin, and Alexander Y. Rudensky. 2003. “Foxp3 Programs the Development and Function of CD4+CD25+ Regulatory T Cells.” *Nature Immunology 2003 4:4* 4 (4): 330–36. https://doi.org/10.1038/ni904.

Fujita, Pauline A, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, et al. 2011. “The UCSC Genome Browser Database: Update 2011.” *Nucleic Acids Research* 39 (Database issue): D876-82. https://doi.org/10.1093/nar/gkq963.

Gates, Leah A, Charles E Foulds, and Bert W O’Malley. 2017. “Histone Marks in the ‘Driver’s Seat’: Functional Roles in Steering the Transcription Cycle.” *Trends in Biochemical Sciences* 42 (12): 977–89. https://doi.org/10.1016/j.tibs.2017.10.004.

Germain, Ronald N. 2002. “T-Cell Development and the CD4–CD8 Lineage Decision.” *Nature Reviews Immunology 2002 2:5* 2 (5): 309–22. https://doi.org/10.1038/nri798.

Guo, Yan, Fei Ye, Quanghu Sheng, Travis Clark, and David C. Samuels. 2014. “Three-Stage Quality Control Strategies for DNA Re-Sequencing Data.” *Briefings in Bioinformatics* 15 (6): 879–89. https://doi.org/10.1093/BIB/BBT069.

Hach, Faraz, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E Eichler, and S Cenk Sahinalp. 2010. “MrsFast: A Cache-Oblivious Algorithm for Short-Read Mapping.” *Nature Methods* 7 (8): 576. https://doi.org/10.1038/NMETH0810-576.

Hahne, Florian, and Robert Ivanek. 2016. “Visualizing Genomic Data Using Gviz and Bioconductor.” In , 335–51. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-3578-9_16.

Handy, Diane E., Rita Castro, and Joseph Loscalzo. 2011. “Epigenetic Modifications.” *Circulation* 123 (19): 2145–56. https://doi.org/10.1161/CIRCULATIONAHA.110.956839.

Harmanci, Arif, Joel Rozowsky, and Mark Gerstein. 2014. “MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments Using a Mappability-Corrected Multiscale Signal Processing Framework.” *Genome Biology* 15 (10): 474. https://doi.org/10.1186/s13059-014-0474-3.

Hatem, Ayat, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek. 2013. “Benchmarking Short Sequence Mapping Tools.” *BMC Bioinformatics* 14 (1): 184. https://doi.org/10.1186/1471-2105-14-184.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89. https://doi.org/10.1016/J.MOLCEL.2010.05.004.

Hori, Shohei, Takashi Nomura, and Shimon Sakaguchi. 2017. "Control of Regulatory T Cell Development by the Transcription Factor Foxp3." *Journal of Immunology* 198 (3): 981–85. https://doi.org/10.1126/SCIENCE.1079490.

Hung, Jui-Hung, and Zhiping Weng. 2017. "Analysis of Microarray and RNA-Seq Expression Profiling Data." *Cold Spring Harbor Protocols* 2017 (3): pdb.top093104. https://doi.org/10.1101/pdb.top093104.

"Initial Sequencing and Analysis of the Human Genome." 2001. *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Israel, Jennifer W., Grace A. Chappell, Jeremy M. Simon, Sebastian Pott, Alexias Safi, Lauren Lewis, Paul Cotney, et al. 2018. "Tissue- and Strain-Specific Effects of a Genotoxic Carcinogen 1,3-Butadiene on Chromatin and Transcription." *Mammalian Genome* 29 (1–2): 153–67. https://doi.org/10.1007/s00335-018-9739-6.

Ji, Hongkai, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. 2008. "An Integrated Software System for Analyzing ChIP-Chip and ChIP-Seq Data." *Nature Biotechnology* 26 (11): 1293–1300. https://doi.org/10.1038/nbt.1505.

Josefowicz, Steven Z., Li-Fan Lu, and Alexander Y. Rudensky. 2012. "Regulatory T Cells: Mechanisms of Differentiation and Function." *Annual Review of Immunology* 30 (April): 531. https://doi.org/10.1146/ANNUREV.IMMUNOL.25.022106.141623.

