Combining epigenomic and transcriptomic data for the analysis of allele-specific regulation in a heart failure model

Julian Rummel

Thesis for obtaining the academic degree

Master of Science

in the course of study Bioinformatics
of the faculty information technology and mathematics
of Goethe University Frankfurt am Main

Julian Rummel, 6673334
Goethe University Frankfurt am Main
March 2023

First reader:

Prof. Dr. Marcel H. Schulz

Computational Epigenomics and Systems Cardiology

Institute of Cardiovascular Regeneration

Uniklinikum and Goethe University Frankfurt am Main

Second reader:

Dr. Katharina Zarnack

RNA Bioinformatics

Buchmann Institut for Molecular Life Sciences

Goethe University Frankfurt am Main

Contents

List of Figures III

List of Tables IV

List of Listings V

Zusammenfassung VI

Abstract VIII

List of Abbreviations XII

1. Introduction 1

1.1. Background . . . 1
1.2. Current State of Research . . . 3
1.3. Experimental Setup . . . 8
1.4. Motivation . . . 12
1.5. Structure of Master Thesis . . . 14

2. Theory and Methods 15

2.1. Quality Control . . . 15
2.2. snakePipes . . . 19
2.3. DESeq2 . . . 24
2.4. Functional Classification of Differentially Expressed Genes . . . 25
2.5. Manual ATAC-seq Data Analysis Workflow . . . 27
2.6. Enhancer-Gene Interactions and Motif Enrichment Analysis . . . 34

3. Results 38

3.1. Quality Control . . . 39
3.2. mRNA-seq Analysis . . . 50
3.3. ATAC-seq Analysis . . . 59
3.4. Combining Epigenomic and Transciptomic Data . . . 62

4. Discussion and Outlook **67**
4.1. Discussion 67
4.2. Outlook 78
4.3. Conclusion and Perspective 79

Literature **81**

Appendix **93**

A. Commands and R-scripts to Perform mRNA-seq and ATAC-seq Analysis **93**
A.1. Commands for Quality Control on Console Level 93
A.2. Commands for snakePipes on Console Level - mRNA-seq 93
A.3. R-script Extract for Analysis of Differentially Expressed Genes - mRNA-seq 94
A.4. Commands for Manual Analysis of ATAC-seq Data on Console Level 95
A.5. R-script for Analysis of Differential Accessibility Regions - ATAC-seq 95
A.6. Command for Obtaining Enhancer-Gene Interactions and Motif Enrichment Analysis on Console Level 96
A.7. Session Info - R 96

B. Readme - Novogene Rawdata Explanation mRNA-seq and ATAC-seq **99**

C. Blacklist File - Mus musculus Index **102**

D. Ignore for Normalization File - Mus musculus Index **108**

List of Figures

1.1. Feeding Behaviour of Different Mouse Strains 9
1.2. Weight Gain Under Standard and High-Fat Diet (+ L-NAME) 10
1.3. HFpEF Development After 10 and 16 Weeks of High-Fat Diet 11

2.1. Workflow to Analyse ATAC-seq Data 29

3.1. Adapter Content mRNA-seq 40
3.2. Adapter Content ATAC-seq 41
3.3. Sequence Duplication Level mRNA-seq 43
3.4. Sequence Duplication Level ATAC-seq 44
3.5. Overrepresented Sequences mRNA-seq 45
3.6. featureCounts: Assignments for All Strains mRNA-seq 46
3.7. Overall Alignment Rate mRNA-seq 49
3.8. Overall Alignment Rate ATAC-seq 50
3.9. Venn Diagram of Differentially Expressed Genes Treated vs. Untreated 52
3.10. MA-plot Differentially Expressed Genes BL6 and CC 53
3.11. Heatmap Cluster of Differentially Expressed Genes 55
3.12. Linegraph for Each Cluster Treated vs. Untreated 56
3.13. Highlighted Terms of Functional Enrichment Analysis for Cluster 8 . 57
3.14. Highlighted Terms of Functional Enrichment Analysis for Cluster 4 . 58
3.15. Statistics - Peak Calling for BL6 and CC Samples 60
3.16. MA-plot Differential Accessibility Regions BL6 and CC 62

List of Tables

3.1. Motif Enrichment Analysis - BL6 64
3.2. Motif Enrichment Analysis - CC 65

Listings

A.1. Command to Execute FastQC 93
A.2. Command to Execute MultiQC 93
A.3. Command to Build Genome Index - snakePipes 93
A.4. Command to Execute mRNA-seq Workflow - snakePipes 94
A.5. R-Script: Obtain Differentially Expressed Genes using DESeq2 94
A.6. Command to Execute ATAC-seq Workflow 95
A.7. R-Script: Obtain Differential Accessibility Regions using DiffBind . . 95
A.8. Command to Execute STARE 96
A.9. Command to Perform Motif Enrichment Analysis - HOMER 96

Zusammenfassung

Das Verständnis der genetischen Grundlagen menschlicher Vielfalt und ihrer Beziehung zu Krankheiten ist ein wichtiges Ziel der medizinischen Forschung. Die Genexpressionsanalyse misst sowohl die Häufigkeit von Genprodukten auf transkriptionaler Ebene in einer bestimmten Zelle, als auch die in der Zelle herrschenden Umstände. Epigenetische Mechanismen regulieren genomische Prozesse wie die Reparatur, Replikation und Transkription von Desoxyribonukleinsäure, indem sie die Zugänglichkeit von Chromatin kontrollieren. Daher ist die Integration von 'Assay for Transposase-Accessible Chromatin with high-throughput sequencing' und 'RNA-sequencing' ein effizienter Ansatz zur Untersuchung des Zusammenhangs zwischen Veränderungen der Genexpressionsmuster und Genomstruktur, insbesondere durch Enhancer-Gen-Interaktionen. Neben ökologischen und physiologischen Faktoren spielt zusätzlich vertikaler Gentransfer eine entscheidende Rolle bei der Expression von Genen, welche aufgrund von unzulänglicher Datenqualität in dieser Arbeit nicht ausgiebig untersucht werden kann.

In dieser Studie wurden die unterschiedlichen Genexpressionsprofile von zwei Unterarten von Mäusen, *Mus musculus domesticus* und *Mus musculus castaneus*, und ihre Hybridisierung unter einem durch fettreiche Ernährung induzierten Herzinsuffizienzmodell verglichen. Die Analyse ergibt, dass *Mus musculus domesticus* eine signifikant stärkere Reaktion auf die fettreiche Ernährung und die Hemmung mitochondrialer reaktiver Sauerstoffspezies durch den L-NAME-Inhibitor zeigt. Eine Untersuchung funktioneller Gemeinsamkeiten von geclusterten Genexpressionsprofilen identifiziert biologische Prozesse, die für die metabolischen und physiologischen Reaktionen auf eine fettreiche Ernährung in beiden Stämmen verantwortlich sind. Die Motif-Anreicherungsanalyse von Enhancer-Gen-Interaktionen zeigt Transkriptionsfaktoren, die wiederum die beobachteten differenziell exprimierten Gene regulieren. Dazu gehören Zinkfinger Transkriptionsfaktoren, wie Sp1, die in beiden Stämmen identifiziert wurden, aber auch Stamm-spezifische Transkriptionsfaktoren, wie p53. *Mus musculus domesticus* weist eine Hochregulierung des Fettsäurestoffwechsels als Reaktion auf eine fettreiche Ernährung auf, während dies bei *Mus mus-*

culus castaneus nicht der Fall ist. Im Gegensatz dazu zeigt *Mus musculus castaneus* eine Hochregulierung von kardioprotektiven Mechanismen, wie Cadherin-vermittelte Adhäsions- und Steroidhormonrezeptorwege, die bei *Mus musculus domesticus* nicht signifikant erhöht sind.

Diese Ergebnisse liefern erste Einblicke in die unterschiedliche Anfälligkeit für Herzkrankheiten zwischen beiden Mausstämmen, die mit der genetischen Vererbung zusammenhängen, und können als Grundlage für weitere Studien zur allelspezifischen Expression und Prägung dienen. Dies kann letztlich zur Identifizierung neuer Ziele für therapeutische Interventionen sowie zur Entwicklung personalisierter medizinischer Ansätze für die Behandlung von Herzinsuffizienz führen.

Die Gesamtheit der Ergebnisse, einschließlich derjenigen, die nicht in dieser Arbeit vorgestellt werden, wurde zusammen mit dem entsprechenden Code in einem GitHub-Repository (`https://github.com/jurummel/Masterthesis_JRummel`) archiviert.

Abstract

Understanding the genetic basis of human variation and its relationship to diseases is a crucial goal in medical research. Gene expression analysis measures the abundance of gene products at the level of transcription in a specific cell and under particular conditions. Epigenetic mechanisms regulate genomic processes, such as desoxyribonucleic acid repair, replication, and transcription, by controlling chromatin accessibility. Therefore, the integration of Assay for Transposase-Accessible Chromatin with high-throughput sequencing and RNA-sequencing is a time-efficient approach to investigate the correlation between changes in gene expression patterns and genome structure, particularly by enhancer-gene interactions. In addition to ecological and physiological factors, genetic inheritance also plays a crucial role in gene expression, which can not be extensively studied in this work due to insufficient data quality.

In this study, the differential gene expression profiles of two subspecies of mice, *Mus musculus domesticus* and *Mus musculus castaneus*, and their hybridization under a high-fat diet-induced heart failure model are compared. The analysis reveals that *Mus musculus domesticus* shows a significantly stronger response to the high-fat diet and mitochondrial reactive oxygen species inhibition by L-NAME inhibitor. Functional enrichment analysis of clustered gene expression profiles identifies biological processes responsible for the metabolic and physiological responses in both strains. Motif enrichment analysis of enhancer-gene interactions exhibits transcription factors responsible for the observed differentially expressed genes. These include zinc-finger transcription factors, such as Sp1, identified in both strains, but also strain-specific transcription factors, such as p53. *Mus musculus domesticus* shows upregulation of fatty acid metabolism in response to a high-fat diet, while *Mus musculus castaneus* does not. In contrast, *Mus musculus castaneus* demonstrates upregulation of cardioprotective mechanisms, such as cadherin-mediated adhesion and steroid hormone receptor pathways, which are not significantly increased in *Mus musculus domesticus*.

These findings provide initial insights into the heart disease susceptibility in both

mouse strains linked to genetic inheritance and may serve as a foundation for further studies on allele-specific expression and imprinting. This can ultimately lead to the identification of novel targets for therapeutic intervention, as well as the development of personalized medical approaches for the treatment of heart failure.

The entirety of the results, including those that are not presented in this thesis, along with the corresponding code, were archived in a GitHub repository (`https://github.com/jurummel/Masterthesis_JRummel`).

Erklärung zur Abschlussarbeit

gemäß § 35, Abs. 16 der Ordnung für den Masterstudiengang Bioinformatik vom 17. Juni 2019:

Hiermit erkläre ich,

Herr Julian Rummel

Die vorliegende Arbeit habe ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Ebenso bestätige ich, dass diese Arbeit nicht, auch nicht auszugsweise, für eine andere Prüfung oder Studienleistung verwendet wurde.

Zudem versichere ich, dass die von mir eingereichten schriftlichen gebundenen Versionen meiner Masterarbeit mit der in elektronischer Form eingereichten Version dieser Masterarbeit übereinstimmen.

Frankfurt am Main, den

Acknowledgments

Before I address the actual topic of this master thesis I would like to take this time to thank all people who made all of this possible and helped me with all my questions.

First of all I would like to thank Prof. Dr. Marcel Schulz who gave me the possibility to edit this topic and took over supervision and assessment for my thesis. He always had the right solution ready for any difficulties that arose. For your excellent work and advice I want to thank you very much.

Also I would like to express a big thank you to Tamer Ali (rer. nat.) and Phillip Grote (PhD) who supervised the experiments and provided all data sets used in this study. They further had an open ear for all of my concerns and gave me a perfect introduction to the topic. Thank you very much.

Finally I would like to thank my family and friends who helped me a lot in this stressful phase, by always being there for me and proofreading my Master's thesis.

List of Abbreviations

ABC Activity-by-Contact

ATAC-seq Assay for Transposase-Accessible Chromatin with high-throughput sequencing

BL6 C57BL/6J - homozygous

bp base pair

BWA Burrows-Wheeler aligner

CC CAST/EiJ - homozygous

cDNA complementary DNA

ChIP-seq Chromatin ImmunoPrecipitation DNA-Sequencing

DAR Differential Accessibility Region

DEG Differentially Expressed Gene

DMR Differentially Methylated Region

DNA Desoxyribonucleic Acid

ECM extracellular matrix

FACS Fluorescence-Activated Cell Sorting

FDR False Discovery Rate

FRiP Fraction of Reads in Peaks

GO Gene Ontology

GTF General Transfer Format

HFpEF Heart Failure With Preserved Ejection Fraction

HTML Hypertext Markup Language

KEGG Kyoto Encyclopedia of Genes and Genomes

LFC log_2 fold change

miRNA microRNA

mRNA-seq messengerRNA-sequencing

mRNA messengerRNA

NGS Next Generation Sequencing

PCR Polymerase Chain Reaction

RNA-seq RNA-sequencing

RNA Ribonucleic Acid

ROS Mitochondrial reactive oxygen species

RRBS Reduced Representation Bisulfite-Seq

scRNA-seq Single cell RNA-Sequencing

SHR steroid hormone receptor

SNP Single Nucleotide Polymorphism

TAD Topologically Associating Domain

TF transcription factor

UMI Unique Molecular Identifier

WGBS Whole-Genome Bisulfite-sequencing

XB C57BL/6J (maternal) / CAST/EiJ (paternal) - heterozygous

YB C57BL/6J (paternal) / CAST/EiJ (maternal) - heterozygous

1. Introduction

1.1. Background

The heart failure syndrome is described to be an emerging epidemic. Nowadays, the number of heart failure patients continues to increase due to a growing and aging population. A global meta-analysis of echocardiographic screening studies of people aged 65 and over found that the prevalence of all types of heart failure is approximately 4.2%. The burden of risk factors and comorbidities is substantial and increasing, particularly for the elderly. Even though heart failure primarily affects older people, recent studies suggest that the burden of heart failure may be increasing among young people. Therefore, understanding underlying biological processes involved in heart failure and thereby finding treatment methods is irrefutably important. [1]

Following this thought, decoding the genetic basis of human variation and other organisms is a major goal in medical research, in order to further understand and cure diseases. Measuring the differences in gene expression levels is a good approach to identify variation within and between species. [2] Besides ecological or physiological factors, another relevant influence on gene expression is genetic inheritance.

A diploid organism has an equal genetic contribution from male and female parents due to vertical transmission through sexual reproduction. Thus, it appears that the genetic information of the offspring is shaped equally by maternal and paternal alleles. However, some genes deviate from the uniform distribution of parental alleles and preferentially express a maternal or paternal allele. This phenomenon is known as allele-specific expression. [3] One reason for an imbalanced allelic expression could be genetic imprinting, a process in which a germline marker suppresses

one of the parental allele during the development [4]. Hence, imprinting and the resulting allele-specific expression can lead to alteration of gene expression levels and phenotypic changes [5]. This in turn can affect the response to diseases, such as cardiovascular regeneration.

Gene expression profiles provide the abundance of genes in a cell at given points in time. Therefore, these profiles serve as a fundamental tool for comprehending triggered cellular processes [6]. Consequently, gene expression profiling plays an important role in gene expression data analysis. Especially for the distinction of diseased and non-diseased sample groups, gene expression reveals crucial information about underlying biological processes and cell states.

Epigenetic information, such as Desoxyribonucleic Acid (DNA) methylation, regulate genomic processes such as DNA repair, DNA replication, and transcription by controlling chromatin accessibility [7]. The function of genes is thus influenced by various epigenetic changes. By combining the analysis of gene expression with the analysis of chromatin accessibility, a good overall picture of the genome and the biological background processes can be obtained. Consequently, gene expression and open chromatin analysis prove to be important steps for allele-specific genomic analysis and finding significant differences between two sample groups.

Mice serve as model organisms especially in cardiovascular research [8], because they are biologically very similar to humans and get many of the same diseases for the same genetic reasons [9]. Therefore, gene expression profile data and chromatin accessibility are studied in this work for two different species of mice, *Mus musculus domesticus* (C57BL/6J), *Mus musculus castaneus* (CAST/EiJ), and their hybridization, by messengerRNA-sequencing (mRNA-seq) and Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq). The procedure consists of dividing the different strains into two groups, one treated and one untreated, and deliberately induce heart failure in the treated group with a high-fat diet. The untreated group serves as a control. By measuring gene expression and analysing epigenetic information, allele-specific expression or species-specific responses to the high-fat diet can be studied to reveal the underlying biological processes.

1.2. Current State of Research

1.2.1. Gene Expression

Gene expression is considered to be a fundamental activity of all organisms. Proteins, which are a product of gene expression, control most activities of an organism, including the process of gene expression itself [10]. Measurement of gene expression can be done by gene expression profiling, which is a method to identify all of the genes in a cell or tissue that produce messengerRNA (mRNA), thus carrying the genetic information for building proteins [11].

The regulation of genes defines the subsequent identity and status of a cell. Gene expression consists of two sequential steps: the transcription of DNA into Ribonucleic Acid (RNA) followed by the translation of RNA into proteins. Simply spoken, gene expression turns raw genetic information into functional units - proteins. In addition to protein synthesis, RNA is involved in further biological processes. RNAs that do not code for proteins are referred to as non-coding parts of the genome. These non-coding parts may be involved in diverse functions in regulating gene expression through silencing e.g. microRNAs (miRNAs), binding to proteins, and more [12]. Understanding gene expression helps to comprehend the evolution of life and to find cures for various diseases [13].

1.2.2. Gene Expression Analysis

Gene expression is the measurement of the RNA produced by a gene in a cell at specific points in time. The analysis of gene expression provides data to understand the behavioral processes of genes in the genome, which is crucial for current biological research [14]. Thousands of genes are analysed simultaneously to examine cell drug response, signal transduction, oncology, cardiovascular diseases, and others.

Basically, gene expression data is realized through RNA-sequencing (RNA-seq) to make quantitative and qualitative statements about gene activity. Thus, RNA samples are the data input and the resulting output is gene expression data. RNA-seq can be further differentiated into bulk RNA-seq and Single cell RNA-Sequencing

(scRNA-seq).

Bulk Expression Analysis - Bulk RNA-sequencing

Bulk RNA-seq experiments provide the analysis of gene expression of an entire sample without differentiating among cell-types within the sample [15]. These experiments indicate the presence and quantity of RNA in a sample at a given moment and thus measure the average expression level. Particularly, the Next Generation Sequencing (NGS)-method is useful for quantifying expression signatures from ensembles. For example, in disease studies and similarly in comparative transcriptomics [16]. General steps of bulk RNA-seq are RNA extraction by cell lysis (RNA purification), enrichment of the RNA of interest, fragmentation of base pairs, synthesis of complementary DNA (cDNA), library preparation to obtain cDNA with adapters, and sequencing [17].

Single Cell Analysis - Single cell RNA-Sequencing

NGS for genomics, transcriptomics, and epigenomics, increasingly sets its focus on the characterization of individual cells. In contrast to bulk sequencing, scRNA-seq is able to uncover complex and rare cell populations. In addition, it facilitates the identification of regulatory relationships between genes. Single-cell isolation is the first step towards acquiring transcriptome information from a single cell. The most efficient and most common method for single-cell isolation is Fluorescence-Activated Cell Sorting (FACS). [18] FACS is a specialized type of flow cytometry and sorts biological cells into different containers based on specific light scattering and fluorescent characteristics [19]. The following steps do not differ from bulk RNA-seq: reverse transcription into first strand cDNA, cDNA amplification, scRNA-seq library generation, and sequencing [18].

The following analysis involves a number of crucial steps to process the resulting data. It begins with sequencing reads from the sequencer, which are then aligned to a reference genome, if available. Afterwards, the read fragments mapped to each gene are counted, which leads to gene expression quantification. Consequently,

this results in a table of counts, which are normalized and statistically analysed in order to obtain the Differentially Expressed Genes (DEGs) between two conditions. Along with this workflow, the observer should perform a quality control to inspect the quality of the data and possibly determine laboratory changes to improve the quality of future datasets.[20]

There are many tools that can complete some of the steps listed. One of these is snakePipes [21].