Kaisers, Wolfgang, Heiner Schaal, and Holger Schwender. 2015. "Rbamtools: An R Interface to Samtools Enabling Fast Accumulative Tabulation of Splicing Events over Multiple RNA-Seq Samples." *Bioinformatics* 31 (10): 1663–64. https://doi.org/10.1093/bioinformatics/btu846.

Karlić, Rosa, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. 2010. "Histone Modification Levels Are Predictive for Gene Expression." *Proceedings of the National Academy of Sciences of the United States of America* 107 (7): 2926–31. https://doi.org/10.1073/pnas.0909344107.

Karmodiya, Krishanpal, Saurabh J. Pradhan, Bhagyashree Joshi, Rahul Jangid, Puli Chandramouli Reddy, and Sanjeev Galande. 2015. "A Comprehensive Epigenome Map of Plasmodium Falciparum Reveals Unique Mechanisms of Transcriptional Regulation and Identifies H3K36me2 as a Global Mark of Gene Suppression." *Epigenetics & Chromatin* 8 (1): 32. https://doi.org/10.1186/s13072-015-0029-1.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12 (6): 996–1006. https://doi.org/10.1101/gr.229102.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12 (6): 996–1006. https://doi.org/10.1101/gr.229102.

Khattri, Roli, Tom Cox, Sue-Ann Yasayko, and Fred Ramsdell. 2003. "An Essential Role for Scurfin in CD4+CD25+ T Regulatory Cells." *Nature Immunology 2003 4:4* 4 (4): 337–42. https://doi.org/10.1038/ni909.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology* 14 (4): 1–13. https://doi.org/10.1186/GB-2013-14-4-R36/FIGURES/6.

Kornberg, Roger D. 1974. "Chromatin Structure: A Repeating Unit of Histones and DNA." *Science* 184 (4139): 868–71. https://doi.org/10.1126/science.184.4139.868.

Kouzarides, Tony. 2007. “Chromatin Modifications and Their Function.” *Cell* 128 (4): 693–705. https://doi.org/10.1016/J.CELL.2007.02.005.

Laajala, Teemu D, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. 2009. “A Practical Comparison of Methods for Detecting Transcription Factor Binding Sites in ChIP-Seq Experiments.” *BMC Genomics* 10 (1): 618. https://doi.org/10.1186/1471-2164-10-618.

Landt, Stephen G, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, et al. 2012. “ChIP-Seq Guidelines and Practices of the ENCODE and ModENCODE Consortia.” *Genome Research* 22 (9): 1813–31. https://doi.org/10.1101/gr.136184.111.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. “Software for Computing and Annotating Genomic Ranges.” Edited by Andreas Prlic. *PLoS Computational Biology* 9 (8): e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

Lee, Pearl, Navdeep S. Chandel, and M. Celeste Simon. 2020. “Cellular Adaptation to Hypoxia through Hypoxia Inducible Factors and Beyond.” *Nature Reviews Molecular Cell Biology* 21 (5). https://doi.org/10.1038/s41580-020-0227-y.

Levene, P.A. 1919. “THE STRUCTURE OF YEAST NUCLEIC ACID.” *Journal of Biological Chemistry* 40 (2): 415–24. https://doi.org/10.1016/s0021-9258(18)87254-4.

Li, H., and N. Homer. 2010. “A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing.” *Briefings in Bioinformatics* 11 (5): 473–83. https://doi.org/10.1093/bib/bbq015.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 25 (14): 1754. https://doi.org/10.1093/BIOINFORMATICS/BTP324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng, Jue Ruan, and Richard Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores.” *Genome Research* 18 (11): 1851–58. https://doi.org/10.1101/GR.078212.108.

Liang, Kun, and Sündüz Keleş. 2012. “Detecting Differential Binding of Transcription Factors with ChIP-Seq.” *Bioinformatics* 28 (1): 121–22. https://doi.org/10.1093/bioinformatics/btr605.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2013. “The Subread Aligner: Fast, Accurate and Scalable Read Mapping by Seed-and-Vote.” *Nucleic Acids Research* 41 (10): e108. https://doi.org/10.1093/NAR/GKT214.

———. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/BIOINFORMATICS/BTT656.

Lio, Chan-Wang Joaquim, and Chyi-Song Hsieh. 2008. “A Two-Step Process for Thymic Regulatory T Cell Development.” *Immunity* 28 (1): 100–111. https://doi.org/10.1016/J.IMMUNI.2007.11.021.

Liu, Bin, Jimmy Yi, Aishwarya SV, Xun Lan, Yilin Ma, Tim HM Huang, Gustavo Leone, and Victor X Jin. 2013. “QChIPat: A Quantitative Method to Identify Distinct Binding Patterns for Two Biological ChIP-Seq Samples in Different Experimental Conditions.” *BMC Genomics* 14 (Suppl 8): S3. https://doi.org/10.1186/1471-2164-14-S8-S3.

Liu, Xiaoyu, Chenfei Wang, Wenqiang Liu, Jingyi Li, Chong Li, Xiaochen Kou, Jiayu Chen, et al. 2016. “Distinct Features of H3K4me3 and H3K27me3 Chromatin Domains in Pre-Implantation Embryos.” *Nature 2016 537:7621* 537 (7621): 558–62. https://doi.org/10.1038/nature19362.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Mahony, Shaun, Matthew D. Edwards, Esteban O. Mazzoni, Richard I. Sherwood, Akshay Kakumanu, Carolyn A. Morrison, Hynek Wichterle, and David K. Gifford. 2014. "An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding." Edited by Ilya Ioshikhes. *PLoS Computational Biology* 10 (3): e1003501. https://doi.org/10.1371/journal.pcbi.1003501.

Marinov, Georgi K, Anshul Kundaje, Peter J Park, and Barbara J Wold. 2014. "Large-Scale Quality Analysis of Published ChIP-Seq Data." *G3 (Bethesda, Md.)* 4 (2): 209–23. https://doi.org/10.1534/g3.113.008680.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17 (1): 10. https://doi.org/10.14806/ej.17.1.200.

Maybin, Jacqueline A., Alison A. Murray, Philippa T. K. Saunders, Nikhil Hirani, Peter Carmeliet, and Hilary O. D. Critchley. 2018. "Hypoxia and Hypoxia Inducible Factor-1α Are Required for Normal Endometrial Repair during Menstruation." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-017-02375-6.

Mazur, Erik C., Yasmin M. Vasquez, Xilong Li, Ramakrishna Kommagani, Lichun Jiang, Rui Chen, Rainer B. Lanz, Ertug Kovanci, William E. Gibbons, and Francesco J. DeMayo. 2015. "Progesterone Receptor Transcriptome and Cistrome in Decidualized Human Endometrial Stromal Cells." *Endocrinology* 156 (6): 2239–53. https://doi.org/10.1210/en.2014-1566.

McKinnon, Katherine M. 2018. "Flow Cytometry: An Overview." *Current Protocols in Immunology* 120 (February): 5.1.1. https://doi.org/10.1002/CPIM.40.

McLean, Cory Y, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." *Nature Biotechnology 2010 28:5* 28 (5): 495–501. https://doi.org/10.1038/nbt.1630.

Micsinai, Mariann, Fabio Parisi, Francesco Strino, Patrik Asp, Brian D. Dynlacht, and Yuval Kluger. 2012. "Picking ChIP-Seq Peak Detectors for Analyzing Chromatin Modification Experiments." *Nucleic Acids Research* 40 (9): e70–e70. https://doi.org/10.1093/nar/gks048.

Nelson, Christopher J., Helena Santos-Rosa, and Tony Kouzarides. 2006. "Proline Isomerization of Histone H3 Regulates Lysine Methylation and Gene Expression." *Cell* 126 (5): 905–16. https://doi.org/10.1016/J.CELL.2006.07.026.

Okitsu, Cindy Yen, John Cheng Feng Hsieh, and Chih-Lin Hsieh. 2010. "Transcriptional Activity Affects the H3K4me3 Level and Distribution in the Coding Region." *Molecular and Cellular Biology* 30 (12): 2933–46. https://doi.org/10.1128/mcb.01478-09.