1.2.3. ATAC-seq

DNA is folded around histone proteins to form nucleosomes and compact them into chromatin. This hierarchical packaging compartmentalizes inactive genomic regions and makes biologically active regions accessible for transcription. This includes promoters, enhancers or other regulatory elements. Epigenetic information, such as DNA methylation, nucleosome positioning, histone composition, and many more, determine cellular phenotypes. [22] Therefore, understanding epigenetic structure, i.e., investigating transcription factor binding, positions of modified nucleosomes, and accessibility of chromatin at regulatory elements, is necessary to obtain information about a cell [23]. ATAC-seq uses the hyperactive Tn5 transposase to concurrently cut and ligate adapters for high-throughput sequencing at accessible regions. This produces multidimensional assays of the regulatory landscape of chromatin. These data sets can then be divided into reads of two sizes, one shorter than the canonical length generally protected by a nucleosome and the other corresponding to the approximate length of DNA protected by a nucleosome. This separation provides the position of nucleosomes and nucleosome-free regions, or more conveniently, it shows the accessible regions. [22]
ATAC-seq follows this protocol [22]:

1. Tagmentation through Tn5
2. Purification of tagged DNA fragments
3. Polymerase Chain Reaction (PCR)-amplification

4. Sequencing using NGS

Analysis of the data may reveal even more information about the different mouse species and can be combined with mRNA-seq for a multiomic approach to study gene expression [24]. Finding open chromatin regions requires the use of a peak caller like MACS2 [25] to find peaks in the mapped reads and then overlap those peaks between two or more samples to group them into active regions [26].

The newly obtained epigenetic information coupled with various gene expression analyses such as allele-specific variations and other discoveries can then assemble a picture of the different mouse strains. These insights can, for example, uncover imprinting effects and ultimately take a step towards understanding and treating cardiovascular diseases.

1.2.4. Accessible Regions in Chromatin

Chromatin accessibility refers to the extent to which nuclear macromolecules interact with chromatin-wrapped DNA. This interaction is organized by nucleosomes and other chromatin-binding factors and results in varying degrees of wrapping by histones throughout the genome. In some regions, such as facultative and constitutive heterochromatin, histones wrap tightly around DNA. However, at regulatory loci, including intergenic regions and transcribed gene bodies, histones are depleted. Chromatin accessibility is a dynamic and critical aspect of the genome that influences chromatin organization and function. In addition, it reflects the regulatory potential of different regions of the genome. [27] Thus, the accessibility of binding sites for transcription factors (TFs) and the transcription machinery is affected by the placement of nucleosomes in chromatin. This in turn has implications on DNA-related processes such as transcription, DNA repair, replication, and recombination. The understanding of the relationship between nucleosome positioning and gene expression has been extended by experiments. Thereby it could be shown, that the process of transcriptional activation involves changes in nucleosome position, whereas the regulation of transcription in eukaryotes requires the dynamic repositioning of nucleosomes. [28] Therefore, profiling chromatin accessibility on

a genome-wide scale is an excellent tool for mapping putative regulatory elements in a cell type or cell state. Post-translational chemical modifications of chromatin, such as DNA methylation, are commonly related to chromatin accessibility and may reflect specific functions of genomic regions in the context of regulating gene expression. The initial alterations in chromatin accessibility are initiated by the attachment of TFs, which either displace histones or preferentially bind to their recognition sequence in nucleosomal DNA. These initial TFs, known as pioneer factors, create a nucleosome-depleted region that enables further TF binding and stabilization of the site. This can lead to the regulation of gene expression of target genes. Therefore, studying TF binding sites in accessible chromatin regions provides valuable insight into cell type-specific lineage factors and gene regulatory networks. [29]

ATAC-seq provides a simple and efficient solution to reveal specific chromatin configuration characteristic of a given cell type and to analyse changes in this configuration due to perturbation or disease [30]. Thus, to determine the differences between conditioned groups, e.g. treated and untreated, an appropriate analysis is to search for Differential Accessibility Regions (DARs).

1.2.5. Differential Analysis of Open Chromatin Regions

Identification of open chromatin regions requires a peak caller such as MACS2, which assesses local enrichment against the genomic background to generate peaks in the regions [31]. Peaks in multiple samples can be combined and subsequently analysed through differential analysis by estimating read count differences in open chromatin regions between two groups. Differential analysis of ATAC-seq data is usually performed using methods developed for DEGs analysis, such as edgeR [32] and DESeq2 [33], because the basic principles of differential analysis are the same for ATAC-seq and RNA-seq studies. First of all, most of the open chromatin regions are the same under the two conditions and only a small fraction of these regions can be identified as significantly different. Further, the distribution of reads among the open chromatin regions adheres a specific distribution, i.e. a negative binomial distribution. [34] DiffBind [35] is a suitable R package [36] to accomplish this task.

1.3. Experimental Setup

1.3.1. General Information

The house mouse *Mus musculus* is a basic laboratory model and one of the first taxa suitable for gene diversity studies [37]. It has also served as a primary model in biomedical, ecological, and evolutionary research [38]. In the past, studies of the taxonomic status, origin, and relationships among the various components of *Mus musculus* have slowly crystallized into a picture of a taxon with at least three major subspecies: *Mus musculus domesticus*, *Mus musculus musculus*, and *Mus musculus castaneus* [37]. More recently, some studies have taken advantage of the increasing geographic distribution of *Mus musculus* to investigate the genetics of phenotypic changes and adaptations associated with range expansion [39]. Furthermore, the analysis of two different subspecies and their hybridization is a suitable approach to study allele-specific changes and imprinting phenomena. *Mus musculus domesticus* (C57BL/6J) and *Mus musculus castaneus* (CAST/EiJ) are known to be separated from one another for about 0.6 to 1 million years [40], with approximately 22.6 million Single Nucleotide Polymorphisms (SNPs). In addition, there is a difference in heart function between these species. [41] Hence, the comparison of these species is suitable for the study of Heart Failure With Preserved Ejection Fraction (HFpEF) (diastolic heart failure), i.e. heart failure with preserved left ventricular function [42]. In terms of genomic differences, this is the main research objective of this work.

1.3.2. Diet

To acquire both treated and untreated samples, four distinct experimental setups were conducted by the Grote LAB [43] in cooperation with Novogene [44]. C57BL/6J - homozygous (BL6), CAST/EiJ - homozygous (CC), C57BL/6J (maternal) / CAST/EiJ (paternal) - heterozygous (XB), and C57BL/6J (paternal) / CAST/EiJ (maternal) - heterozygous (YB). For each strain there is a control group with a standard diet and a treated group with a high-fat diet and Mitochondrial reactive oxygen species (ROS) inhibition by L-NAME inhibitor. The feeding

behaviour in the different mouse strains is shown in fig. 1.1.

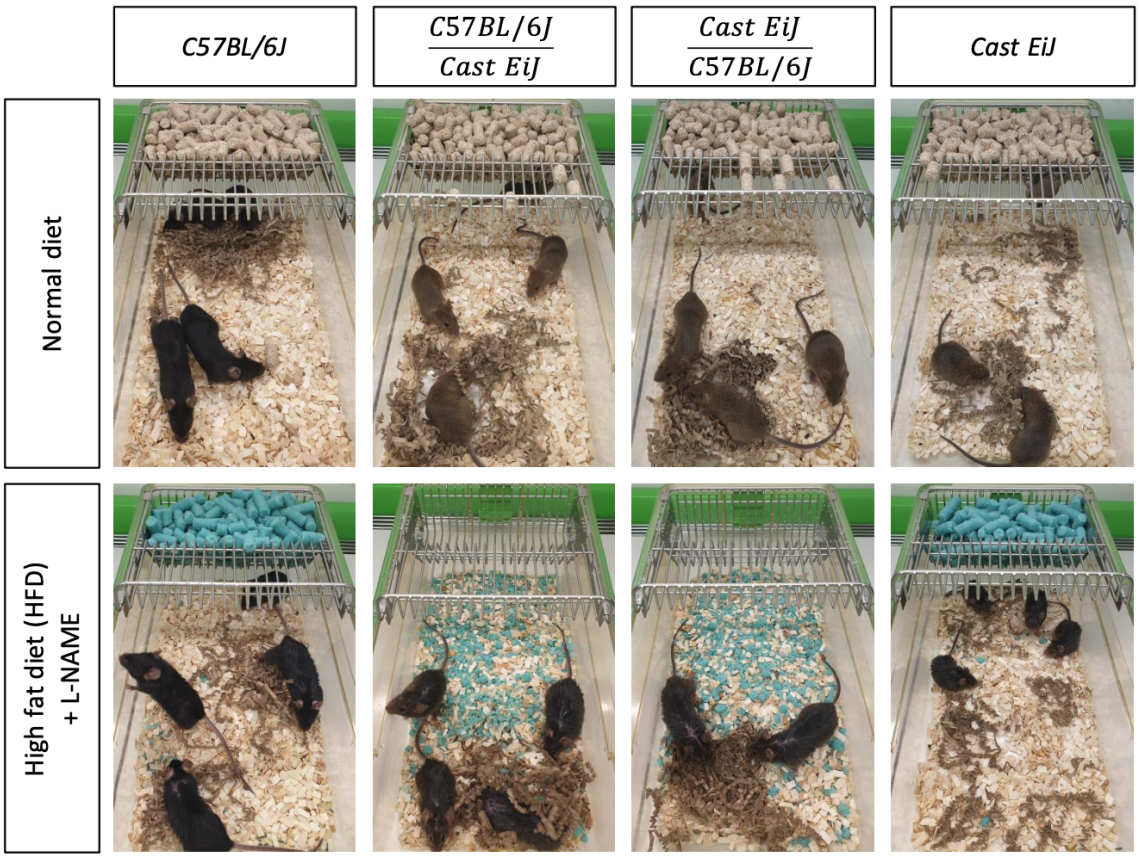


Figure 1.1.: Feeding Behaviour of Different Mouse Strains [41]: This figure shows the feeding behaviour of all four different mouse strains. The top line displays the mice with a normal diet and the bottom line those with a high-fat diet and ROS Inhibition by L-NAME Inhibitor. BL6, CC, XB, and YB were investigated. (This figure is provided by the Grote LAB [43].)

The high-fat diet has a significant effect on the mice, as clearly indicated by the weight gain (see fig. 1.2) and the development of HFpEF after 10 to 16 weeks (see fig. 1.3).

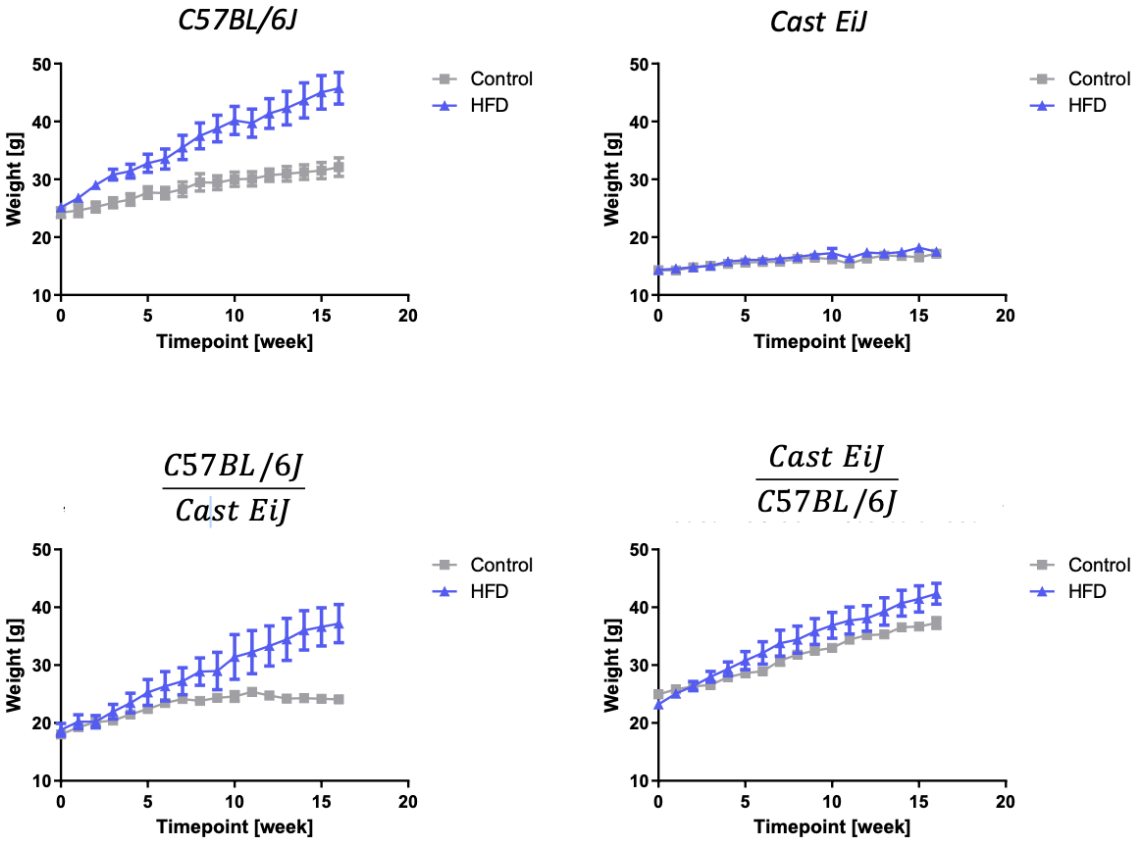
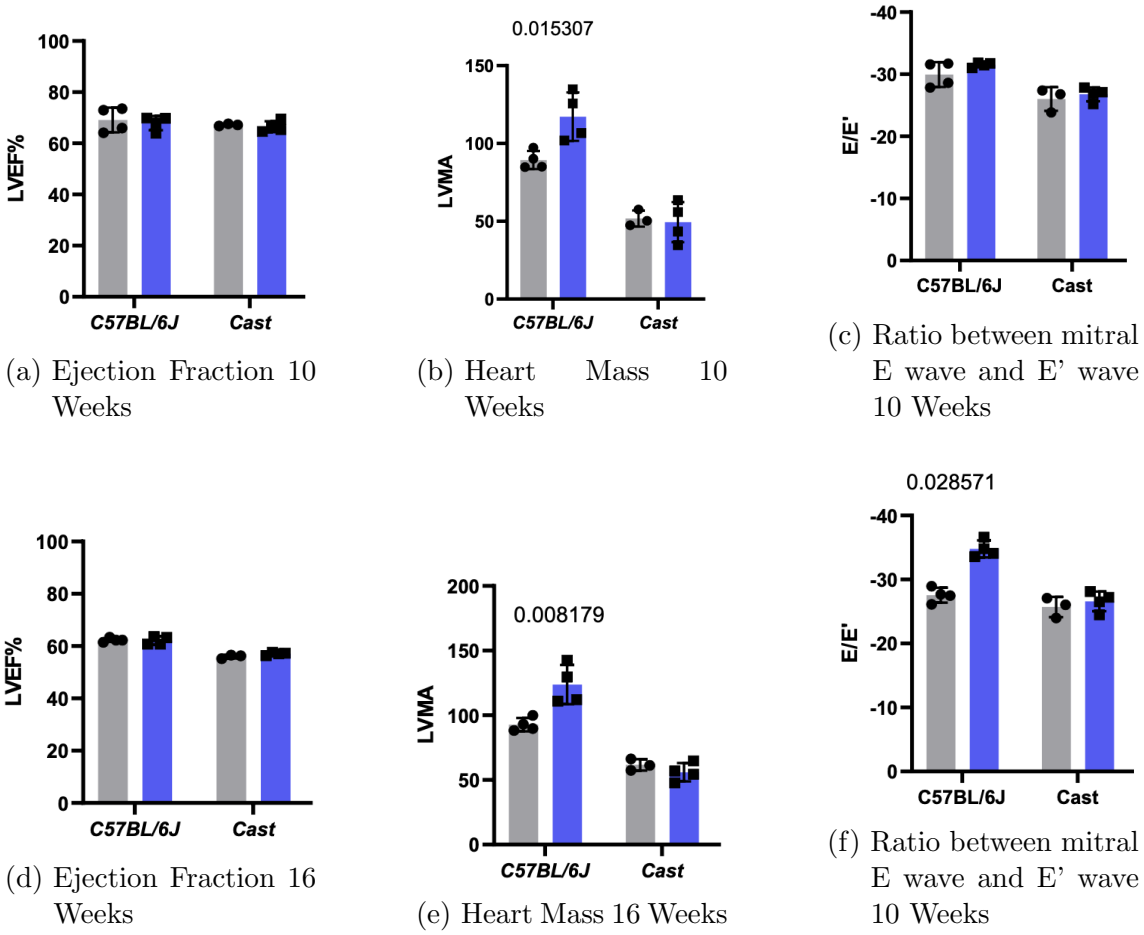


Figure 1.2.: Weight Gain Under Standard and High-Fat Diet (+ L-NAME) [41]: This figure shows the weight gain in grams of all four mouse strains at specific time points in weeks. The blue lines represent the samples with the high-fat diet and the grey lines are the samples with a standard diet. While the CC samples show no changes and the YB samples show small differences, BL6 and XB differ more between the control group and the high-fat diet group. (This figure is provided by the Grote LAB [43].)

Various methods can be utilized to detect and characterize patterns related to cardiovascular diseases, including HFpEF, in order to obtain information about heart function. For example, the ejection fraction (see fig. 1.3a and fig. 1.3d), which measures the amount of blood pumped by the heart during each contraction, the heart mass (see fig. 1.3b and fig. 1.3e), and the ratio between mitral E wave and E' wave or the ratio between early mitral inflow velocity and mitral annular early diastolic velocity to assess left ventricular filling pressure, respectively (see fig. 1.3c and fig. 1.3f) [45].



(a) Ejection Fraction 10 Weeks

(b) Heart Mass 10 Weeks

(c) Ratio between mitral E wave and E' wave 10 Weeks

(d) Ejection Fraction 16 Weeks

(e) Heart Mass 16 Weeks

(f) Ratio between mitral E wave and E' wave 10 Weeks

Figure 1.3.: HFpEF Development After 10 and 16 Weeks of High-Fat Diet [41]: The blue lines are the samples with the high-fat diet and the grey lines are the samples with a standard diet. (a) and (d) show the ejection fraction of the left ventricle measured as the difference between end-diastolic volume and end-systolic volume at 10 and 16 weeks. There is no major difference between the control groups and treated groups. (b) and (e) show the heart mass measured by left ventricular mass at 10 and 16 weeks. C57BL/6J shows a higher mass in the high-fat diet group, in contrast to CAST/EiJ, where there is no obvious difference. (c) and (f) show the ratio between early mitral inflow velocity and mitral annular early diastolic velocity to assess left ventricular filling pressure at 10 and 16 weeks. BL6 shows a higher ratio in the high-fat diet group, in contrast to CC, where no higher ratio is detected in the treated group. (This figure is provided by the Grote LAB [43].)

For further analysis, mRNA-seq and ATAC-seq were performed by Novogene [44] and provided for all samples to obtain a comprehensive understanding of gene expression and chromatin accessibility in all samples.

1.3.3. Dataset

The resulting dataset for subsequent analysis are mRNA-seq data and ATAC-seq data for 29 samples. There are eight samples for strain BL6, consisting of four treated and four control samples, designated by the abbreviation *BL6-XXX*. In addition, there are six samples for strain CC, consisting of four treated and two control samples, designated by the abbreviation *F2-XXXCC*. Also available are seven samples representing the strain XB, consisting of four treated and three control samples, designated by the abbreviation *F2-XXXXB*. There are eight samples covering the strain YB, consisting of four treated and four control samples, designated by the abbreviation *F2-XXXYB*. The raw sequencing data is in FASTQ format, compressed with gzip. *_1.fq.gz* and *_2.fq.gz* contain read 1 and read 2 for paired-end sequencing. Furthermore, a text file (*MD5.txt*) is provided, with the hash for the compressed FASTQ files. The MD5 hash can be used to verify the integrity of a file. If a file has been modified as a result of a corrupt file transfer, its MD5 hash would be changed (see appendix B). In this thesis, the terms *untreated* and *control* are used interchangeably.

1.4. Motivation

1.4.1. Benefits of Exploring ATAC-seq and mRNA-seq Data

The study of genetic differences between two subspecies of mice, *Mus musculus domesticus* (C57BL/6J) and *Mus musculus castaneus* (CAST/EiJ), and their hybridization may provide a deeper understanding of allele-specific changes and imprinting phenomena. Research has shown differences in cardiac function between these two species, making this analysis suitable for studying cardiovascular function and the underlying processes of heart failure. [41]

In addition to phenotypic changes, such as variation in heart mass and weight gain, deeper insights can only be achieved through genomic studies, especially ATAC-seq and mRNA-seq.

It is well known that ATAC-seq experiments help to understand the epigenetic

structure of a cell by identifying transcription factor binding sites, modified nucleosome positions, and accessible chromatin regions [23]. This in turn provides information about the gene function affected by the described interaction with chromatin, while the DNA sequence itself is not changed [46]. Furthermore, mRNA-seq experiments provide insights into the transcriptome of a cell. In particular, the resulting gene expression analysis is essential for interpreting the functional elements of the genome and understanding developments and diseases [47]. Combining these two assays in a multiomic approach can yield a more complete picture of the effects of heart failure, HFpEF in this case, in different mouse species. Especially for the comparison of two different groups, such as diseased and non-diseased.[24]

Categorizing DEGs into functional categories, including signaling pathways or biochemical pathways, can help to draw biological conclusions about the impact of heart failure on different mouse species [48]. This can be further evaluated by analysing the regulation of said DEGs by TFs. However, this requires sequencing assays such as ATAC-seq and mRNA-seq to be performed beforehand.