Ooi, Wen Fong, Manjie Xing, Chang Xu, Xiaosai Yao, Muhammad Khairul Ramlee, Mei Chee Lim, Fan Cao, et al. 2016. "Epigenomic Profiling of Primary Gastric Adenocarcinoma Reveals Super-Enhancer Heterogeneity." *Nature Communications* 7 (September): 12983. https://doi.org/10.1038/ncomms12983.

Park, Peter J. 2009. "ChIP–Seq: Advantages and Challenges of a Maturing Technology." *Nature Reviews Genetics* 10 (10): 669–80. https://doi.org/10.1038/nrg2641.

Park, Shinae, Go Woon Kim, So Hee Kwon, and Jung-Shin Lee. 2020. "Broad Domains of Histone H3 Lysine 4 Trimethylation in Transcriptional Regulation and Disease." *The FEBS Journal* 287 (14): 2891–2902. https://doi.org/10.1111/FEBS.15219.

Park, Sung Ho, Kyuho Kang, Eugenia Giannopoulou, Yu Qiao, Keunsoo Kang, Geonho Kim, Kyung-Hyun Park-Min, and Lionel B Ivashkiv. 2017. "Type I Interferons and the Cytokine TNF Cooperatively Reprogram the Macrophage Epigenome to Promote Inflammatory Activation." *Nature Immunology* 18 (10): 1104–16. https://doi.org/10.1038/ni.3818.

Pekowska, Aleksandra, Touati Benoukraf, Joaquin Zacarias-Cabeza, Mohamed Belhocine, Frederic Koch, Hélàne Holota, Jean Imbert, Jean Christophe Andrau, Pierre Ferrier, and Salvatore

Spicuglia. 2011. “H3K4 Tri-Methylation Provides an Epigenetic Signature of Active Enhancers.” *EMBO Journal* 30 (20): 4198–4210. https://doi.org/10.1038/emboj.2011.295.

Pepke, Shirley, Barbara Wold, and Ali Mortazavi. 2009. “Computation for ChIP-Seq and RNA-Seq Studies.” *Nature Methods 2009 6:11* 6 (11): S22–32. https://doi.org/10.1038/nmeth.1371.

Phanstiel, Douglas H, Alan P Boyle, Carlos L Araya, and Michael P Snyder. 2014. “Sushi.R: Flexible, Quantitative and Integrative Genomic Visualizations for Publication-Quality Multi-Panel Figures.” *Bioinformatics (Oxford, England)* 30 (19): 2808–10. https://doi.org/10.1093/bioinformatics/btu379.

Piovesan, Allison, Maria Chiara Pelleri, Francesca Antonaros, Pierluigi Strippoli, Maria Caracausi, and Lorenza Vitale. 2019. “On the Length, Weight and GC Content of the Human Genome.” *BMC Research Notes* 12 (1): 106. https://doi.org/10.1186/s13104-019-4137-z.

Planet, Evarist, Camille Stephan-Otto Attolini, Oscar Reina, Oscar Flores, and David Rossell. 2012. “HtSeqTools: High-Throughput Sequencing Quality Control, Processing and Visualization in R.” *Bioinformatics* 28 (4): 589–90. https://doi.org/10.1093/bioinformatics/btr700.

Polit, Lélia, Gwenneg Kerdivel, Sebastian Gregoricchio, Michela Esposito, Christel Guillouf, and Valentina Boeva. 2021. “CHIPIN: ChIP-Seq Inter-Sample Normalization Based on Signal Invariance across Transcriptionally Constant Genes.” *BMC Bioinformatics* 22 (1): 1–14. https://doi.org/10.1186/S12859-021-04320-3/FIGURES/5.

Qin, Qian, Shenglin Mei, Qiu Wu, Hanfei Sun, Lewyn Li, Len Taing, Sujun Chen, et al. 2016. “ChiLin: A Comprehensive ChIP-Seq and DNase-Seq Quality Control and Analysis Pipeline.” *BMC Bioinformatics 2016 17:1* 17 (1): 1–13. https://doi.org/10.1186/S12859-016-1274-4.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/BIOINFORMATICS/BTQ033.