In conclusion, ATAC-seq and mRNA-seq are valuable tools for understanding significantly different biological processes between two sets of samples. This is demonstrated by their ability to classify gene groups and provide insights into gene expression and epigenetic structure.

1.4.2. Goals for Master Thesis

This thesis aims to explore the fundamental differences between the cardiac function of two subspecies of mice, *Mus musculus domesticus* (C57BL/6J) and *Mus musculus castaneus* (CAST/EiJ). In-depth analysis of various sequencing assays will provide important genomic insights to elucidate the differential response of the two mouse strains to the high-fat diet. Additionally, it is important to draw attention to this entire topic as it not only possesses significant future potential and relevancy for medical topics but for everyday life.

1.5. Structure of Master Thesis

This master thesis is divided into four chapters. The introduction (chapter 1) gives an insight into the topic and its relevance. In the following chapter 2, Theory and Methods, different analysis tools are compared and snakePipes among others is explained. Chapter 3 - Results, presents the quality of all given ATAC-seq and mRNA-seq samples and shows the results of differential analysis of gene expression and open chromatin regions. Chapter 4, Discussion and Outlook, discusses the results of this master's thesis and provides an outlook on how these results can be used for further studies.

2. Theory and Methods

2.1. Quality Control

NGS can be affected by various artifacts that occur during library preparation and the sequencing process, which can negatively impact the quality of the obtained data [49]. This in turn may have implications for subsequent analysis and biological conclusions. Therefore, it is crucial to be aware of possible errors in the data, either to correct them or consider them for biological interpretation and further analysis.

2.1.1. FastQC

FastQC [50] is a tool for quality control of raw sequencing data obtained by high-throughput sequencing. It provides several analysis functions that potentially present issues that should be considered before performing further analysis. Any variant of BAM, SAM or FASTQ files is accepted as input. The generated results can be exported to a permanent Hypertext Markup Language (HTML)-based report that contains summary graphs and tables for various areas such as basic statistics, per base sequence quality, per sequence quality score, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, kmer Content, and more. [50] More important for the present data, and consequently described in more detail, are the following quality features.

Adapter Content

Adapters are technical sequences and are ligated to both sides of the fragment of interest in paired-end experiments to serve as binding sites for priming [51]. Knowing

if the library contains a significant amount of adapters will be useful in assessing whether adapter trimming is necessary. If the insert-size, which is the size of the sequence of interest, of paired-end reads is shorter than the read length, the adapter-sequence will appear in the sequencing output [52]. This can lead to errors and a worse alignment rate [53]. Therefore, removing adapter contamination is a crucial step in downstream analysis. FastQC performs a dedicated search for a set of defined kmers and gives insight into the total proportion of the sequencing library that contains these kmers, specified as adapters [50].

This module emits a warning if any adapter occurs in more than 5% of all reads, and an error message if any adapter is present in more than 10% of all reads [50].

Sequence Duplication Level

Duplicates in a sequencing library are either technical duplicates caused by PCR artefacts, or biological duplicates that are natural collisions where various copies of the exact identical sequence have been randomly selected. Both types are equally reported as duplicates in this module. [50] PCR artefacts are polymerase errors, which generate sequence changes that are not present in the original samples [54]. These errors can propagate to later cycles of PCR, creating duplicates [55]. A high level of duplication could indicate some kind of enrichment bias and should hence be considered as a quality characteristic. However, sequences from different transcripts in the source population are present in very variable amounts in RNA-seq libraries. Thus, in order to be able to observe low-expressed transcripts, highly expressed transcripts are usually greatly over-sequenced. This leads to a large number of duplicates and thus to peaks in the duplication level diagram. These duplications originate from physically contiguous regions, but it is not possible to distinguish between over-sequencing and general technical duplication using FASTQ raw files. For this reason, a high duplication level in RNA-seq samples should neither be treated as a problem nor removed by deduplication. [50]

For memory efficiency, this module only analyses the sequences that occur in the first 100,000 sequences of each file, which should give a good indication of the duplication level in the entire file. Additionally, all sequences with more than 10

duplicates appear in grouped bins to provide a clear impression of the overall duplication level without requiring the presentation of each individual duplication value. A duplication is reported in the case of an exact sequence match across the entire sequence. Furthermore, all reads longer than 75 base pairs (bps) are trimmed to 50 bps. [50]

This module emits a warning if non-unique sequences account for more than 20.0% of the total and an error message if non-unique sequences account for more than 50.0% of the total. [50]

Overrepresented Sequences

Overrepresented sequences are individual sequences that occupy a significantly larger portion of the entire sequence set. A high level of representation may have a number of implications. On the one hand, it could indicate a highly biologically significant presence of the sequence. On the other hand, it could imply that the library is contaminated or that the diversity of the library is not as high as anticipated. However, in RNA-seq data it is expected that some transcripts are sufficiently abundant that they register as overrepresented. [50]

In this module, sequences that account for more than 0.1% of the total set are considered to be overrepresented. Similar to the sequence duplication level, only sequences that occur in the first 100,000 sequences are tracked to the end of the file and all reads longer than 75 bps are trimmed to 50 bps. Therefore, some sequences may be incorrectly classified as not overrepresented. [50]

This module provides a warning if any sequence represents more than 0.1% of the total, and an error message if any sequence represents more than 1.0% of the total [50].

FastQC can be executed directly at console level, as shown in listing A.1 (see appendix A.1), or even embedded in a workflow of snakePipes [21].

2.1.2. MultiQC

MultiQC [56] can be used to merge FastQC results from multiple samples and add more statistics, such as the overall alignment score of various aligners. For this, MultiQC scans given analysis directories for log files and quality reports. Thus, it automatically detects suitable files and creates concise plots. [56] The following quality characteristics are relevant to the present data and are therefore described in more detail.

featureCounts: Assignments

Essentially, featureCounts compares the mapping position of each read or fragment base to the genomic region spanned by each feature, considering any gaps (insertions, deletions, exon-exon junctions, or fusions) that may be present in the read. If there is an overlap of at least one base between the read or fragment and a feature, it is considered a hit and the read is then assigned to the corresponding gene. However, in some case the read may not be assigned for various reasons. One reason is the possibility of multiple overlaps with different features, complicating the determination of the true target gene. The lack of function or ambiguous nature of the read may be another reason for non-assignment. [57]

To assess the biological significance of the results, it is important to know the assignment percentage. This may explain unexpected results, e.g. from differential analysis.

Overall Alignment Rate

Aligning is a crucial step in any sequencing analysis, such as mRNA-seq and ATAC-seq. The statistics for paired-end data in Bowtie2 [58] include the following categories: The percentage of reads that are uniquely and concordantly aligned, reads that are concordantly aligned at more than one location, reads that are concordantly aligned exactly 1 time, and reads that are neither concordantly nor discordantly aligned 0, 1, or more than 1 time. If reads are aligned concordantly, it means that read 1 is in forward direction and read 2 in reverse, and the distance

between the reads is in the range of 500 bps. The overall alignment rate consists of all the single reads that are aligned at least once. [59] The mapping statistics of STAR [60] include reads that are mapped uniquely, mapped to multiple loci, mapped to too many loci, and reads that are not mapped for various reason. These statistics are calculated for each read (single- or paired-end) and then summed or averaged across all reads. Paired-end reads are counted as one read in STAR. [60] MultiQC processes this information and displays it as a barplot for each sample.

Comparable to the featureCount assignments, these information might reveal insights about the quality of the performed alignment and thus should be considered for interpretation.

MultiQC can be executed directly at console level to scan the provided path for suitable files or also embedded in a workflow of snakePipes [21], as shown in listing A.2 (see appendix A.1).

2.2. snakePipes

2.2.1. Motivation for Using snakePipes

SnakePipes is a software that contains a large number of different workflows for processing large datasets. It is very flexible in its ability to analyse data as it provides a set of workflows for processing, quality control, and downstream analysis of data from assays used in transcriptomic and epigenomic studies: scRNA-seq, mRNA-seq, ATAC-seq, Whole-Genome Bisulfite-sequencing (WGBS), Chromatin ImmunoPrecipitation DNA-Sequencing (ChIP-seq), Hi-C, and noncoding-RNA-seq [21].

In contrast to other software, snakePipes has very well-written documentation about the setup, various workflows, and more. snakePipes is also quite widely used, which makes it easy to get started and opens the possibility to get help and report issues in a forum (GitHub [61]). Moreover, it does not matter what kind of organism is being analysed, thanks to the possibility of creating any desired index in addition to some pre-defined ones. This makes it very practical to use for

almost any kind of application, and highlights the lack of a limitation. Furthermore, it is very straightforward to operate snakePipes at the console level, as there are console command examples for each workflow available. [21] Thus, working with the bioinformatics cluster (`https://schulzlab.github.io/Docu/`), which is needed to process a large amount of data, is not a big hurdle.

2.2.2. Workflows

Single cell RNA-Sequencing

This pipeline processes Unique Molecular Identifier (UMI)-based data and expects the cell barcode and UMI in Read 1, next to the cDNA sequence in Read 2. There are two analysis workflows available, one using STAR solo for mapping and quantification (STAR solo mode) and the other using Salmon [62] to create the count matrix (Alevin mode [63]). The STAR solo mode quantifies genic read counts at the single cell level and reads supporting spliced and unspliced transcripts in each cell (velocyto). This leads to seurat [64] objects for genic counts. Next to various quality control tables and plots, the output mainly consists of three files: barcodes.tsv, features.tsv, and matrix.mtx. On the other hand, the Alevin mode maps and generates a readcount matrix, estimates uncertainty of gene counts, performs general quality control, and quantifies spliced and unspliced read counts in each cell. The output provides, next to quality control files, raw and boostrapped count matrices, sample specific matrices, column data (barcodes), row data (genes), and Alevin spliced/unspliced counts for RNA velocity. [21]

messengerRNA-sequencing

This pipeline is of major importance for the analysis of the data in the present work and enables users to process single or paired-end mRNA-seq FASTQ files to obtain gene/transcript-counts and differential expression data. Additionally, it allows full allele-specific mRNA-seq analysis using the allelic-mapping mode. The required input is a directory of gzipped FASTQ files, containing either single or paired-end files. In addition to annotation, bamCoverage, and quality control files, the most

important file generated for this work is the featureCounts file, containing gene-level counts on the filtered General Transfer Format (GTF) files. The featureCounts file can be used for differential expression analysis. [21]

Assay for Transposase-Accessible Chromatin with high-throughput sequencing

The ATAC-seq workflow requires one or more BAM files (generated by an integrated DNA mapping workflow) in order to locate accessible regions and, if desired, a sample sheet to test for differential binding. BAM files are filtered to contain only correctly paired-end reads of appropriate fragment size. In addition, the differential open chromatin analysis generates a HTML report summarizing the analysis of DARs. This pipeline also produces various quality control files and files containing peaks detected by a selected peak caller (MACS2 [25], HMMRATAC [65], and Genrich [66]). [21]

Whole-Genome Bisulfite-sequencing

Reads are mapped to a reference genome using the bisulfite-specific aligner Burrows-Wheeler aligner (BWA)-meth [67]. FASTQ files and a genome alias are required for this purpose. Furthermore, the user can pass a sample sheet with grouping information to perform a differential methylation analysis. This analysis, together with a blacklist of genomic positions, maps Differentially Methylated Regions (DMRs) to known SNPs masking CpG methylation levels. Many quality control metrics, such as conversion rate, flagstat, depth of coverage, GC bias, and methylation bias calculations are generated during this workflow. As a result, the user receives BAM files with bwa-meth and PCR duplicate removal, as well as several DMR files. [21]

Chromatin ImmunoPrecipitation DNA-Sequencing

One or multiple BAM files are required to find peaks. Additionally, multiple samples and a sample sheet can be specified to retrieve differential peaks. These files can be generated by an integrated DNA mapping workflow. Other optional features include spikein normalisation, which provides spikein-derived scaling factors for ChIP

samples, and differential binding analysis between two groups of samples. Analogous to the ATAC-seq workflow, the output contains several quality control files and Genrich, MACS2, or histoneHMM [68] peaks, depending on the user's settings. [21]

Hi-C

This pipeline processes Hi-C data to obtain corrected Hi-C matrices and to call Topologically Associating Domains (TADs). The Hi-C workflow accepts FASTQ files with paired-end reads as input and assigns each read to a user-specified reference genome. BWA [69] is used for mapping, followed by HiCExplorer [70] for analysis. Important outputs are the contact matrices, the quality control measurements, the corrected matrix, and information about the called TADs (boundaries, domains, and scores). [21]

noncoding-RNA-sequencing

The noncoding-RNA-seq pipeline allows users to process FASTQ files containing single or paired-end ribosomal-depleted RNA-seq reads to the point of gene-, transcript-, repeat-element counts, and differential expression. The only aligner available is STAR. Differential expression can only be performed by loading a sample sheet. Besides quality checks such as insert size distribution and mapping statistics, bamCoverage files and various DESeq2 [33] files are generated if specified. [21]

2.2.3. Data Processing with snakePipes

Preparation

The snakePipes pipeline provides flexibility in selecting an appropriate reference genome for downstream analysis. In the present study, *Mus musculus domesticus* (C57BL/6J) and *Mus musculus castaneus* (CAST/EiJ) mRNA-seq data are available and a user defined reference index is generated for subsequent analysis. This requires two inputs: a FASTA file or URL containing the genome sequence and a GTF file or URL containing the gene annotation [21]. The genome and GTF file

are obtained from the ensembl genome browser (`http://www.ensembl.org/Mus_musculus/Info/Index`). Furthermore, a blacklist file (see appendix C) is specified, besides a file listing the chromosomes to be ignored during normalisation steps (see appendix D). This information could be utilised to create the index by running a command on console level, as defined in listing A.3 (see appendix A.2). This enables the allele-specific downstream analysis of the mRNA-seq data using the snakePipes mRNA-seq workflow.

Analysis

The analysis of the mRNA-seq data is conducted using the snakePipes pipeline described in section 2.2.2. A console level command is executed for each sample to prepare the required files for subsequent analysis, as shown in listing A.4 (see appendix A.2).

The input directory (*-i*) contains the paired-end FASTQ files for the sample, while the output directory (*-o*) specifies the location for the results. In addition, a file containing SNPs (*--VCFfile*) is provided to map the previously established genome index to the *Mus musculus domesticus* (C57BL/6J) and *Mus musculus castaneus* (CAST/EiJ) subspecies (*--strains*) and generating a N-masked genome. N-masking is the process of masking polymorphic genomic sites using the ambiguity nucleobase 'N', conducting a singular alignment to the N-masked genome, and subsequently assigning reads based on the nucleotide sequence identified beneath the masked positions [71]. The SNP-file is provided by Sanger (`https://www.sanger.ac.uk/data/mouse-genomes-project/`). The allelic-mapping mode, specified with the flag (*--mode*), initiates an alignment process to the allele-masked genome, followed by the partitioning of mapped files according to allele-specific information [21].

Several outputs are produced by this analysis. The most important file for further investigation is the featureCounts output, which contains gene-level counts on the filtered GTF files and can be used for differential expression analysis.

2.3. DESeq2

2.3.1. Idea

To further analyse the data generated by snakePipes, it is valuable to examine DEGs, which may provide insights into potential regulatory mechanisms underlying the observed differences in the high-fat diet response of BL6 and CC. A gene is differentially expressed if there is a statistically significant difference in read counts between two experimental conditions or genomes [72]. Thus, it reveals insights about biological differences between healthy and diseased states, for example, or allele-specific variations [73]. To accomplish this, it is crucial to perform a comprehensive analysis of the featureCount tables generated by the mRNA-seq workflow. An R-script [36] prepares the data and applies the R-package DESeq2 to perform the differential expression analysis.

2.3.2. Description

DESeq2 is a tool for differential analysis of count data, using various statistical methods such as shrinkage estimation and log_2 fold changes (LFCs) to detect dispersions and to improve the stability and interpretability of estimates [33]. The authors of DESeq2 aim to facilitate gene ranking and visualization based on stable estimation of effect sizes (LFCs) and to incorporate user-defined thresholds for biological significance in testing differential expression. This allows DESeq2 to perform consistently across a wide range of data types and is suitable for both small studies with few replicates and large observational studies. DESeq2 is available for R and therefore easy to integrate and work with. Thus, this tool is suitable for the gene-level analysis of mRNA-seq data and facilitates the creation of graphs and lists as output. [33]

This software package is utilised to obtain DEGs of the data provided (section 2.2.3), both in the genomes and between the treated and untreated sample groups. The application of DESeq2 is demonstrated in listing A.5 (see appendix A.3).

2.4. Functional Classification of Differentially Expressed Genes

Genes regulate biological functions and control the mechanisms of various diseases, including HFpEF. To explore underlying biological mechanisms of a disease, it is essential to study the transcriptome, i.e. classifying and clustering genes based on their expression patterns. DEGs are genes whose biological patterns change between healthy and diseased states and thus behave differently in disease conditions. [74]

One approach to classify a large number of DEGs is to cluster these genes into groups. However, there are many different algorithms that can be used for this task and one particular algorithm is hierarchical clustering.

2.4.1. Hierarchical Clustering

Clustering algorithms can either be hierarchical or partitional. The former method finds successive cluster by either considering each element as a separate cluster and merging them into successively larger clusters (agglomerative clustering) or dividing the entire set into successively smaller clusters (divisive clustering). [75]

A mandatory step is the selection of a measure to calculate the distance between two data points. A fairly common distance measure, that is used in this work, is the Euclidean distance, which calculates the square root of the sum of the squares of the differences between corresponding values, as denoted in Eq. (2.1) [75]:

$$d = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2} \tag{2.1}$$

In the case of two data points X and Y the distance is analogous to the determination of the length of the hypotenuse in a triangle [75].

At each step, the algorithm clusters the two elements that are closest in terms of distance measures and gradually builds a hierarchy from the individual elements [75]. A function performing agglomerative clustering in R is hclust, which is part of the 'stats' package [36]. This function uses the complete linkage method for determining

distances between cluster as default. In this particular method, the cluster distance between two clusters is the maximum Euclidean distance between their individual components. At each stage, the distances between clusters are recalculated using the Lance-Williams dissimilarity update formula. [36]

Clustering stops when the user defines that the clusters are too far apart to be merged or when there is a sufficiently small number of clusters [75].

2.4.2. Functional Enrichment Analysis

In this work, the clusters obtained represent groups of genes that are similar in terms of expression for each individual sample group. These clusters can be classified into simplified functional categories, usually representing signaling pathways or biochemical pathways curated based on information from the literature. Functional enrichment analysis and its validity depends on stringent statistical methods as well as both accurate and recent gene functional annotations. [48] A popular tool for identifying significantly enriched biological functions and pathways from well established data sources is gProfiler2 [76].

gProfiler2

This R package can handle gene lists for more than 600 species and strains. It identifies biological functions and pathways from reliable and regularly updated annotation data sources such as Gene Ontology (GO) [77][78][79], Kyoto Encyclopedia of Genes and Genomes (KEGG) [80], Reactome [81], WikiPathways [82], miRTarBase [83], TRANSFAC [84], Human Protein Atlas [85], protein complexes from CORUM [86], and Human Phenotype Ontology [87].

To begin with, the function 'gost' from gprofiler2 identifies the biological functions and pathways that are significantly enriched in a gene set. The output of the 'gost' function is a named list where the 'result' element contains a data frame with the enriched functions and associated statistics, such as the False Discovery Rate (FDR) corrected p-value. A Manhattan plot allows these results to be further visualised, even interactively. Each circle in this plot corresponds to a single term on the x-axis and the adjusted p-values on $-log_{10}$ scale on the y-axis. The most

important terms can be highlighted by the user. [76]

The results obtained provide information about the differences between the treated and untreated samples and indicate significantly affected biological functions and pathways.

2.5. Manual ATAC-seq Data Analysis Workflow

As mentioned in section 1.3.3, not only mRNA-seq but also ATAC-seq data are the result of the experimental setup. Accurate identification of regulatory regions involves analysing open chromatin regions [88]. This requires numerous critical steps to determine chromatin accessibility across the genome and to ensure that the resulting data is reliable and informative [22]. It is essential that quality control is performed beforehand, e.g. with FastQC, as written in section 2.1. Subsequently, the data can be further modified by trimming. This is recommended if the quality control step indicates a high level of PCR adapters [89]. One of the tools desinged for this purpose is Trim Galore [90] (see section 2.5.1). To determine the origin of the reads within the original genome, the reads obtained from the trimming step, e.g. the resulting FASTQ files, need to be aligned to a reference genome (*Mus musculus domesticus* and *Mus musculus castaneus*) [91]. Due to the ever-increasing computational burden of longer reads [92] and larger amounts of data, an efficient DNA mapping tool is essential [91], with Bowtie2 (see section 2.5.2) being a good choice. The aligner creates a BAM file that contains a header and an alignment section with information for each read pair [93]. High-throughput sequencing methods rely on PCR, which, among other biases, can lead to PCR duplicates, i.e. reads derived from the same original cDNA molecule via PCR [55]. Hence, it might be necessary to further modify the BAM file. SAMtools [94] comprises many functions including SAMtools fixmate or SAMtools markdup for sorting and filtering BAM files (see section 2.5.3). Finally, in order to identify open chromatin regions that are statistically significant after performing an ATAC-seq analysis, a peak-calling program such as MACS2 is required [95] (see section 2.5.4). To correlate the open chromatin regions

of the current analysis with the DEGs of the mRNA-seq analysis and to compare samples with two or more conditions, DiffBind [35] can be applied. This is suitable for determining DARs, in this case between treated and untreated samples, and can potentially reveal allele specific differences. Using DiffBind requires the provision of a sample sheet containing the information acquired during the analysis, such as BAM files and peaks for each sample, and additional information like condition and the peak caller used (see section 2.5.5). All differential peaks are displayed in a table containing the chromosome, start and end coordinates, and other important statistics including the p-value and FDR [35]. This table can be further filtered for statistically significant DARs and then compared to the DEGs from the mRNA-seq analysis. The entire workflow is shown in fig. 2.1.