R. Stark, G. Brown. 2011. “DiffBind: Differential Binding Analysis of ChIP-Seq Data.” 2011. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.

Rao, Mohan S., Terry R. Van Vleet, Rita Ciurlionis, Wayne R. Buck, Scott W. Mittelstadt, Eric A. G. Blomme, and Michael J. Liguori. 2019. “Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies.” *Frontiers in Genetics* 9 (January): 636. https://doi.org/10.3389/fgene.2018.00636.

Rashid, Naim U, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. 2011. “ZINBA Integrates Local Covariates with DNA-Seq Data to Identify Broad and Narrow Regions of Enrichment, Even within Amplified Genomic Regions.” *Genome Biology* 12 (7): R67. https://doi.org/10.1186/gb-2011-12-7-r67.

Rhee, Ho Sung, and B. Franklin Pugh. 2012. “ChIP-Exo: A Method to Identify Genomic Location of DNA-Binding Proteins at Near Single Nucleotide Accuracy.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 0 21 (SUPPL.100). https://doi.org/10.1002/0471142727.MB2124S100.

Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. “Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel SequencingNo Title.” *Nature Methods* 4 (8): 651–57. https://doi.org/10.1038/nmeth1068.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26. https://doi.org/10.1038/nbt.1754.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D, and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Ross-Innes, Caryn S., Rory Stark, Andrew E. Teschendorff, Kelly A. Holmes, H. Raza Ali, Mark J. Dunning, Gordon D. Brown, et al. 2012. "Differential Oestrogen Receptor Binding Is Associated with Clinical Outcome in Breast Cancer." *Nature* 481 (7381): 389–93. https://doi.org/10.1038/nature10730.

Rozowsky, Joel, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. 2009. "PeakSeq Enables Systematic Scoring of ChIP-Seq Experiments Relative to Controls." *Nature Biotechnology* 27 (1): 66–75. https://doi.org/10.1038/nbt.1518.

Rye, Morten Beck, Pål Sætrom, and Finn Drabløs. 2011. "A Manually Curated ChIP-Seq Benchmark Demonstrates Room for Improvement in Current Peak-Finder Programs." *Nucleic Acids Research* 39 (4): e25–e25. https://doi.org/10.1093/nar/gkq1187.

Sakaguchi, Shinya, Matthias Hombauer, Ivan Bilic, Yoshinori Naoe, Alexandra Schebesta, Ichiro Taniuchi, and Wilfried Ellmeier. 2010. "The Zinc-Finger Protein MAZR Is Part of the Transcription Factor Network That Controls the CD4 versus CD8 Lineage Fate of Double-Positive Thymocytes." *Nature Immunology 2010 11:5* 11 (5): 442–48. https://doi.org/10.1038/ni.1860.

Sanger, F, S Nicklen, and A R Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Schweikert, Gabriele, Botond Cseke, Thomas Clouaire, Adrian Bird, and Guido Sanguinetti. 2013. "MMDiff: Quantitative Testing for Shape Changes in ChIP-Seq Data Sets." *BMC Genomics* 14 (1): 826. https://doi.org/10.1186/1471-2164-14-826.

Semenza, Gregg L. 2012. "Hypoxia-Inducible Factors in Physiology and Medicine." *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2012.01.021.

Shao, Zhen, Yijing Zhang, Guo-Cheng Yuan, Stuart H Orkin, and David J Waxman. 2012. "MAnorm: A Robust Model for Quantitative Comparison of ChIP-Seq Data Sets." *Genome Biology* 13 (3): R16. https://doi.org/10.1186/gb-2012-13-3-r16.

Shen, Li, Ning-Yi Shao, Xiaochuan Liu, Ian Maze, Jian Feng, and Eric J. Nestler. 2013. "DiffReps: Detecting Differential Chromatin Modification Sites from ChIP-Seq Data with Biological Replicates." Edited by Roberto Mantovani. *PLoS ONE* 8 (6): e65598. https://doi.org/10.1371/journal.pone.0065598.