To execute this workflow, several command-line level tasks and an R-script [36] are required, as shown in listing A.6 and listing A.7 (see appendix A.4 and appendix A.5).

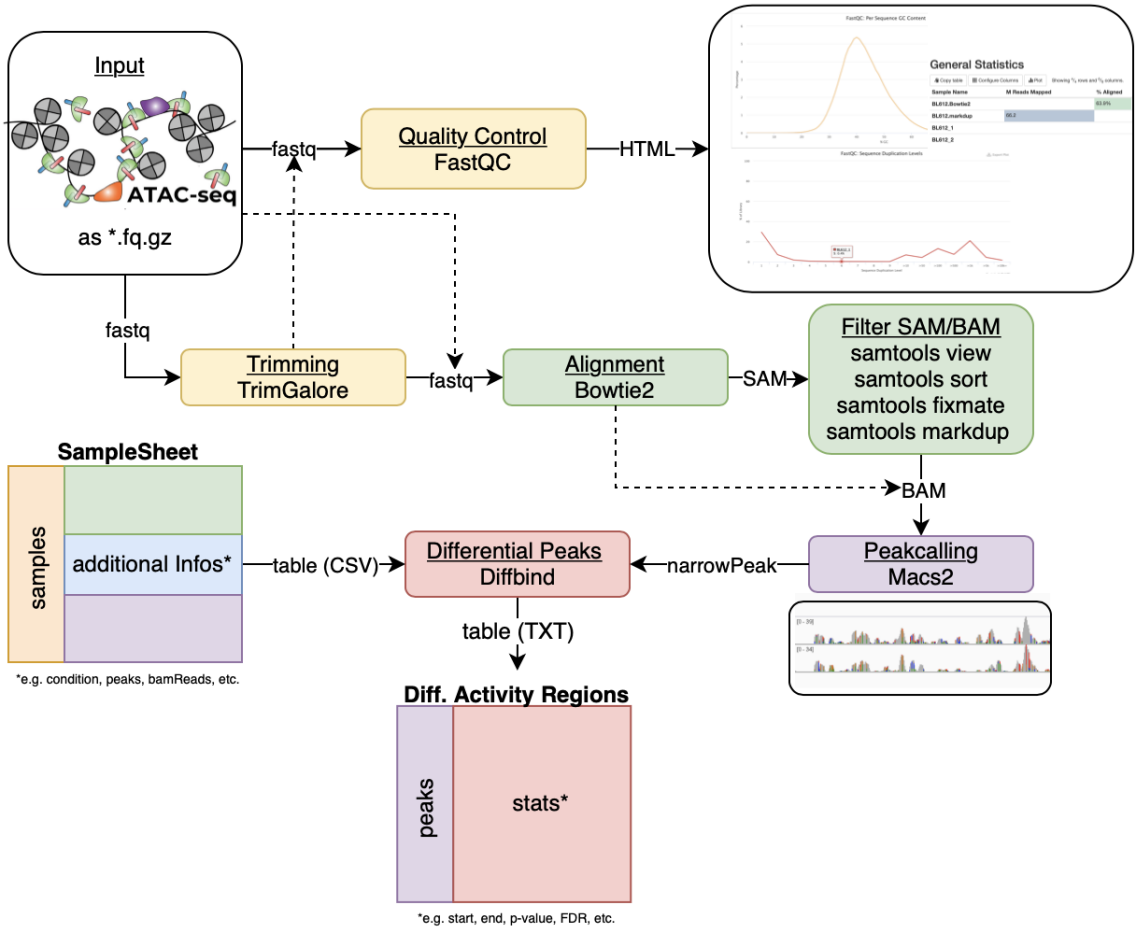


Figure 2.1.: Workflow to Analyse ATAC-seq Data: This figure shows the steps of ATAC-seq data analysis to determine open chromatin regions and DARs. The FASTQ files generated by ATAC-seq are checked for errors using FastQC. The extracted findings lead to adapter trimming with Trim Galore, followed by alignment with Bowtie2, resulting in a SAM file. To filter further, for example for PCR duplicates, SAMtools is used. The final BAM file is then analysed with MACS2 to generate peaks. Combined with a sample sheet, these peaks can be analysed with DiffBind to yield DARs between treated and untreated samples. Some steps, such as trimming and filtering, are not always necessary, but are essential for the data in this study. (Constructed with draw.io [96], external images used [97][98].)

2.5.1. Trim Galore

Trim Galore is a wrapper script around FastQC and Cutadapt [99] to enable quality and adapter trimming to be applied consistently to NGS FASTQ files. The integration of the previously mentioned software packages makes it suitable for automated adapter and quality trimming at once. As default, Trim Galore uses the first 13 bps

of Illumina standard adapters ('AGATCGGAAGAGC'). Other adapters are also accepted, such as the Nextera transposase sequence ('CTGTCTCTTATA'), which is more relevant to this work. The tool automatically detects common adapter sequences, which are then trimmed in further steps. Several trimming options are available for quality filtering, including trimming based on the Phred quality score, trimming of short sequences, and specialized trimming to a fixed length. For example, for paired-end sequencing samples, Trim Galore removes entire sequence pairs if one or both reads are shorter than the specified length limit. Furthermore, this tool accepts and can generate standard or gzip compressed FASTQ files. [90]

Besides the main functions, it provides additional features for Reduced Representation Bisulfite-Seq (RRBS) sequence files, where quality and adapter trimming is performed in two consecutive steps. Thus, 2 additional bases can be removed that contain a cytosine that was artificially introduced in the end-repair step during library preparation. [90]

In summary, for almost all sequencing applications, Trim Galore rises to the challenge of taking appropriate action to correct problems identified by thorough quality control. [90]

2.5.2. Bowtie2

Aligning reads to a reference genome, and thus finding the exact regions in the genome from which the reads originated, is a highly computationally intensive task, making fast and memory efficient tools essential for analysing NGS data [91]. Many aligners use a genome index to narrow down the corresponding regions in the reference genome by looking for all possibilities to mutate the read string into a string that exists in the reference [58]. For instance, an efficient index is the full-text minute index [100]. However, most existing tools, such as BWA, are only efficient for ungapped alignments of short reads, where gaps are defined as locations caused by sequencing errors or insertions and deletions. Bowtie2 extends the full-text minute index approach with a gapped extension stage that uses dynamic programming to achieve an effective combination of sensitivity, accuracy, and speed across a wide range of read lengths and sequencing technologies. Consequently, Bowtie2 is an

ultrafast and memory-efficient software for aligning reads of about 50 to 1000 characters in length, and is particularly good at aligning relatively long genomes. An alignment is provided by four steps. Step one extracts seed substrings from the read and its reverse complement, which are then aligned to the reference genome without gaps using the full-text minute index in the next step. In the third step, the positions of the seed alignments in the reference genome are calculated using the index. Finally, the seeds are extended to complete alignments by dynamic programming. The output is a set of alignments in SAM format. [58]

In conclusion, Bowtie2, with its highly efficient alignment protocol, achieves a very fast and memory efficient gapped alignment of sequencing reads [58].

2.5.3. SAMtools

Different alignment tools produce outputs in different formats, complicating downstream processing. Therefore, the SAM format has been introduced, which is a common alignment format that provides support for all sequence types and aligners, consisting of a header section marked with '@' and an alignment section. In many cases, further processing of the alignment files is necessary to ensure that the data is suitable for subsequent analysis. SAMtools is a library and software package for this purpose, with various functions such as converting from other alignment formats, sorting and merging alignments, removing PCR duplicates, and many more. A small selection of useful commands used in this work is described in the following. [101]

SAMtools view

This command is used for format conversion, file filtering, and extraction of sequence ranges of SAM/BAM/CRAM files [102]. In this study it is mainly required to convert the SAM format (Bowtie2 output) to the BAM format [102], which is the binary representation of SAM, thus increasing the performance [101].

SAMtools sort

Simply put, this command sorts SAM/BAM/CRAM files. However, there are different sorting methods, e.g. sorting alignments by leftmost coordinates or by read name, depending on the subsequent task. In this work this command is used twice. First, the BAM file is sorted by read name to be able to use SAMtools fixmate. Later, the alignment is sorted by leftmost coordinates to prepare it for duplicate removal with SAMtools markdup. [103]

SAMtools fixmate

SAMtools fixmate fills in the mate coordinates, insert sizes, and mate related flags from an alignment that has previously been read name sorted with SAMtools sort. In other words, this tool complements information about paired-end reads to the corresponding other read. The execution is a prerequisite for PCR duplicate removal with SAMtools markdup. [104]

SAMtools markdup

PCR duplicates are defined as copies made by PCR of the same template fragments. Hence, if a pair of reads from paired-end data maps to the exact same location as another pair of reads in the reference genome, it can be assumed that one of them is a PCR duplicate [105]. This can have major implications for subsequent analysis and biological interpretation of the results. Thus, it is necessary to remove these duplicates. SAMtools markdup marks duplicates from a coordinate sorted alignment and offers the possibility to remove them. [106]

In summary, SAMtools provides a variety of tools to further modify alignments to prepare the data for downstream analysis [94].

2.5.4. MACS2

In genome-enrichment assays, the sequencing and read alignment step is usually followed by peak-calling to characterize the significantly enriched genomic regions

[31]. In terms of ATAC-seq, peak-calling identifies the accessibility of chromatin to detect regulatory elements and understand the transcriptional regulation [107]. One and the most popular peak-calling program is MACS2. Although optimized for ChIP-seq assays, it is also suitable for ATAC-seq data and is even the default peak caller in the ENCODE ATAC-seq pipeline [108] [107].

Two key features of MACS2 are the empirical modeling of d, which is the distance between the modes of the Watson and Crick peaks in the alignment, accompanying the tag shift by $d/2$ to putative protein-DNA interaction sites and the use of a dynamic λ_{local} to capture local biases in the genome [25].

By default, MACS2 calculates the number of duplicate reads at a given position, determined by the sequencing depth, and removes redundant reads that exceed this number. Alternatively, there are options to keep these reads. This is useful if duplicates have already been removed in previous steps. MACS2 identifies enriched regions by modeling the distance between paired forward and reverse strand peaks detected by sliding a window across the genome and comparing the number of reads in the window to the expected background. The software uses a fold enrichment cutoff and a set number of enriched regions (1000 by default) to build a peak detection model, expanding reads in the 3' direction based on the fragment length of the model. The genome is then rescanned with a window size of twice the fragment length to select potential peaks. MACS2 also calculates p-values to capture local biases and q-values for peak candidates using the Benjamini-Hochberg correction. [25]

To optimise the use of this tool for the analysis of ATAC-seq data, appropriate input arguments are required. The consensus of the bioinformatics community and the recommendations of the MACS2 developers suggest using the flag '-f BAMPE' for the analysis of paired-end ATAC-seq data to allow MACS2 to accumulate the whole fragments in a general way [109].

2.5.5. DiffBind

DiffBind is an R Bioconductor package that provides functions for e.g. ChIP-seq data or open chromatin assays such as ATAC-seq. It works with aligned sequence

reads, i.e. the output of Bowtie2, and peaks called by a peak caller, for example MACS2. The main goal of this package is to find differentially bound sites between sample groups of a ChIP-seq experiment. This process involves normalising the experimental data and specifying a model design and one or more contrasts. [35] Subsequently, the core analysis routines are executed using statistical routines developed for RNA-seq, primarily edgeR [32] and DESeq2. This results in the assignment of a p-value and FDR adjusted p-values for each binding candidate, giving an indication of the degree of confidence that the sites are differentially bound, presented as a table. However, this tool is also suitable to analyse DARs in ATAC-seq experiments. DiffBind also offers a range of reporting and visualisation features, such as MA-plots and volcano plots. [35]

The results can be further compared to the DEGs from the mRNA-seq analysis, by studying enhancer-gene interactions.

2.6. Enhancer-Gene Interactions and Motif Enrichment Analysis

Understanding the mechanisms involved in regulating gene expression represents a fundamental goal of epigenomics. Enhancers play a central role in this process, as they represent open chromatin regions within the genome that are capable of binding TFs in a sequence-specific manner [110], which, among other functions, affects gene expression [111]. Therefore, the study of enhancer-gene interactions is of paramount importance for a comprehensive understanding of epigenomic landscapes and differential gene expression in disease models. STARE [110] is a useful tool to investigate such interactions.

2.6.1. Evaluate Enhancer-Gene Interactions with STARE

STARE is a tool that quantifies enhancer-gene interactions based on chromatin accessibility data, utilizing a scoring method known as the Activity-by-Contact (ABC)-score [110]. The ABC-score considers enhancer activity measurements and chro-

matin contact data in a gene-centric manner, without accounting for all candidate target genes of an enhancer [112]. STARE offers a generalized ABC-score, which characterizes enhancer activity in a gene-specific manner by incorporating information from all annotated transcription start sites of a given gene, as denoted in Eq. (2.2) [110]:

$$gABC_{r,g} = \frac{\sum\limits_{t \in TSS_g} (A_{r,t} \cdot C_{r,t})}{\sum\limits_{i \in R_g} \sum\limits_{t \in TSS_g} (A_{i,t} \cdot C_{i,t})}, \tag{2.2}$$

where TSS_g are all annotated transcription start sites of gene g, $A_{r,t}$ describes the approximated regulatory activity and $C_{r,t}$ is the relative number of contacts of an enhancer's candidate target genes. By including contact information for a larger number of potentially relevant transcription sites and avoiding the selection of a single transcription start site, more comprehensive information can be obtained with the proposed method. [110]

STARE is capable of assessing the regulatory effect of a TF on a gene, after detecting genomic regions that affect the gene's regulation through ABC-scoring. However, this function is not further relevant in this study and therefore not described in more detail. [110]

The independently executable function 'STARE_ABCpp' requires a bed file containing all candidate regions (*-b*), e.g. the MACS2 narrowPeak file, and a gene annotation file in GTF format (*-a*). In addition, the path to a directory of normalised chromatin contact files in coordinate format (*-f*), a file of gene IDs/symbols to constrain the output (*-u*), which is beneficial for integrating this analysis with DEGs, and a specific column in the bed file indicating the activity of the region (*-n*) can be specified, as depicted in listing A.8 (see appendix A.6). [110]

STARE provides two outputs for each activity column specified in the command (*-n*), one with all the enhancer-gene interactions, and one summarising a set of features for each gene [110]. The first file may be subjected to additional modifications to conduct a motif enrichment analysis.

2.6.2. Motif Enrichment Analysis with HOMER

A major goal of molecular biology is to identify the mechanisms that control gene transcription [113]. Motifs are short recurring patterns or sequences of nucleotides, which in this context are responsible for binding TFs [114]. Motif enrichment analysis aims to identify the TFs that regulate the transcription of a given set of genes by detecting overrepresentation of known DNA-binding motifs in the regulatory regions of these genes [113]. These regions are obtained by mapping the ATAC-seq data to DEGs using STARE. A suitable tool to perform motif enrichment analysis is HOMER [115].

The 'findMotifs.pl' function of HOMER accepts FASTA formatted files to perform differential motif discovery. It is recommended to pass both target and background regions. If no background regions are defined, HOMER will scramble the target input sequences to generate a background. The method used by HOMER for motif enrichment analysis includes the ZOOPS scoring approach, which allows either zero or one occurrence per sequence. In addition, hypergeometric or binomial enrichment calculations are applied to identify overrepresented motifs in the regulatory regions of a number of genes. After the automated background selection, GC normalisation is performed. Sequences in the target and background sets are grouped into bins based on their GC content, typically at 5.0% intervals. To ensure that the background sequences have a similar GC content distribution as the target sequences, the background sequences are weighted accordingly. This strategy is particularly useful to avoid bias in motif discovery towards GC-rich regions, such as CpG islands. Subsequently, autonormalisation is employed to identify and remove imbalances in the sequence content other than GC%. This is achieved by assigning weights to the background sequences with the aim of minimising the difference in the frequency of short oligo sequences between the target and background datasets. The next step is to load a library of known motifs to scan each sequence for motif occurrences. By evaluating how many target sequences are considered bound compared to the background sequences, the final motif enrichment is calculated. The enrichment of known motifs is displayed as a HTML file. HOMER also possesses the functionality to discover motifs de novo. However, this is not further relevant for this study and

therefore not described in more detail. [115]

The function former described can be executed at console level, as shown in listing A.9 (see appendix A.6)

3. Results

mRNA-seq and ATAC-seq samples contain a lot of interesting information about the genome and between different groups (conditions) of genomes, respectively. To obtain this information, various methods and tools can be applied, as mentioned in chapter 2. For example, DESeq2 [33] for mRNA-seq data, which provides DEGs for given count data. This makes it a powerful tool for studying allele-specific variation and cardiovascular diseases such as HFpEF. Similarly, DiffBind [35] is a useful option to obtain DARs between two groups of samples for ATAC-seq data. Analysis of DEGs and DARs can provide previously unknown insights into differences between two genomes or even merely between treated and untreated samples.

In the subsequent analysis, quality control was performed for all samples obtained (BL6, CC, XB, and YB) from the initial mRNA-seq experiment (see section 3.1). For the ATAC-seq experiment, only BL6 and CC samples were evaluated for quality control due to unexpected results from the differential expression analysis (see section 3.2). Exclusively these sample groups BL6 and CC are presented in the results of the mRNA-seq analysis. The examination of the BL6 and CC-ATAC-seq samples uncover further issues and unexpected results (see section 3.3), leading to the preparation of new BL6 and CC-ATAC-seq samples without high-fat diet treatment. The new samples were then considered for enhancer-gene interaction analysis (see section 3.4). A detailed discussion surrounding the impact of these observations on the results and the potential root cause are presented in section 4.1.

The entirety of the results, including those that are not presented in this thesis, have been archived along with the corresponding code and can be accessed in a GitHub repository (`https://github.com/jurummel/Masterthesis_JRummel`).

3.1. Quality Control

As mentioned in section 2.1, the quality of the obtained data can be negatively affected by various artifacts during the sequencing process [49], which may have implications for subsequent analysis and biological conclusions. Therefore, raising awareness about the data quality and possibly correcting it is a mandatory prerequisite in biological data analysis.

3.1.1. Adapter Content

Adapter content provides information about whether trimming with Trim Galore [90] is necessary. Figure 3.1 and fig. 3.2 exhibit the adapter content from all BL6, CC, XB, and YB samples for the mRNA-seq experiment, while only BL6 and CC samples for the ATAC-seq experiment are presented.

mRNA-seq

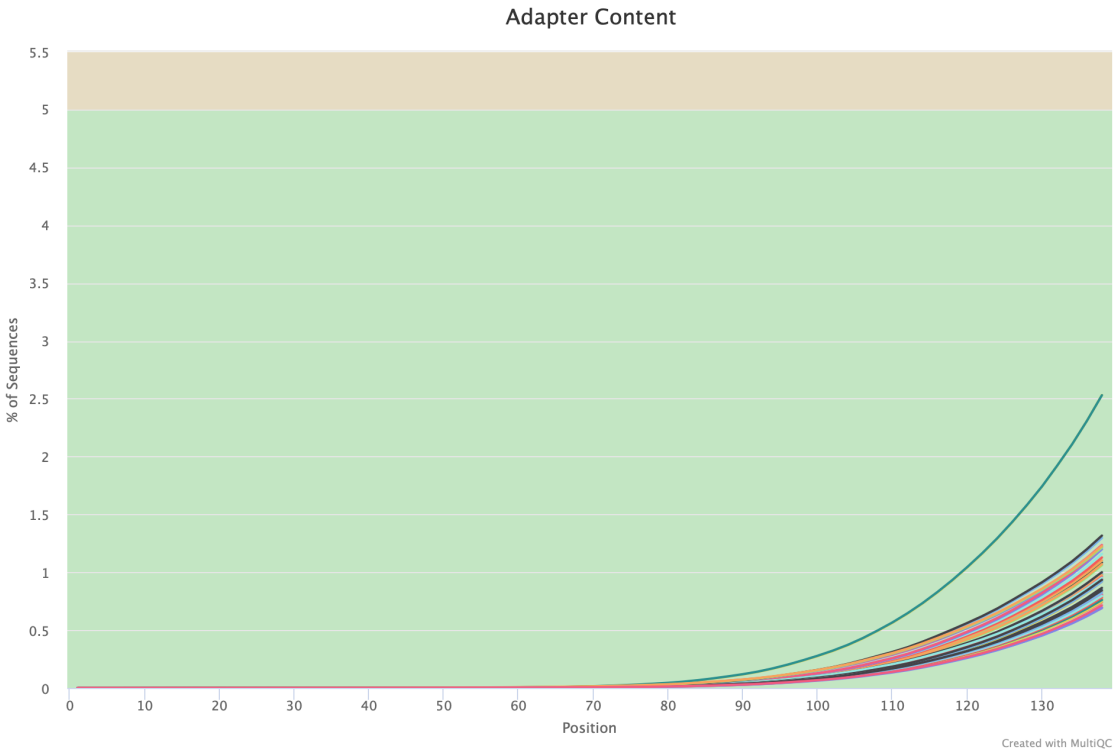


Figure 3.1.: Adapter Content mRNA-seq: This plot shows the adapter content of all samples of the mRNA-seq experiment. It shows the cumulative percentage count of the proportion of the library that has seen each of the adapter sequences at each position [50]. Each color represents one specific sample. For instance, sample BL611, comprising BL611_1 and BL611_2, which exhibit almost identical adapter content and are hence described as a single sample, namely BL611. From base 1 to 66 there is 0.0% of adapter content. It slowly raises up to 1.08% at base 138. The green colored area, describes adapter content smaller than 5.0% and therefore no warning is displayed. Above this threshold, yellow colored (here 5.0% to 5.5%), a warning gets displayed. (This plot is generated by FastQC [50].)

Figure 3.1 indicates little to none adapter contamination for every sample. Thus, no adapter sequences were trimmed before starting further analysis.

ATAC-seq

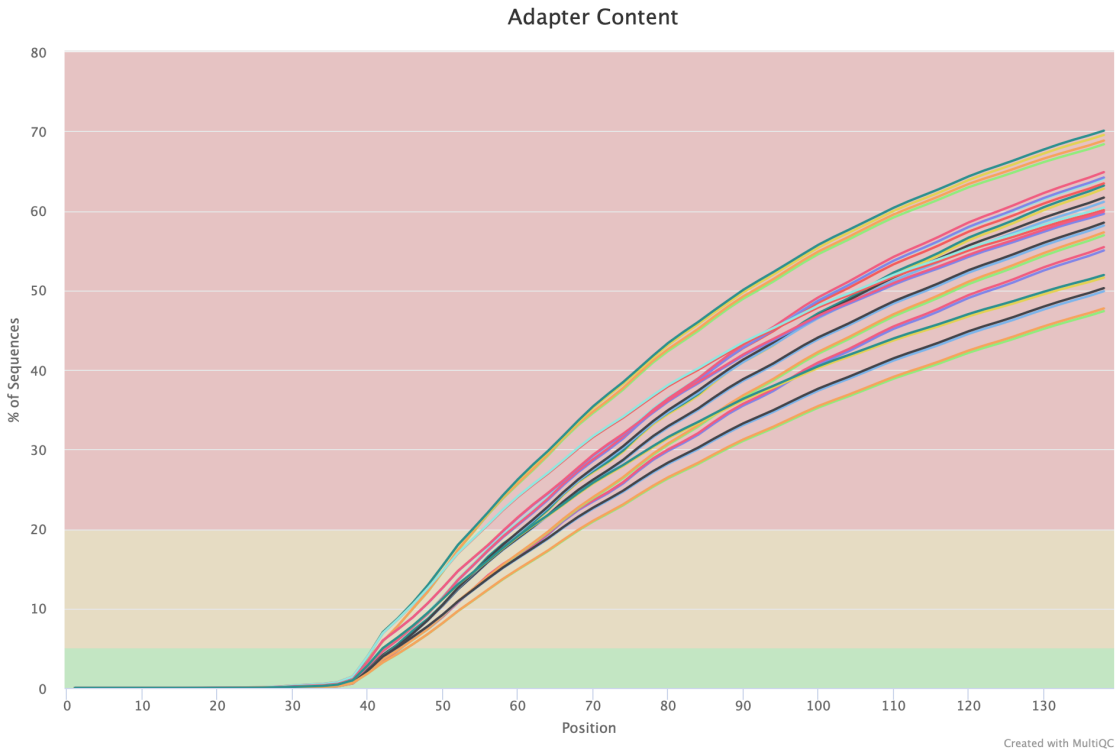


Figure 3.2.: Adapter Content ATAC-seq: This figure shows the adapter content for all samples of the ATAC-seq experiment. It shows the cumulative percentage count of the proportion of the library that has seen each of the adapter sequences at each position [50]. Each colored line represents one specific sample. For example, in sample BL611, from base 1 to 20 there is 0.00% of adapter content. The curve rises slowly to 0.74% to base 38. After this position the adapter content sharply increases to 58.5% at base 138. The green colored area describes adapter content smaller than 5.0% and therefore no warning is displayed. Above this threshold, yellow colored (here 5.0% to 20.0%), a warning is displayed. Adapter content higher than 20.0% results in an error output. BL611_1 and BL611_2 are almost identical in terms of adapter content and are therefore described together as BL611.(This plot is generated by FastQC [50].)

Figure 3.2 shows that there is a high amount of adapter contamination in all ATAC-seq samples, because all of them exceed a critical adapter content of 20.0%. After trimming the Nextera adapter sequence ('CTGTCTCTTATA') with Trim Galore, no samples were found with any adapter contamination higher than 0.1% and therefore no plot is generated [50]. In total, about 24.0% to 36.0% base pairs got trimmed.

3.1.2. Sequence Duplication Level

Another feature that should be considered when checking for errors and assessing data quality is the sequence duplication level. This module contains the degree of duplication for each sequence in a library. Sequence duplication diagrams generated by FastQC [50] show the percentage of the entire library for each duplication level presented up to 10,000. Duplicates are either technical duplicates created by PCR artefacts, or biological duplicates that are natural collisions where different copies of the exact same sequence are randomly selected [50]. The present figures (fig. 3.3 and fig. 3.4) visualize the sequence duplication from all BL6, CC, XB, and YB samples for the mRNA-seq experiment, while only BL6 and CC samples for the ATAC-seq experiment are presented.

mRNA-seq

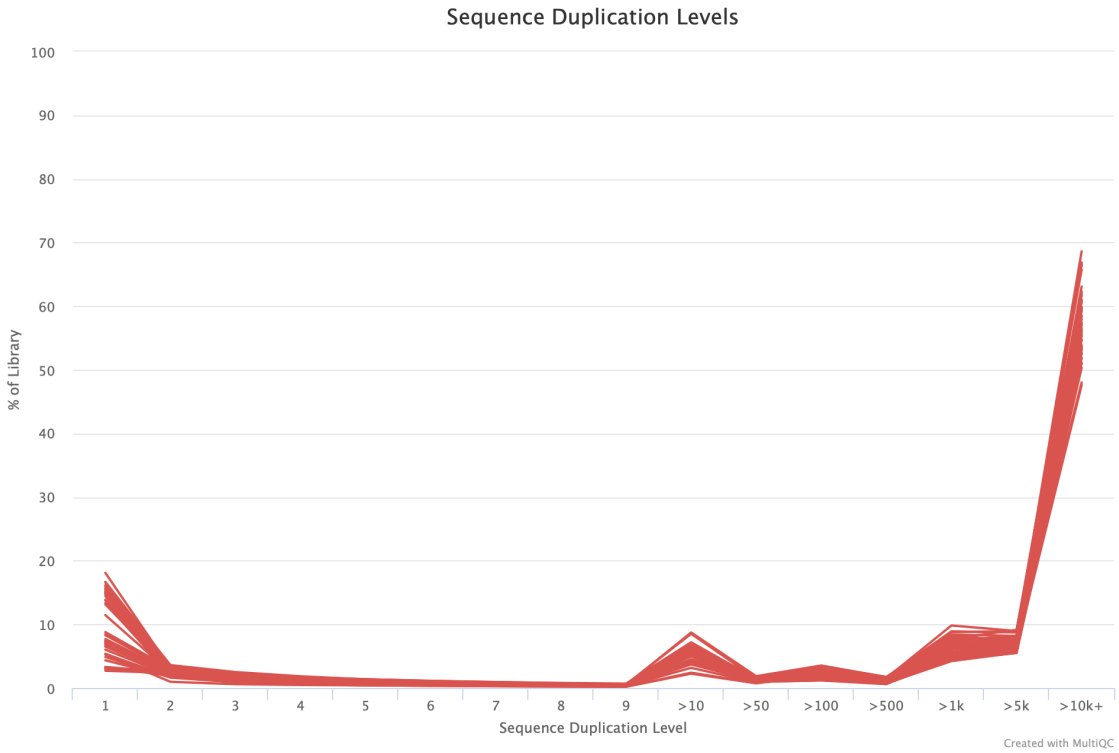


Figure 3.3.: Sequence Duplication Level mRNA-seq: This figure shows the sequence duplication level of all mRNA-seq samples. The X-axis shows the sequence duplication level and the Y-axis the percentage of the library. For example, 5.3% of the library of BL611_1 and 13.3% of BL611_2 are unique. Furthermore, there is a small peak at >10, which means that about 8.5% of the reads are duplicated between 10 and 50 times. There is also a notable peak at the end (>10k+), which implies that 60.9% of BL611_1 and 55.5% of BL611_2 are duplicated more than 10,000 times. (This plot is generated by FastQC [50].)

Figure 3.3 displays a very high level of duplication, since 47.6% (F2019XB_2) to 68.6% (BL615_2) of all reads are duplicated more than 10,000 times.

ATAC-seq

It is important to note that the following plots show the quality measurements for the trimmed ATAC-seq samples.

As illustrated in Figure 3.4, all ATAC-seq samples have similar behaviour and possess a rather high level of duplicates for a ATAC-seq experiment, as 10.0% to

25.0% of all reads are duplicated between 1,000 and 5,000 times. Therefore, to avoid skewing the results, these PCR duplicates are removed in the upcoming analysis.

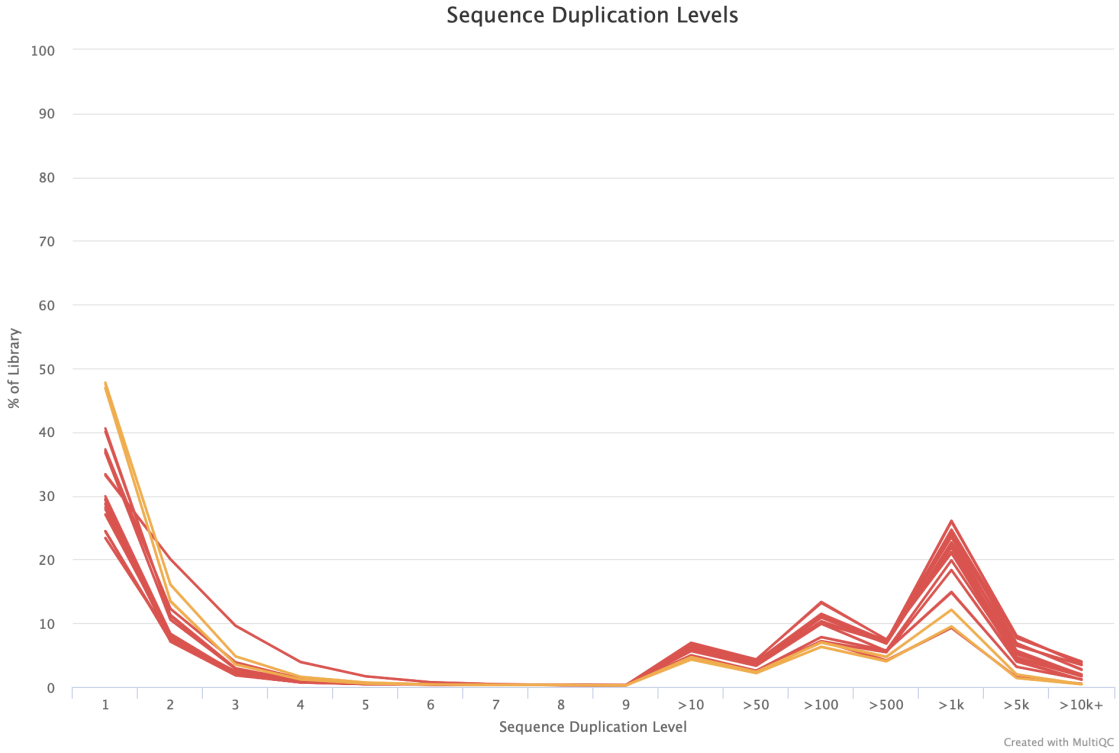


Figure 3.4.: Sequence Duplication Level ATAC-seq: This plot shows the sequence duplication level of all ATAC-seq samples. The X-axis shows the sequence duplication level and the Y-axis the percentage of the library. Red lines are samples that are indicated as errors by FastQC, while the orange lines only represent a warning. As an example, for BL611 this plot indicates that 27.1% of the library is unique. The line slowly starts rising at a duplication level of nine until resulting in a peak at >1k, which means that 24.2% of BL611 is duplicated between 1,000 and 5,000 times. BL611_1 and BL611_2 are almost identical in terms of sequence duplication level and are therefore described together as BL611. (This plot is generated by FastQC [50].)

3.1.3. Overrepresented Sequences

Overrepresented sequences are individual sequences that make up a proportionally large part of the total set. Sequences that account for more than 0.1% are classified as overrepresented. This means either that the sequence has a high biological significance, that the library is contaminated, or not as diverse as expected [50]. All BL6, CC, XB, and YB samples are shown in the mRNA-seq experiment. For

clarity and to avoid redundancy, in fig. 3.5 only Read '_1' is included as a label on the Y-axis, while Read '_2' is shown directly below. (Important note: There are no overrepresented sequences in the ATAC-seq samples.)

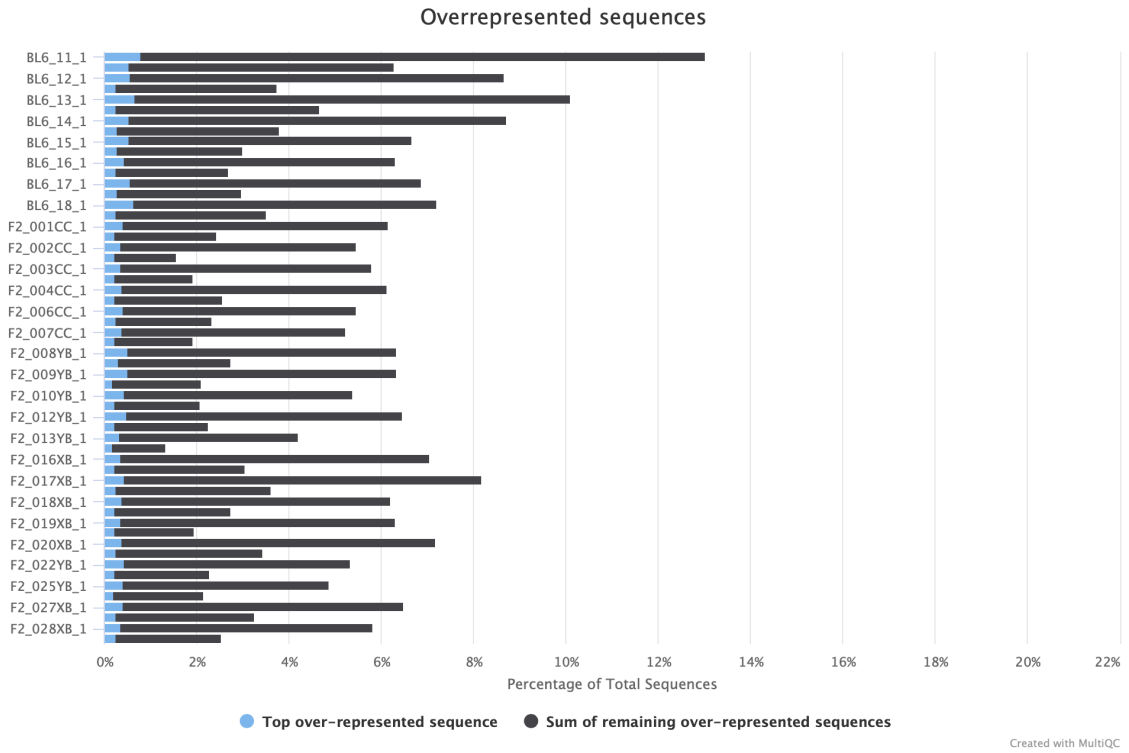


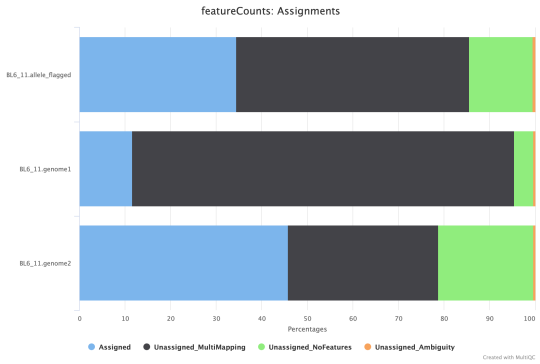
Figure 3.5.: Overrepresented Sequences mRNA-seq: This plot shows the total amount of overrepresented sequences found in each library for all mRNA-seq samples. The overrepresented reads that come from the single most common sequence are displayed in blue and the total count from all remaining overrepresented sequences which make up more than 0.1% of the total are shown in grey. For instance, in BL611_1 0.78% of all sequences are represented by one particular sequence. 12.23% of the whole sequence is represented by the remaining overrepresented sequences. In BL611_2 one particular sequence represents 0.55% of all sequences. However, just 5.73% of the whole sequence is represented by the remaining overrepresented sequences. (This plot is generated by FastQC [50].)

Overrepresented sequences go along with the sequence duplication level and can be further compared to the adapter content, as many adapter sequences are very similar to each other and can therefore be reported which is not technically correct, but have a very similar sequence to the actual match [50].

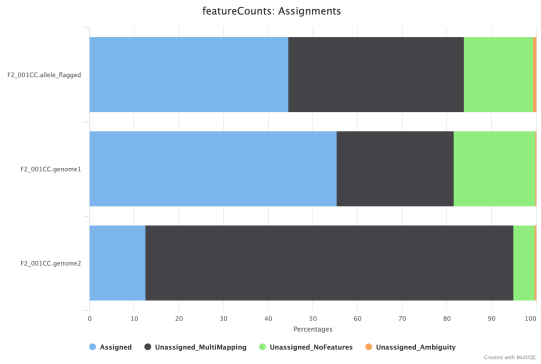
3.1.4. featureCounts: Assignments (in %)

featureCounts accurately assigns reads by comparing the mapping position of each read or fragment base to the genomic region spanned by each feature. Thus, each read can be specifically assigned to a gene. [57]

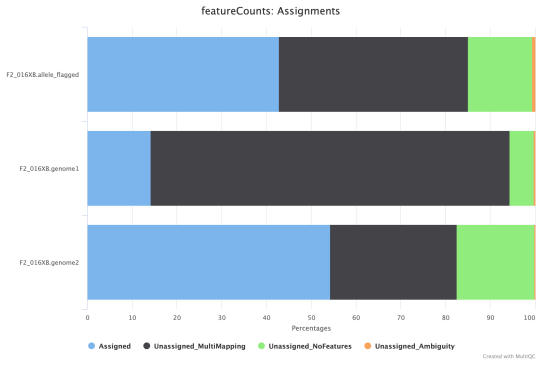
Because of the large differences between the various strains (BL6, CC, XB, and YB), one representative of each is shown in fig. 3.6. All other samples of each strain behave very similarly to BL611, F2001CC, F2016XB, and F2008YB and can hence also be described by these results. (Important note: This module is only evaluable for the mRNA-seq samples.)



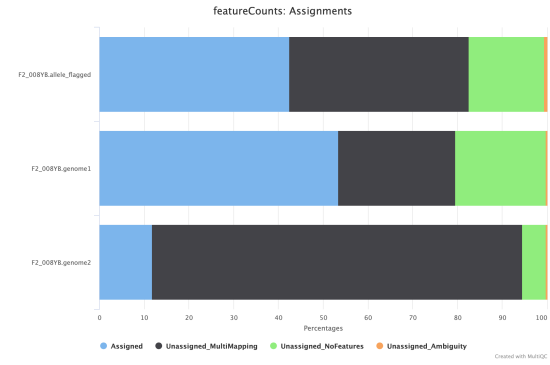
(a) featureCounts: Assignments for BL611 mRNA-seq



(b) featureCounts: Assignments for F2001CC mRNA-seq



(c) featureCounts: Assignments for F2016XB mRNA-seq



(d) featureCounts: Assignments for F2008YB mRNA-seq

Figure 3.6.: featureCounts: Assignments for All Strains mRNA-seq: This plot illustrates the assigned reads for the allele-flagged BAM file and for both allele-specific BAM files (BL6 and CC). The blue bar describes all assigned reads. Unassigned reads are displayed as follows: Multimapping in grey, no features in green, and ambiguity in orange. (This plot is generated by MultiQC [56].)