Skene, Peter J., and Steven Henikoff. 2017. "An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites." *ELife* 6 (January). https://doi.org/10.7554/ELIFE.21856.

Skidmore, Zachary L., Alex H. Wagner, Robert Lesurf, Katie M. Campbell, Jason Kunisaki, Obi L. Griffith, and Malachi Griffith. 2016. "GenVisR: Genomic Visualizations in R." *Bioinformatics* 32 (19): 3012–14. https://doi.org/10.1093/bioinformatics/btw325.

Smolle, Michaela, and Jerry L. Workman. 2013. "Transcription-Associated Histone Modifications and Cryptic Transcription." *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1829 (1): 84–97. https://doi.org/10.1016/J.BBAGRM.2012.08.008.

Soneson, Charlotte, and Mauro Delorenzi. 2013. "A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data." *BMC Bioinformatics* 14 (1): 91. https://doi.org/10.1186/1471-2105-14-91.

Song, Qiang, and Andrew D. Smith. 2011. "Identifying Dispersed Epigenomic Domains from ChIP-Seq Data." *Bioinformatics* 27 (6): 870–71. https://doi.org/10.1093/bioinformatics/btr030.

spicuglia, salvatore, and Laurent Vanhille. 2012. "Chromatin Signatures of Active Enhancers." *Nucleus* 3 (2): 126–31. https://doi.org/10.4161/nucl.19232.

Starks, Rebekah R., Anilisa Biswas, Ashish Jain, and Geetu Tuteja. 2019. "Combined Analysis of Dissimilar Promoter Accessibility and Gene Expression Profiles Identifies Tissue-Specific Genes and Actively Repressed Networks." *Epigenetics & Chromatin* 12 (1): 16. https://doi.org/10.1186/s13072-019-0260-2.

Steinhauser, Sebastian, Nils Kurzawa, Roland Eils, and Carl Herrmann. 2016. “A Comprehensive Comparison of Tools for Differential ChIP-Seq Analysis.” *Briefings in Bioinformatics* 17 (6): bbv110. https://doi.org/10.1093/bib/bbv110.

Strahl, Brian D., and C. David Allis. 2000. “The Language of Covalent Histone Modifications.” *Nature*. Nature Publishing Group. https://doi.org/10.1038/47412.

Sturm, Marc, Christopher Schroeder, and Peter Bauer. 2016. “SeqPurge: Highly-Sensitive Adapter Trimming for Paired-End NGS Data.” *BMC Bioinformatics* 17 (1): 208. https://doi.org/10.1186/s12859-016-1069-7.

Sun, Kun. 2020. “Ktrim: An Extra-Fast and Accurate Adapter- and Quality-Trimmer for Sequencing Data.” Edited by Inanc Birol. *Bioinformatics* 36 (11): 3561–62. https://doi.org/10.1093/bioinformatics/btaa171.

Suomi, Tomi, Fatemeh Seyednasrollah, Maria K. Jaakkola, Thomas Faux, and Laura L. Elo. 2017. “ROTS: An R Package for Reproducibility-Optimized Statistical Testing.” Edited by Timothée Poisot. *PLOS Computational Biology* 13 (5): e1005562. https://doi.org/10.1371/journal.pcbi.1005562.

Szalkowski, A. M., and C. D. Schmid. 2011. “Rapid Innovation in ChIP-Seq Peak-Calling Algorithms Is Outdistancing Benchmarking Efforts.” *Briefings in Bioinformatics* 12 (6): 626–33. https://doi.org/10.1093/bib/bbq068.

Thomas, Reuben, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard. 2016. “Features That Define the Best ChIP-Seq Peak Calling Algorithms.” *Briefings in Bioinformatics* 18 (3): bbw035. https://doi.org/10.1093/bib/bbw035.

Tsompana, Maria, and Michael J Buck. 2014. “Chromatin Accessibility: A Window into the Genome.” *Epigenetics & Chromatin* 7 (1): 33. https://doi.org/10.1186/1756-8935-7-33.