As shown in fig. 3.6a, for BL611 in the allele flagged file, 34.5% of all reads are assigned, 51.0% are unassigned due to multimapping, 14.0% are unassigned due to lack of function, and 0.5% are unassigned due to ambiguity. Across genome 1, 11.6% of all CC-specific reads are assigned, 83.8% are unassigned due to multimapping, 4.2% are unassigned due to lack of function, and 0.3% are unassigned due to ambiguity. In genome 2, 45.8% of all BL6-specific reads are assigned, 32.9% are unassigned due to multimapping, 20.9% are unassigned due to lack of function, and 0.4% are unassigned due to ambiguity. To summarize, the percentage of assigned reads in BL6 is much higher than in CC. In addition, the number of total reads labeled BL6-specific is approximately five times higher than the number of reads specific for CC. This is not surprising as a homozygous BL6 sample is analysed (see section 4.1).

For F2001CC, as depicted in fig. 3.6b, in the allele-flagged file, 44.6% of all reads are assigned, 39.3% are unassigned as a result of multimapping, 15.5% are unassigned due to lack of function, and 0.7% are unassigned due to ambiguity. In genome 1, 55.4% of all CC-specific reads are assigned, 26.2% are unassigned due to multimapping, 18.3% are unassigned due to lack of function, and 0.2% are unassigned due to ambiguity. Furthermore, in genome 2, 12.6% of all BL6-specific reads are assigned, 82.3% are unassigned due to multimapping, 4.9% are unassigned due to lack of function, and 0.2% are unassigned due to ambiguity. To conclude, the assigned read percentage in CC is much higher than in BL6. Furthermore, the number of total reads that are flagged as specific for CC is about four times higher than the amount of reads identified as specific for BL6. Nevertheless, this is reasonable since a homozygous CC sample is analysed (see section 4.1).

As depicted in fig. 3.6c, for F2016XB in the allele-flagged file, 42.8% of all reads are assigned, while 42.3% are unassigned due to multimapping, 14.4% are unassigned due to having no function, and only 0.6% are unassigned due to ambiguity. In genome 1, 14.2% of all CC-specific reads are assigned, while 80.1% are unassigned due to multimapping, 5.4% are unassigned due to lack of function, and 0.3% are unassigned due to ambiguity. However, in genome 2, 54.2% of all BL6-specific reads are assigned, 28.3% are unassigned due to multimapping, 17.3% are unassigned due

to no function, and 0.2% are unassigned due to ambiguity. As a result, there was a significantly higher percentage of assigned reads in BL6 compared to CC, which was unexpected. The ratio of genome-specific assigned reads is expected to be more similar for heterozygous samples, as further discussed in section 4.1.

According to fig. 3.6d, of the total reads for F2008YB in the allele-flagged file, 43.8% are successfully assigned, while 38.8% are unassigned due to issues with multimapping. Additionally, 16.7% are unassigned due to a lack of function and 0.6% are unassigned due to ambiguity. In genome 1, 53.7% of all reads specific to CC are assigned, while the rest are unassigned, with 26.5% due to multimapping, 19.6% due to a lack of function, and 0.2% due to ambiguity. In contrast, in genome 2, only 15.3% of all BL6-specific reads are assigned, with 77.8% being unassigned due to multimapping, 6.7% unassigned due to a lack of function, and 0.2% unassigned due to ambiguity. This results in an unexpected outcome, where CC has a much higher percentage of assigned reads compared to BL6. It is expected that the ratio of genome-specific assigned reads will be more comparable for heterozygous samples, as further discussed in section 4.1.

3.1.5. Overall Alignment Rate (in %)

The overall alignment rate provides insight into the percentage or the total number of reads that are mapped to the reference genome by an aligner such as Bowtie2 [58] for the ATAC-seq data or STAR [60] for the mRNA-seq data. This in turn provides insight into the quality of the alignment and the biological relevance of the result. Figure 3.7 and fig. 3.8 illustrate the overall alignment rate from all BL6, CC, XB, and YB samples for the mRNA-seq experiment, while only BL6 and CC samples are shown for the ATAC-seq experiment.

mRNA-seq

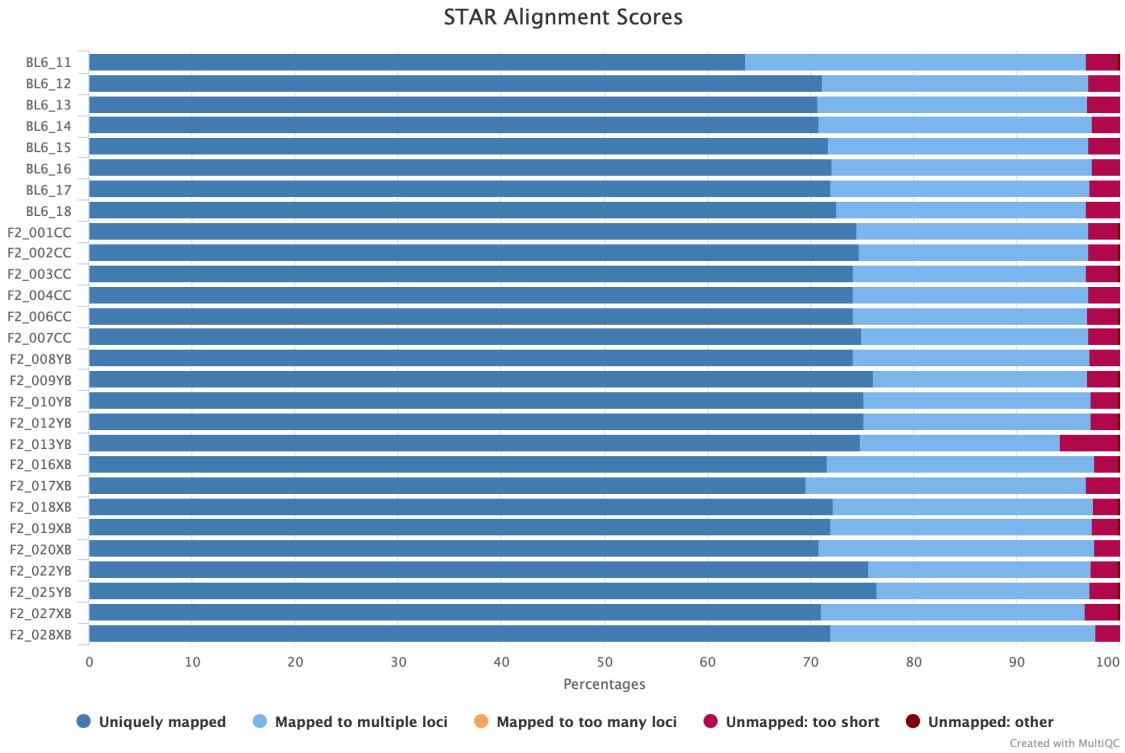


Figure 3.7.: Overall Alignment Rate mRNA-seq: This plot shows the overall alignment rate of STAR for all mRNA-seq samples. All uniquely mapped reads are depicted in dark blue, while the reads that are mapped to multiple loci are displayed in light blue. Unmapped reads that are too short are shown in red and others in dark red. For example, in sample BL611, 63.7% of all reads are mapped uniquely, 33.0% are mapped to multiple loci, and 3.2% are not mapped. (This plot is generated by MultiQC [56].)

Figure 3.7 clearly indicates a similar alignment rate for all mRNA-seq samples, ranging from 63.8% (BL611) to 76.1% (F2009YB).

ATAC-seq

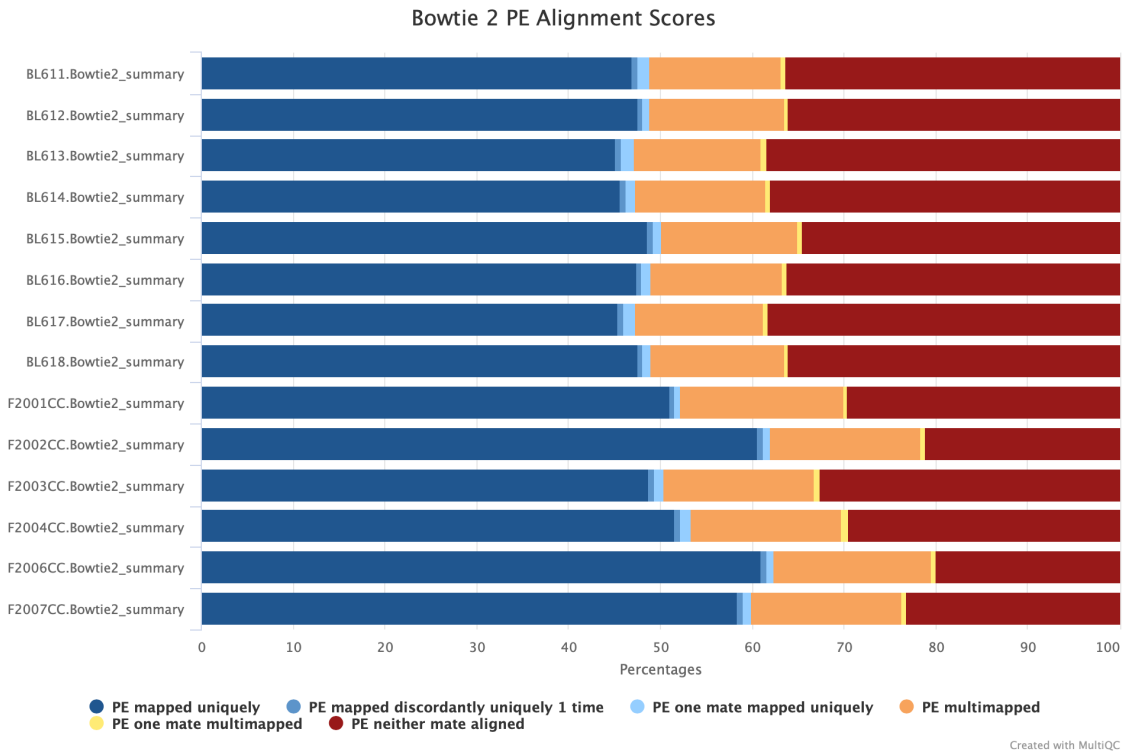


Figure 3.8.: Overall Alignment Rate ATAC-seq: This plot shows the overall alignment rate of Bowtie2 for all ATAC-seq samples. The blue lines denote all paired-end reads that are uniquely mapped, either 1 time discordant, unique to just one mate, or completely unique. Multimapped reads, either completely or just one mate, are depicted by orange and yellow colors. The red color describes all reads that are not aligned. For example, in sample BL611, 48.9% of all reads are mapped uniquely, 14.9% are mapped multiple times, and 36.1% are not mapped. (This plot is generated by MultiQC [56].)

For the ATAC-seq experiment, the alignment rate is rather similar across all samples in a strain. The alignment rate ranges from 61.6% (BL613) to 80.0% (F2006CC), with CC achieving a higher alignment rate on average than BL6.

3.2. mRNA-seq Analysis

3.2.1. Differential Expression Analysis

It is interesting to consider genetic differences and similarities between control, with a standard diet, and treated samples, with a high-fat diet and ROS inhibition by

L-NAME inhibitor. This analysis reveals DEGs as a consequence of the response of different mouse strains to the high-fat diet [41].

featureCounts Data

The snakePipes [21] mRNA-seq pipeline output (see section 2.2.2) relevant for DEG analysis is a featureCounts file provided in TSV format. This file contains gene-level counts for three categories, as allelic mapping mode is used. The first represents the gene counts of the allele-tagged BAM file, which adds tags to the original SAM file describing whether or not a read can be mapped to a specific allele [116] based on the given SNP information. The other two columns contain the gene counts of the allele-specific BAM files, which are referred to as genome 1 (CC) and genome 2 (BL6).

For the analysis, the featureCounts data, obtained through the mRNA-seq pipeline of snakePipes, is divided into 4 groups: Control and treated samples of each strain. In addition, the values of the individual allele assignments, i.e. genome 1 (CC), genome 2 (BL6), and the allele flagged BAM file, are summed to one value. Furthermore, a metadata file is prepared containing information about which samples or columns in the featureCounts table belong to a control or treated group. This allowed DESeq2 to analyse the count data, applying a FDR correction of 0.05. Consequently, the key information are DEGs, enabling a comparison between the two groups and strains. The analysis yields approximately 15 times more DEGs for BL6 than for CC, as illustrated in fig. 3.9. With merely one DEGs in XB and twelve in YB, these samples are considered insignificant and are therefore excluded from further analysis.

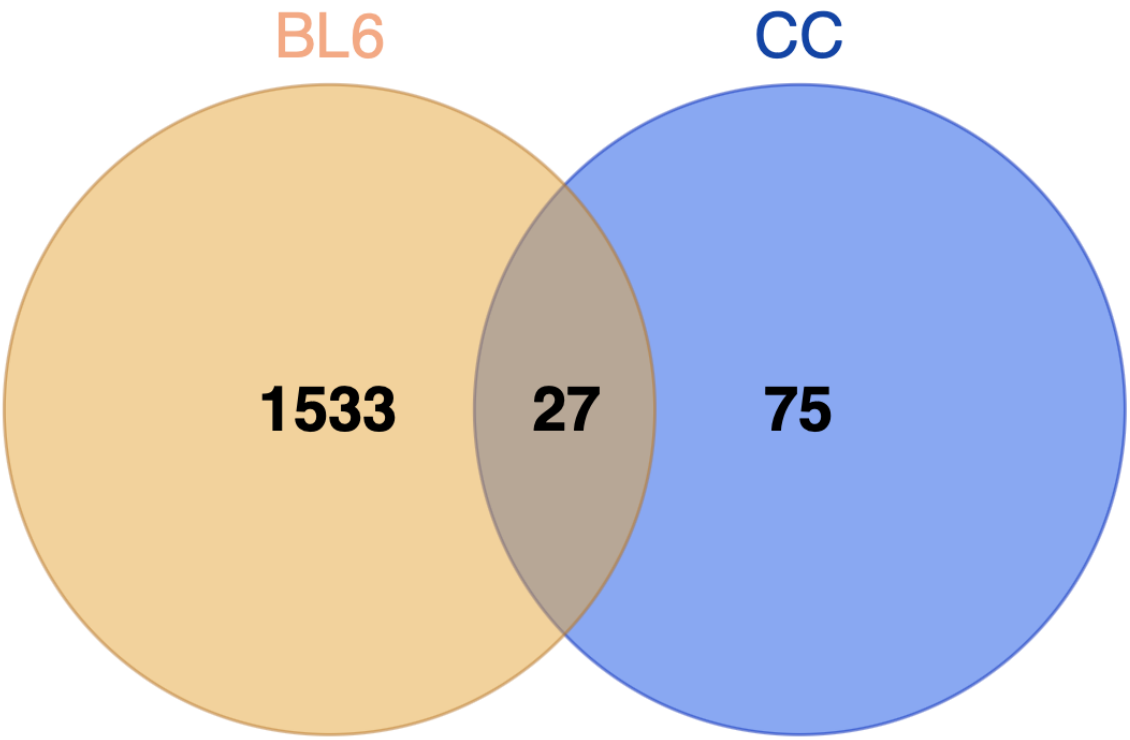
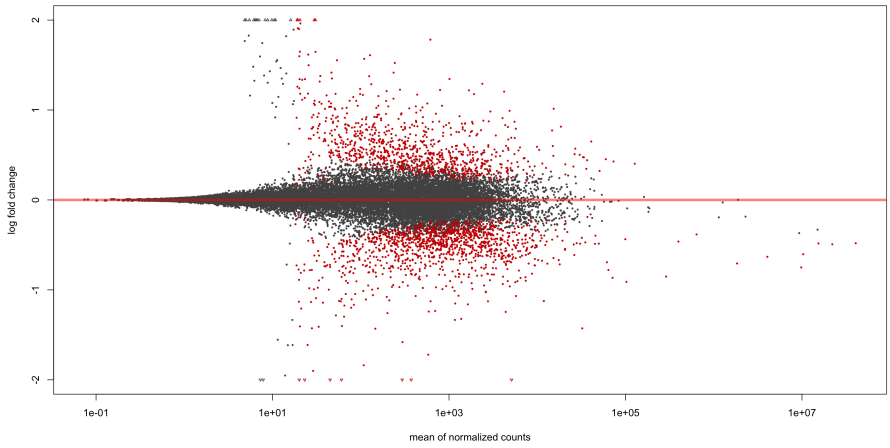
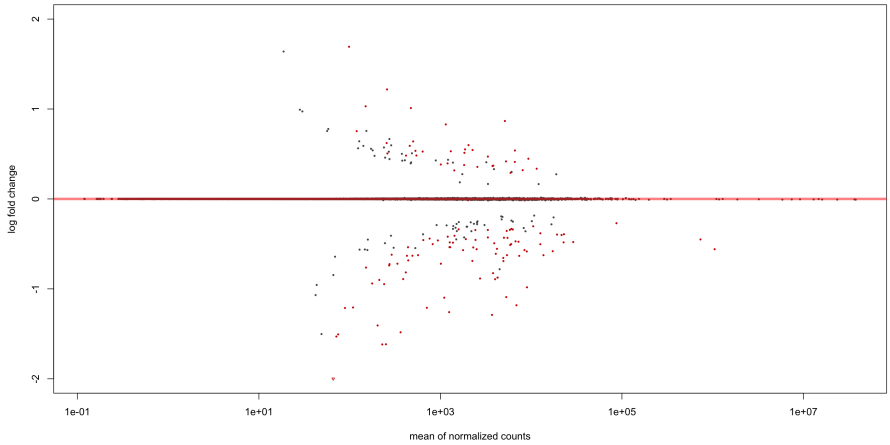


Figure 3.9.: Venn Diagram of Differentially Expressed Genes Treated vs. Untreated: The blue circle represent the number of DEGs in CC and the orange circle represents BL6. BL6 has 1560 DEGs and CC only 102. There are 27 overlapping DEGs between the two strains. (Constructed with draw.io [96].)

The different number of DEGs data discovered is supported by MA-plots displaying similarity between genomes within a strain with the LFC. The LFC is calculated simply as the ratio of the difference between the final value (treated) and the initial value (control) over the original value (control). In this study, a positive LFC value indicates an increase of gene expression, while a negative LFC indicates a decrease in gene expression. [33] In fig. 3.10 it is clearly visible that the investigated genes do not show significant differences for CC, but a greater difference for BL6. LFC-shrinkage was applied since it is useful for gene visualization and ranking [33].



(a) MA-plot BL6



(b) MA-plot CC

Figure 3.10.: MA-plot Differentially Expressed Genes BL6 and CC: This figure shows the MA-plots of DEGs for BL6 and CC. The LFC is plotted on the Y-axis and the mean of normalized expression counts is plotted on the X-axis [117]. Very similar genes approach 0. Similar genes are shown as black dots. Genes that are up- or downregulated are displayed as red dots. It is clearly visible that more genes are differentially expressed in BL6 than in CC, as many genes cluster around zero. LFC-shrinkage was applied. (Constructed with the R package DESeq2 [33].)

BL6 exhibits considerably more up- or downregulated genes than CC. The majority of genes in CC exhibit a very similar expression in the control and treated group as they all sit on the null line, as shown in fig. 3.10b. On the other hand, gene expression appears to be altered between the untreated and treated groups of

BL6, as visualized in fig. 3.10a. These results are consistent with the phenotype of the two strains and are discussed further in section 4.1. Clustering of the DEGs obtained according to the change in expression could shed light on the biological processes underlying the differences between CC and BL6.

3.2.2. Clustering of Differentially Expressed Genes and Functional Enrichment Analysis

Differences Between Cluster - Hierarchical Clustering

As can be seen from fig. 3.9, there are a large number of DEGs, all of which differ in their degree of expression. Classifying these genes into functional groups requires clustering. Since all genes contain raw count data for each sample, normalisation is an essential determinant of the validity and reliability of interpretations [118]. This is very useful, to ensure that all values are on a comparable scale. Thus, what matters is the relative change in expression for each gene. Suitable for this data transformation is the z-score, which can be interpreted as the number of standard deviations away from the mean. [119] The z-score is calculated as denoted in Eq. (3.1) [120]:

$$z = \frac{(x - \mu)}{\sigma} \tag{3.1}$$

In terms of gene expression, x is the raw gene expression of the considered gene, μ describes the overall average gene abundance, and σ represents the standard deviation of all measured values across all samples. Therefore, a z-score above 0 states that the expression of the gene is above the average of all samples, while a value less than 0 means that it is below average. Obviously, a value of zero stands for an exactly average gene expression [119].

The z-score is determined for all genes labeled as differentially expressed in either BL6 or CC for all samples. To obtain a clear overall pattern across the control and treated samples, the mean is calculated across all samples in a group, e.g. BL611, BL612, BL13, and BL614 for the treated group of BL6.

The results are displayed as a heatmap in fig. 3.11 created with the R package ComplexHeatmap [121].

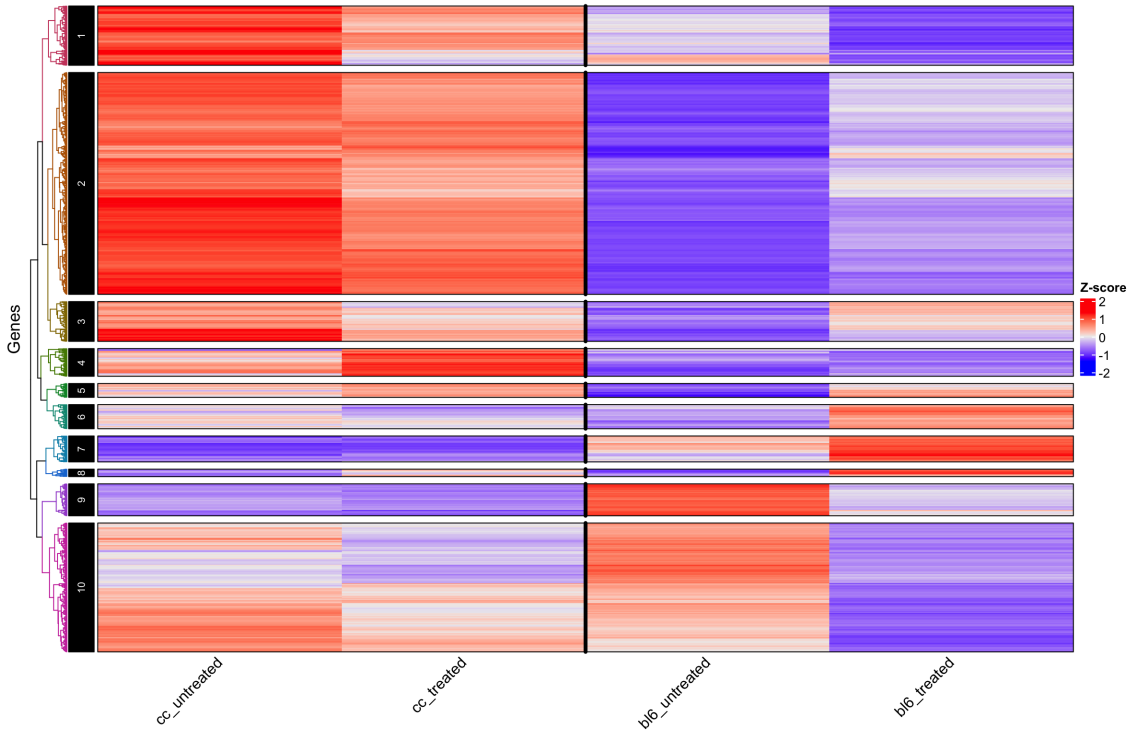


Figure 3.11.: Heatmap Cluster of Differentially Expressed Genes: The heatmap displays the gene expression of all DEGs determined by DESeq2 for the strains BL6 and CC subdivided into treated and untreated. The gene expression is normalized using the z-score, which is calculated for each sample. For every group the mean of all z-scores represent the gene expression for each DEG. The colour-scale is based on the CIELAB color space [122]. Blue shades describe a low gene expression level, while red shades reflect a high gene expression level. It is clearly visible that the clusters are different in size and gene expression of the groups. For example, cluster 9 shows significantly higher expression in the untreated group of BL6 than in the treated group of BL6. In contrast, there are no significant differences between the two groups in CC. Otherwise, in cluster 4, an increase in gene expression in the treated group of CC can be examined, while BL6 shows no differences between the two conditions. (Constructed with the R package ComplexHeatmap [121].)

To emphasize the differences in gene expression between treated and untreated groups in CC and BL6 according to hierarchical clustering, a line plot for each cluster is displayed in fig. 3.12, highlighting the information presented in fig. 3.11.

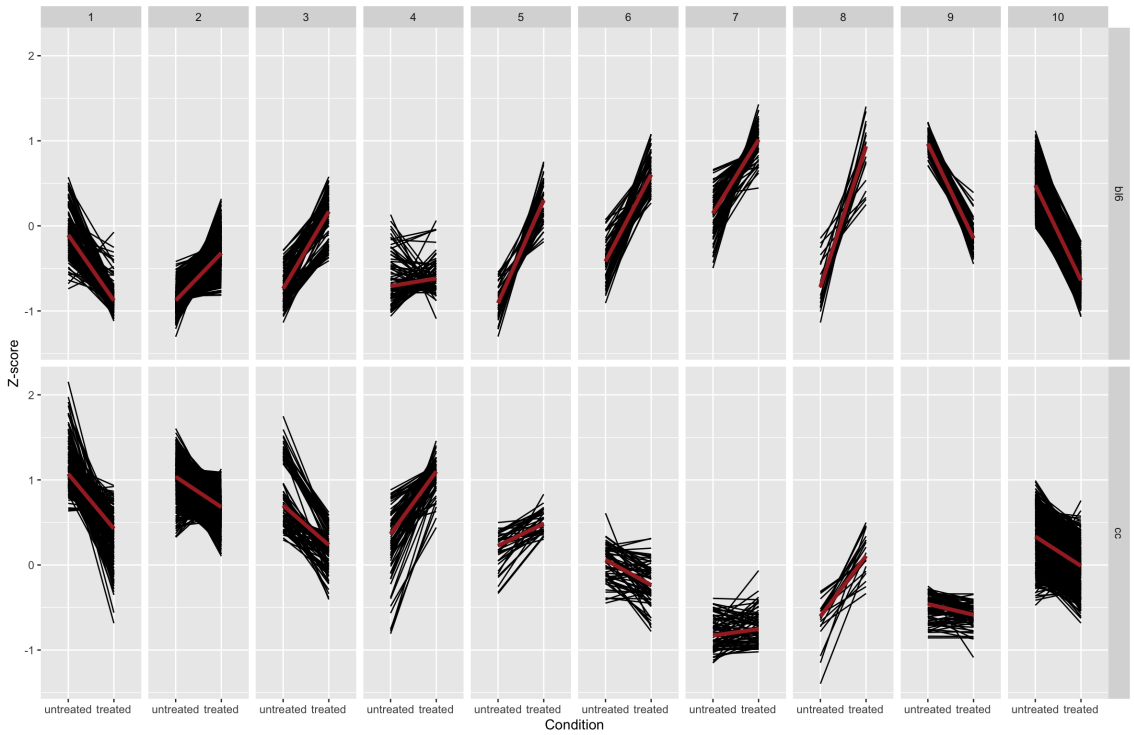


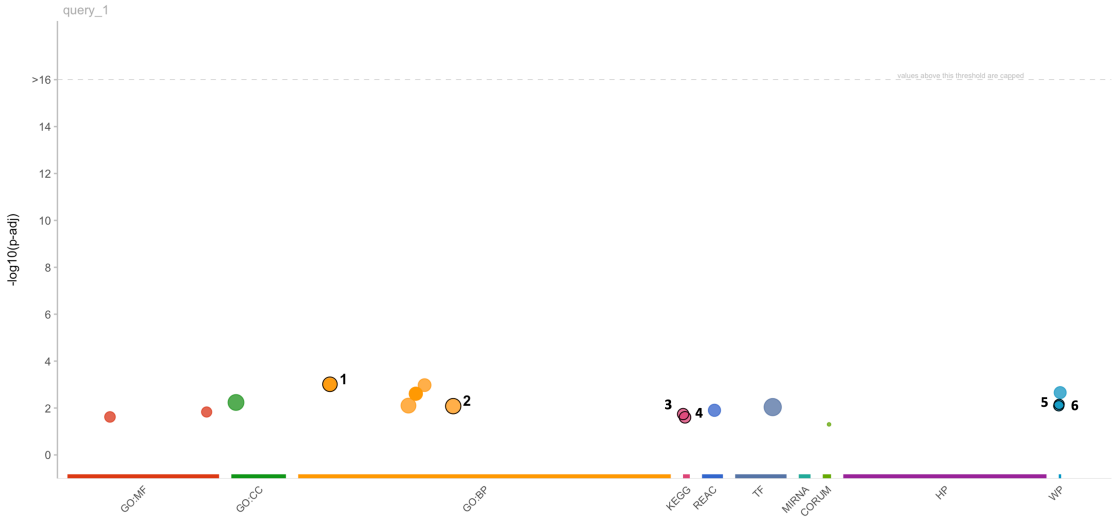
Figure 3.12.: Linegraph for Each Cluster Treated vs. Untreated: These linegraphs display the gene expression of all DEGs determined by DESeq2 for the strains BL6 and CC subdivided into treated and untreated for every cluster. The gene expression is normalized by the z-score, which was calculated for each sample. For every group the mean of all z-scores represent the gene expression for each DEG. Each gene is represented as a black line, showing the course of expression from untreated to treated. The red line represents the trend of all the genes in the particular cluster. For example, cluster 9 shows a decrease in expression from the untreated group of to the treated group of BL6. In comparison, there is a fairly flat trend line in CC. Otherwise, an increase in gene expression from the untreated to the treated group in CC can be examined within cluster 4, while BL6 is represented by a rather flat line between the two conditions. (Constructed with the R package ggplot [123].)

These clusters can now be classified into simplified functional categories accomplished by functional enrichment analysis using gProfiler2 [76].

Functional Enrichment Analysis

The classification of gene clusters and therefore the identification of significantly enriched biological functions and pathways from established data sources is exemplified here. The gene lists of all ten different clusters are analysed with the function 'gost'

of gProfiler2 with respect to the organism *mus musculus*. The identified terms, e.g. from the GO or KEGG database, are then visualized in the form of a Manhattan plot (see fig. 3.13 and fig. 3.14). Terms of interest and importance to this work are highlighted. For clarity, only two example clusters with high contrast are presented. These include cluster 8 (see fig. 3.13), showing higher expression in the treated sample group of BL6, and cluster 4 (see fig. 3.14), demonstrating higher mean expression in the treated sample group of CC.



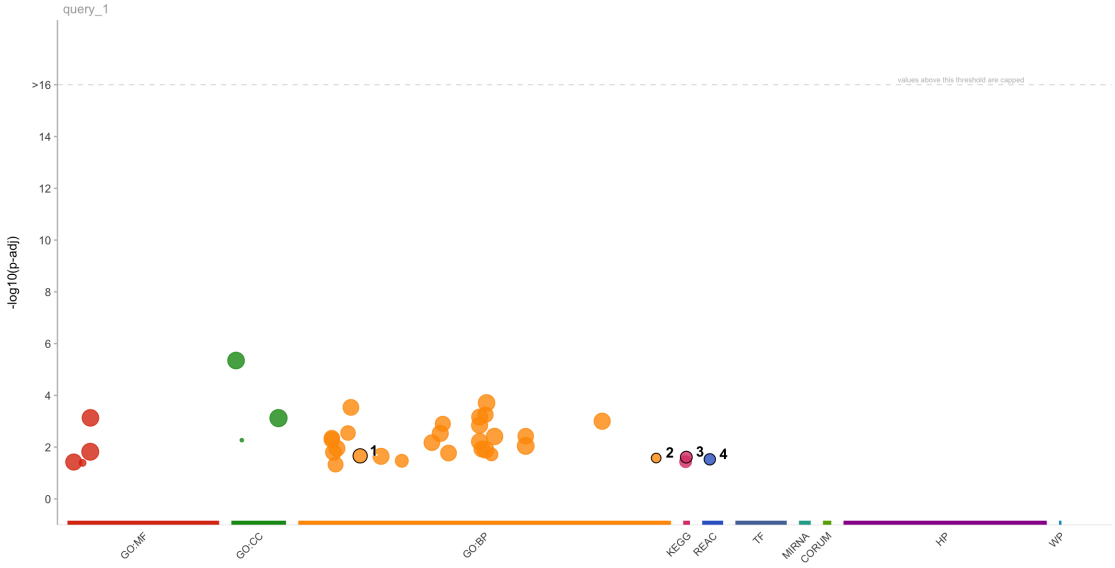
id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0006631	fatty acid metabolic process	456	9.7e-04
2	GO:BP	GO:0044255	cellular lipid metabolic process	1040	8.3e-03
3	KEGG	KEGG:00062	Fatty acid elongation	29	1.8e-02
4	KEGG	KEGG:01040	Biosynthesis of unsaturated fatty acids	34	2.5e-02
5	WP	WP:WP4351	Omega-9 fatty acid synthesis	14	7.0e-03
6	WP	WP:WP4350	Omega-3 / omega-6 fatty acid synthesis	15	8.0e-03

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 3.13.: Highlighted Terms of Functional Enrichment Analysis for Cluster 8: This Manhattan diagram shows each significant term for cluster 8 (see fig. 3.11) from the different data sources as a circle on the X-axis, plotted as the adjusted p-values on the $-log_{10}$ scale on the Y-axis. Each color represents a different data source. There are 6 relevant and therefore highlighted terms from 3 different categories, including biological processes from the GO database, and pathways from the KEGG and WikiPathways database. The highlighted terms with their corresponding adjusted p-values, source, term ID, term name, and term size are listed in the table below the Manhattan diagram. (Constructed with the R package gProfiler2 [76].)

For the genes in cluster 8, 16 significantly enriched processes and pathways are identified. Particularly interesting are two terms in the GO database that are cat-

egorized as biological processes: GO:0006631, the fatty acid metabolic process, and GO:0044255, the cellular lipid metabolic process. However, the rather large term sizes, i.e. the number of genes associated with this term, should be considered in further biological interpretation. The KEGG database also reveals two relevant terms in regard to a heart failure model, namely KEGG:00062, fatty acid elongation, and KEGG:01040, biosynthesis of unsaturated fatty acids. WikiPathways, another database under consideration, points to two more terms that should be examined: WP4351, omega-9 fatty acid synthesis, and WP4350, omega-3 / omega-6 fatty acid synthesis.



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0010942	positive regulation of cell death	667	2.2e-02
2	GO:BP	GO:2000047	regulation of cell-cell adhesion mediated by cadherin	17	2.6e-02
3	KEGG	KEGG:04115	p53 signaling pathway	72	2.4e-02
4	REAC	REAC:R-MMU-3371497	HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand	51	3.0e-02

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 3.14.: Highlighted Terms of Functional Enrichment Analysis for Cluster 4: This Manhattan diagram shows each significant term for cluster 4 (see fig. 3.11) from the different data sources as a circle on the X-axis, plotted as the adjusted p-values on the $-log_{10}$ scale on the Y-axis. Each color represents a different data sources. There are 4 relevant and therefore highlighted terms from 3 different categories, including biological processes from the GO database and pathways from the KEGG and Reactome database. The highlighted terms with their corresponding adjusted p-values, source, term ID, term name, and term size are listed in the table below the Manhattan diagram. (Constructed with the R package gProfiler2 [76].)

Genes in cluster 4 exhibit significant enrichment in 36 biological processes and pathways. In particular, two biological process terms in the GO database, namely GO:0010942 (positive regulation of cell death) and GO:2000047 (regulation of cell-cell adhesion mediated by cadherin), are of great interest. However, the large scope of term GO:0010942 should be carefully considered in further biological interpretation. In addition, one relevant term is identified in the KEGG database, i.e. KEGG:04115 (p53 signaling pathway). Furthermore, one pertinent term in the Reactome database, that is, R-MMU-3371497 (HSP90 chaperone cycle for steroid hormone receptors (SHRs) in the presence of ligand) is highlighted.

Another cluster, cluster 10, shows a significant decrease in gene expression after treatment with a high-fat diet in BL6 mice but not in CC mice, and is associated with a significant enrichment of terms relevant to heart failure analysis, such as R-MMU-2022090, which relates to the assembly of collagen fibrils and other multimeric structures. Cluster 3, in which the gene expression patterns of BL6 and CC mice evolve in opposite directions, contained interesting significant terms mapped to human phenotypes, including HP:0001638, which is related to cardiomyopathy.

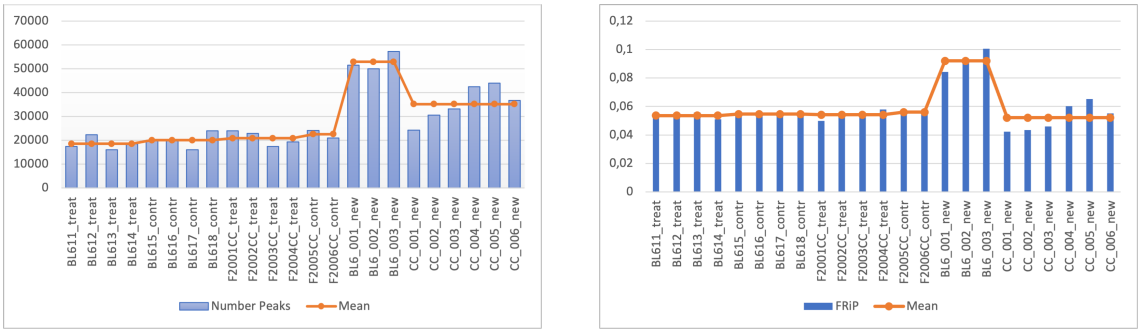
While a full biological interpretation should consider all or most of the significant terms identified in all clusters, for the sake of clarity, this analysis highlights only a few that are particularly relevant for the study of heart failure.

3.3. ATAC-seq Analysis

As outlined in section 4.1, the ATAC-seq samples (BL6 and CC) provided in the disease model are not of sufficient quality for further analysis (referred to as *_treat* and *_contr*). Therefore, the results obtained with these samples are only briefly presented in this section. Instead, attention is directed to a new dataset comprising three non-diseased BL6 and six CC samples, which provides substantially more robust results (referred to as *_new*).

3.3.1. Analysis of Open Chromatin Regions

In the context of ATAC-seq, peak calling serves to identify chromatin accessibility, facilitating the detection of regulatory elements and a better understanding of transcriptional regulation [107]. Various statistical measures, including the total number of peaks identified and the Fraction of Reads in Peaks (FRiP) score, can indicate the efficacy of the experiment. The FRiP score represents the proportion of usable reads in significantly enriched peaks divided by all usable reads, i.e. the proportion of all mapped reads that fall into the named peak regions [124]. These measurements for all BL6 and CC, old and new, are shown in fig. 3.15.



(a) Number MACS2 Peaks BL6 and CC (b) FRiP BL6 and CC

Figure 3.15.: Statistics - Peak Calling for BL6 and CC Samples: This figure displays statistical information related to the called peaks in all samples of the mouse strains BL6 and CC. The samples from the disease model are labelled *_treat* and *_contr* and are not subject to further analysis, while *_new* represents the newly provided samples. The blue bars represent the value for each sample, while the orange line indicates the mean value for each sample group. (a) shows the total number of peaks called by MACS2 for each sample. The data indicate a higher number of peaks in the new samples, particularly in BL6 mice. In contrast, the CC samples show a higher number of peaks in the new samples, with a relatively small difference compared to the old samples. (b) displays the FRiP score for each sample. The results show a twofold increase in the FRiP score for the new BL6 samples compared to the old samples. However, the CC samples present a similar FRiP score, with a slightly higher score for the old samples. (This plot is generated with Excel.)

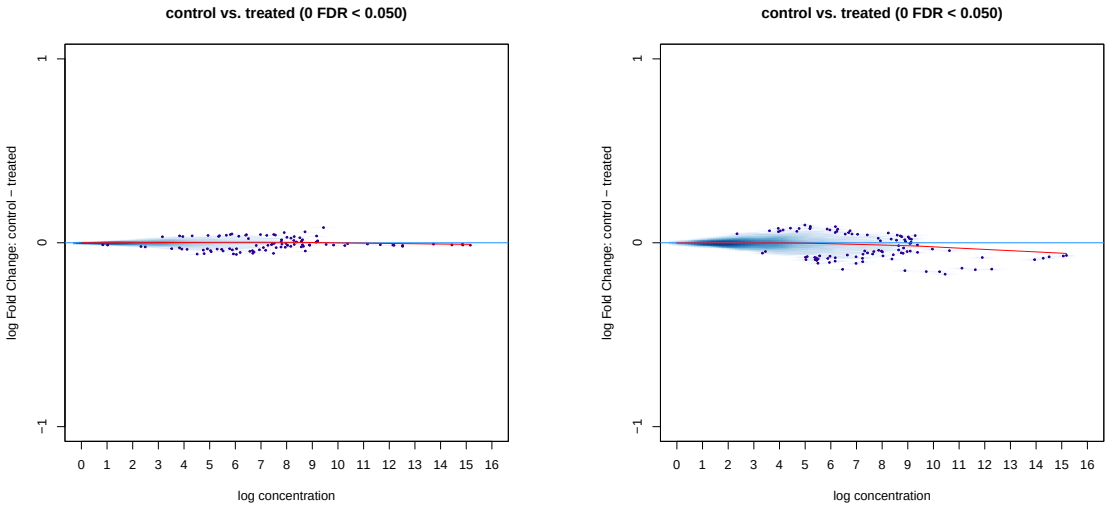
The mean total number of peaks called is compared between treated and untreated BL6 and CC samples and newly provided BL6 and CC samples. There are 18,569 significant peaks for treated BL6 samples, 20,066 for untreated BL6 samples, and 52,930 for new BL6 samples, which represents an increase in called peaks for the

new samples by a factor of approximately 2.85 and 2.64 respectively. In addition, the new BL6 samples also have a higher mean FRiP score of 0.092 compared to 0.054 and 0.055 for the treated and untreated BL6 samples. The difference between the old and new CC samples is not as significant, with a mean of 35,208 called peaks for the new CC samples, which is approximately 1.69 and 1.56 times the number of called peaks for the treated and untreated CC samples, respectively. However, the calculated FRiP score for the old CC samples is slightly higher than that for the new CC samples: 0.054 (treated) and 0.056 (untreated) as opposed to 0.052 for the new CC samples. MACS2 [25] simply pools the peak regions for certain replicates, which results in 39,166 peaks for BL6_treat, 43,395 peaks for BL6_contr, 89,213 peaks for BL6_new, 43,010 peaks for CC_treat, 36,952 peaks for CC_contr, and 91,479 peaks for CC_new.