Tu, Shiqi, Mushan Li, Haojie Chen, Fengxiang Tan, Jian Xu, David J. Waxman, Yijing Zhang, and Zhen Shao. 2021. “MAnorm2 for Quantitatively Comparing Groups of ChIP-Seq Samples.” *Genome Research* 31 (1): 131–45. https://doi.org/10.1101/gr.262675.120.

Tu, Shiqi, and Zhen Shao. 2017. “An Introduction to Computational Tools for Differential Binding Analysis with ChIP-Seq Data.” *Quantitative Biology* 5 (3): 226–35. https://doi.org/10.1007/s40484-017-0111-8.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2013. “A Model Based Criterion for Gene Expression Calls Using RNA-Seq Data.” *Theory in Biosciences = Theorie in Den Biowissenschaften* 132 (3): 159–64. https://doi.org/10.1007/S12064-013-0178-3.

Wang, Charles, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, et al. 2014. “The Concordance between RNA-Seq and Microarray Data Depends on Chemical Treatment and Transcript Abundance.” *Nature Biotechnology* 32 (9): 926–32. https://doi.org/10.1038/nbt.3001.

Watson, J. D., and F. H.C. Crick. 1953. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.

Wilbanks, Elizabeth G., and Marc T. Facciotti. 2010. “Evaluation of Algorithm Performance in ChIP-Seq Peak Detection.” Edited by Gert Jan C. Veenstra. *PLoS ONE* 5 (7): e11471. https://doi.org/10.1371/journal.pone.0011471.

Xu, Shiliyang, Sean Grullon, Kai Ge, and Weiqun Peng. 2014. “Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells.” *Methods in Molecular Biology (Clifton, N.J.)* 1150: 97–111. https://doi.org/10.1007/978-1-4939-0512-6_5.

Xu, Xiao, Yuanhao Zhang, Jennie Williams, Eric Antoniou, W McCombie, Song Wu, Wei Zhu, Nicholas O Davidson, Paula Denoya, and Ellen Li. 2013. “Parallel Comparison of Illumina RNA-Seq and Affymetrix Microarray Platforms on Transcriptomic Profiles Generated from 5-Aza-Deoxy-Cytidine Treated HT-29 Colon Cancer Cells and Simulated Datasets.” *BMC Bioinformatics* 14 (Suppl 9): S1. https://doi.org/10.1186/1471-2105-14-S9-S1.

Yang, Shang-Fang, Chia-Wei Lu, Cheng-Te Yao, and Chih-Ming Hung. 2019. “To Trim or Not to Trim: Effects of Read Trimming on the De Novo Genome Assembly of a Widespread East Asian Passerine, the Rufous-Capped Babbler (Cyanoderma Ruficeps Blyth).” *Genes* 10 (10). https://doi.org/10.3390/genes10100737.

Yang, Yajie, Justin Fear, Jianhong Hu, Irina Haecker, Lei Zhou, Rolf Renne, David Bloom, and Lauren M McIntyre. 2014. “Leveraging Biological Replicates to Improve Analysis in ChIP-Seq Experiments.” *Computational and Structural Biotechnology Journal* 9: e201401002. https://doi.org/10.5936/csbj.201401002.

Yin, Tengfei, Dianne Cook, and Michael Lawrence. 2012. “Ggbio: An R Package for Extending the Grammar of Graphics for Genomic Data.” *Genome Biology* 13 (8): R77. https://doi.org/10.1186/gb-2012-13-8-r77.

Zhang, Yanxiao, Yu-Hsuan Lin, Timothy D. Johnson, Laura S. Rozek, and Maureen A. Sartor. 2014. “PePr: A Peak-Calling Prioritization Pipeline to Identify Consistent or Differential Peaks from Replicated ChIP-Seq Data.” *Bioinformatics* 30 (18): 2568–75. https://doi.org/10.1093/bioinformatics/btu372.

Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137. https://doi.org/10.1186/gb-2008-9-9-r137.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ISBN 978-951-29-9188-4 (PRINT)
ISBN 978-951-29-9189-1 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)

Painosalama Oy, Turku, Finland 2023