Subsequently, to identify regions whose accessibility is altered by the high-fat diet treatment, DARs between the old treated and untreated sample groups are analysed using DiffBind. The new dataset is not included in the results, due to the absence of treated samples.

3.3.2. Analysis of Differential Accessibility Regions

The analysis reveals that 21,724 regions differ in BL6 between control and treatment samples. However, only 7 regions have a p-value less than 0.05. In addition, all regions have a FDR adjusted p-value of 1. A similar trend is observed for CC samples, with 20,662 regions identified and only 158 regions having a p-value less than 0.05. Again, all regions have a FDR adjusted p-value of 1, meaning that no DAR is significant for either BL6 or CC samples. These results are supported by the MA-plots for each species as shown in fig. 3.16.



(a) MA-plot BL6 (b) MA-plot CC

Figure 3.16.: MA-plot Differential Accessibility Regions BL6 and CC: This figure shows the MA-plots of DARs for BL6 and CC. The LFC is plotted on the Y-axis and the mean number of reads across all the samples for a accessible region is plotted on the X-axis [35]. Each dot represents a open chromatin region, while the dark blue color depicts the regions with slightly altered accessibility between control and treated. Very similar peak regions approach 0. The horizontal red curve represents a non-linear loess fit that captures the underlying relationship between coverage levels and LFCs. For BL6 and CC, no significant DARs (FDR adjusted p-value < 0.05) are reported between control and treated samples. (Constructed with the R package DiffBind [35].)

The degree of variability in chromatin accessibility between control and treated samples is slightly higher in CC than in BL6, although not noteworthy. Furthermore, the red curve in the MA-plots, a non-linear loess fit that captures the underlying relationship between coverage levels and LFCs [35], exhibits no significant differences. As stated in section 4.1, this could indicate potential issues with the experimental design.

3.4. Combining Epigenomic and Transciptomic Data

Investigating enhancer-gene interactions between DARs and DEGs can yield valuable insights into the epigenomic landscape of samples in a disease model. This approach can provide initial clues about the contrasting responses of BL6 and CC

to a high-fat diet. However, given the results presented in section 3.3.2, where no significant DARs are identified, the aforementioned analysis cannot be conducted as intended. Instead, this section analyses the newly provided dataset, which has a significantly higher number of called peaks and a higher FRiP score, at least in BL6. Due to the unavailability of treated samples, no DARs analysis is performed, but peaks identified from all BL6 and CC samples are pooled into a consolidated set for enhancer-gene interaction studies. Motif enrichment analysis then identifies TFs that regulate the transcription of all DEGs obtained.

3.4.1. Enhancer-Gene Interactions

For this analysis STARE [110] quantifies enhancer-gene interactions with the generalized ABC-score, for the pooled called peaks in BL6 and CC and their associated GTF file. The output is restricted to all DEGs of the mRNA-seq experiment to filter out regulatory elements that may affect these DEGs and thus be responsible for the observed differential expression. This can therefore explain phenomena in the heart disease model without the need for ATAC-seq data with treated samples.

STARE maps 7,578 open chromatin regions or peak regions for 1,662 DEGs in BL6, corresponding to approximately 4.5 regions per gene, with a generalized ABC-score greater than 0.02. In CC 7,672 regions are mapped to all 1,662 DEGs resulting in an average of 4.6 regions per gene.

The resulting regions are used to generate a FASTA file based on the respective BL6 and CC genome fasta file. This is necessary to identify known motifs in these regions, called motif enrichment analysis.

3.4.2. Motif Enrichment Analysis

The FASTA files, containing all open chromatin regions for BL6 and CC mapped to the DEGs determined by STARE, serve as the input for HOMER [115]. Since there are no proper background sequences, HOMER creates the background by shuffling the passed FASTA file. The ten most significant motifs of the resulting HTML files, which contain the enrichment of known motifs for BL6 and CC, are presented in

table 3.1 and table 3.2.

Table 3.1.: Motif Enrichment Analysis - BL6: This table contains the ten most significant motifs identified by the motif enrichment analysis with HOMER for BL6. The motif, the name of the motif, the p-value, the percentage of sequences assigned to the motif, and the percentage of background sequences assigned to the motif are listed. The motifs are ordered according to their p-value. (Motifs are provided by HOMER [115])

Motif	**Name**	**P-value**	**% of Targets Sequences with Motif**	**% of Background Sequences with Motif**
	GAGA-repeat	1e-464	71.90%	41.01%
	ERG (ETS)	1e-433	47.56%	20.01%
	SCL (bHLH)	1e-388	81.51%	54.61%
	Sp1 (Zf)	1e-350	21.91%	5.11%
	Bcl6 (Zf)	1e-347	32.98%	11.70%
	TF3A (C2H2)	1e-337	36.80%	14.58%
	KLF3 (Zf)	1e-335	28.15%	8.90%
	ZNF416 (Zf)	1e-335	39.84%	16.85%
	Sp5 (Zf)	1e-331	43.08%	19.42%
	Etv2 (ETS)	1e-314	31.69%	11.65%

Table 3.2.: Motif Enrichment Analysis - CC: This table contains the ten most significant motifs identified by the motif enrichment analysis with HOMER for CC. The motif, the name of the motif, the p-value, the percentage of sequences assigned to the motif, and the percentage of background sequences assigned to the motif are listed. The motifs are ordered according to their p-value. (Motifs are provided by HOMER [115])

Motif	**Name**	**P-value**	**% of Targets Sequences with Motif**	**% of Background Sequences with Motif**
	GAGA-repeat	1e-448	71.11%	41.11%
	SCL (bHLH)	1e-404	82.05%	55.09%
	ERG (ETS)	1e-403	47.80%	21.29%
	Sp1 (Zf)	1e-365	22.11%	5.11%
	ERRα (NR)	1e-343	57.34%	31.35%
	ZNF416 (Zf)	1e-339	41.04%	17.88%
	Sp5 (Zf)	1e-334	43.37%	19.82%
	TF3A (C2H2)	1e-331	37.11%	15.18%
	PU.1-IRF	1e-326	33.76%	12.97%
	Bcl6 (Zf)	1e-317	33.61%	13.07%

Table 3.1 and table 3.2 show the top ten of 608 and 609 significantly (FDR adjusted p-value < 0.05) enriched motifs in open chromatin regions of BL6 and CC, respectively. In total, 584 regulatory elements such as ERG (ETS) and SCL (bHLH)

overlap between the two strains, with only 25 unique to BL6 (e.g. GATA3) and 24 unique to CC (e.g. p53). While both tables contain several zinc-finger motifs such as Sp1 and ZNF416, the CC table also includes nuclear receptors like Errα, which have key roles in regulating gene expression [125]. In BL6, on the other hand, other TFs are identified, including ERG and Etv2, as part of the erythroblast transformation-specific family, which are involved in both physiological and pathological conditions [126]. All motifs in BL6 and CC are highly significant with p-values and an FDR adjusted p-value substantially smaller than 0.05. The FDR adjusted p-value is not presented in the table for this reason. Additionally, the percentage target sequence coverage of the top ten motifs ranges from 21.91% (for Sp1) to 81.51% (for SCL) in BL6, and from 22.11% (for Sp1) to 82.05% (for SCL) in CC. The percentage coverage of background sequences ranges from 5.11% (for Sp1) to 54.61% (for SCL) in BL6, and from 5.11% (for Sp1) to 55.09% (for SCL) in CC. These measurements provide additional insights into the significance of the enriched motifs in each dataset.

In both BL6 and CC, highly enriched TFs, including GAGA-repeat, ERG, SCL, Sp1, Bcl6, TF3A, ZNF416, and Sp5, are detected. The TFs KLF3 and Etv2 are exclusively identified in BL6 and the TFs Errα and PU.1-IRF are specifically identified in CC as part of the ten most significant motifs. However, these TFs are observed to be comparatively less significant but still detected in the other mouse strain. For example, the TF p53 in CC and GATA3 in BL6 are completely unique and statistically significant. It should be noted that a comprehensive biological interpretation of these results should include all or most of the significantly enriched motifs, but for clarity only the top ten most relevant motifs according to p-value are displayed in this analysis.

4. Discussion and Outlook

4.1. Discussion

Difficulties Encountered During the Analysis

Given the complexity and relevance of biological data, it is important to emphasize that high-quality data are essential for precise and meaningful analyses. Thus, the quality of the data can significantly impact the accuracy and reliability of the results obtained through the analysis, which highlights the importance of ensuring that the data meet the required standards before proceeding with the analysis.

In this study, quality control measures and subsequent analysis show critical data quality, particularly in the hybrid datasets (XB and YB), as discussed in more detail below. In particular, the hybrid data sets are expected to provide information on allele-specific expression and imprinting phenomena. Therefore, mRNA-seq analysis identifying DEGs in a disease model is anticipated to yield preliminary results. However, the results reveal a surprising absence of DEGs in the XB and YB data sets, which is inconsistent with the observed phenotype, such as weight gain, as shown in fig. 1.2. Notably, both hybrid samples exhibit significant weight gain, especially XB. This suggests that some differences in gene expression should be identified. In addition, the featureCounts assignments to the different genomes, as shown in fig. 3.6c and fig. 3.6d, reveal a highly uneven distribution between the BL6 and CC genome, which is unexpected given the analysis of a pilot dataset. Overall, the aforementioned factors suggest that the sequencing quality of the dataset is insufficient, resulting in the exclusion of all XB and YB samples from the differential expression analysis. In addition, these samples are excluded from the ATAC-seq experiment as the identification of accessible chromatin regions and their association

with DEGs is not achievable, rendering these datasets meaningless for this specific analysis.

A decent number of peaks (approximately > 100,000) and a relatively acceptable FRiP score (approximately > 0.3) is supposed to be achieved by using a peak caller like MACS2 [25] in the analysis of ATAC-seq samples [124]. However, the BL6 and CC samples from the first experiment presented in section 3.3, containing an average of about 20,000 peaks and a FRiP score of around 0.05 (see fig. 3.15), are insufficient for biological interpretation, especially with respect to DEGs. Furthermore, no significant DARs were found. More precisely, there is a very small amount of DARs with a p-value of less than 0.05, as shown in section 3.3.2. The FDR adjusted p-value is 1 for each region, which indicates among all regions called significant, 100.0% of these are truly null [127]. This might suggest issues with the design of the experiment. The observed problems may be attributed to the high percentage of duplicated reads, as visualized in fig. 3.4, which can reach up to 70.0%. Therefore, these samples are excluded from further investigation and a new ATAC-seq dataset with three BL6 and six CC non-diseased samples is created to obtain viable results. Although the quality metrics of the new dataset are not substantially different from those of the old dataset, MACS2 provides a significantly higher number of peaks, up to 60,000, and a FRiP score of 0.092 for BL6. Despite the fact that the FRiP score for the newly obtained CC samples is rather similar to the old dataset, the new data is selected for further analysis due to the higher number of peaks and the critical results of DiffBind [35]. Consequently, the newly identified open chromatin regions can be combined with DEGs obtained by mRNA-seq to provide important insights into heart failure.

In light of these observations, this section highlights fundamental problems associated with the data provided, particularly with respect to the hybrid and ATAC-seq samples. The exclusion of these samples from the analysis in turn complicates the investigation of imprinting phenomena, given the absence of crossed samples. Nevertheless, the analysis still provides intriguing insights into the differential response of BL6 and CC to high-fat diet inducing heart failure.

Quality Control for mRNA-seq and ATAC-seq Data

The analysis of NGS data involves several steps and is not limited to a single approach. Quality control is essential to ensure the availability of uncontaminated data to provide biological significance, as demonstrated by the problems encountered with the data. FastQC [50] and MultiQC [56] are suitable tools to address this challenge.

On the one hand, the adapter content of all mRNA-seq samples, as seen in fig. 3.1, remains in an unproblematic range of under 5.0% of sequences at each position. Therefore, the influence of adapter contamination in these samples on further analytical results, such as alignment rate, is negligible and no adapter trimming is required. On the other hand, there is a problematically high level of adapter contamination in all ATAC-seq samples (see fig. 3.2), with up to 70.0% of all sequences containing the Nextera transposase sequence ('CTGTCTCTTATA'). This can have a major impact on the upcoming analysis, as there is a high risk that untrimmed adapter sequences may interfere with the alignment or mapping of the reads to a reference genome [53]. Consequently, adapter trimming is necessary in order to prevent falsification of analysis results.

Both sequencing experiments show a high number of duplicated sequences, which however should be treated differently. As shown in fig. 3.3, approximately 50.0% to 70.0% of all sequences are duplicated more than 10,000 times. In total, all mRNA-seq samples contain approximately 80.0% to 95.0% of duplicated reads. Although high rates of read duplicates in RNA-seq libraries may appear to be artefacts, it is important to note that they do not necessarily indicate insufficient library complexity caused by low sample input or overamplification. Instead, they may be indicative of a very high abundance of a small number of genes [50]. This is closely related to the overrepresented sequences displayed in fig. 3.5. The fact that some individual sequences make up a larger proportion of the total sequence set can be attributed to the same reason. It is not advisable to remove duplicates computationally during analysis, as the focus of this experiment is the analysis of DEGs, which requires high expression levels [50]. Figure 3.4 indicates that approximately 10.0% to 25.0% of all reads in the ATAC-seq samples are duplicated between 1,000 and 5,000 times.

Overall, there are 40.0% to 70.0% of duplicated reads. In contrast to mRNA-seq experiments, the duplicated reads observed in ATAC-seq experiments are presumably generated by PCR duplicates, which represent artefacts and may introduce bias in downstream analyses [107]. The significant proportion of duplicated reads observed in this study indicates sample preparation issues. To address these problems, which should be considered in the biological interpretation of the data, the SAMtools markdup [106] function is used to mark and remove the duplicates.

A crucial step in the study of differential gene expression is the accurate assignment of reads to specific genes. The percentage of reads assigned to genes is another critical indicator of data quality. In this study, an allele-specific analysis is performed on mRNA-seq data using snakePipes [21]. The assignment of reads to the allele-flagged BAM file and to the two allele-specific BAM files is examined. In all BL6 samples, a significantly higher percentage of reads are assigned to and in the BL6 genome, as shown in fig. 3.6a. Conversely, in all CC samples, a substantially higher percentage of reads are assigned to and in the CC genome, as displayed in fig. 3.6b. This result can be expected as the homozygous samples have a higher match with their corresponding genomes. The heterozygous samples also exhibit a highly uneven distribution of read assignments between the two genomes, as presented in fig. 3.6c and fig. 3.6d. This observation and its implications have been discussed previously.

The overall alignment rate of reads to the reference genome is an essential metric for evaluating the quality and significance of results from different experiments. In particular, a low alignment rate or a high proportion of multimapping reads are indicative of poor data quality and therefore limit the biological interpretability of the results [59]. However, a range of 63.8% to 76.1% uniquely assigned reads for all mRNA-seq samples (see fig. 3.7) can be considered sufficient to draw biologically relevant conclusions. Similarly, the ATAC-seq samples with an alignment rate of 61.6% to 80.0% and an average of 50.0% uniquely assigned reads (see fig. 3.8) are considered suitable for further analyses.

Despite the featureCount assignments for hybrid samples and the unusually high level of duplication for all ATAC-seq samples, the reported quality measurements

warrant further testing of the homozygous samples in subsequent analyses.

Differentially Expressed Genes Between Treated and Untreated Groups of mRNA-seq CC and BL6 Samples

Examining statistically significant differences in read counts for a given gene between two experimental conditions or genomes provides information about the response of samples to treatment and, in this study, the development of HFpEF [74]. As mentioned above, the heterozygous samples do not exhibit any significant differences in contrast to the observed phenotype and are therefore not considered further. In fig. 3.9, a total of 1560 significant DEGs with an adjusted p-value of less than 0.05 are identified in BL6 samples, while 102 are found in CC samples, with 27 DEGs overlapping. This in turn suggests that the high fat diet has a greater effect on BL6 samples compared to CC samples, which is supported not only by MA-plots (fig. 3.10b and fig. 3.10a) but also by phenotypic observations as depicted in fig. 1.2 and fig. 1.3. The weight gain observed within 15 weeks of a high-fat diet was significantly higher in BL6, at approximately 20 grams, than in CC, with little or no weight gain observed in CC. In addition, the increase in left ventricular heart mass may indicate heart failure, as it limits the ability of the heart to pump an adequate amount of blood throughout the body [42]. Another indicator of heart failure is the increase in left atrial pressure with a concomitant decrease in passive left ventricular filling, resulting in a higher mitral E wave to E' wave ratio [45]. Given that these two factors were significantly different in the BL6 samples but relatively similar in the CC samples, it can be assumed that the high-fat diet has a more pronounced effect on BL6 than on CC.

Furthermore, the classification of these DEGs into specific functional groups provides additional evidence for their involvement in underlying biological processes in the heart.

Analysis of Biological Functions Based on Gene Clustering in mRNA-seq

The previously obtained gene expression data is subjected to hierarchical clustering, resulting in ten informative cluster profiles, that are subsequently associated with

annotated biological functions and pathways. The expression of these gene clusters is visualized by a heatmap (see fig. 3.11) and the differences between the untreated and treated sample groups are highlighted using a line plot (see fig. 3.12). Several clusters with interesting differences between BL6 and CC samples are identified for further analysis. Clusters 5, 6, 7, and 8 exhibit a distinct upregulation of gene expression in BL6 mice from untreated to treated, whereas they remain relatively unchanged in CC mice. This result suggests that the annotated biological functions of these cluster profiles are particularly relevant for BL6 mice. Conversely, clusters 9 and 10 contain genes whose expression is significantly reduced in BL6 mice but not in CC mice after treatment with a high-fat diet, indicating biological functions that are downregulated in BL6 mice. Clusters 2 and 3 display contrasting gene expression patterns between BL6 and CC mice, indicating possible divergent underlying processes or pathways. In contrast, cluster 1 shows a similar decrease in gene expression in both mouse strains. Only cluster 4 reveals a significant increase in expression for treated CC samples and consistent expression for BL6 mice, which may be explained by the low number of DEGs and the overall low effect of treatment on CC.

Cluster 8, as a representative of the first described cluster group, contains 21 genes and is associated with 16 significantly enriched processes identified by functional enrichment analysis with gProfiler2 [76]. A significant biological process is GO:0006631, which describes chemical reactions and pathways involving fatty acids, aliphatic monocarboxylic acids liberated from naturally occurring fats and oils by hydrolysis [78] and is a child term of GO:0044255, describing the cellular lipid metabolic process. The upregulation of this process might indicate a shift in the heart's primary source of energy production from glucose to fatty acids [128]. While glucose concentration appears to be an important factor for reducing cardiac damage, a shift to fatty acid metabolism can ultimately lead to the accumulation of oxidative stress, which can contribute to the progression of heart failure [129]. Fatty acid elongation and the biosynthesis of unsaturated fatty acid pathways, KEGG:00062 and KEGG:01040, are closely related to the altered fatty acid metabolism observed in heart failure. The shift in energy metabolism is accompanied by changes in the expression and activity of enzymes involved in fatty acid elongation and unsaturated

fatty acid biosynthesis, induced by a high-fat diet [130]. Therefore, the observation of this particular metabolic phenotype only in BL6 and not in CC can suggest that BL6 is more responsive to the high-fat diet.

In contrast, cluster 4, containing 77 genes and 36 significantly enriched terms, shows upregulated processes in CC but not BL6, including GO:2000047 (regulation of cell-cell adhesion mediated by cadherin) and GO:0010942 (positive regulation of cell death). Cadherins are a superfamily of transmembrane glycoproteins that mediate homophilic and Ca^{2+}-dependent cell-cell adhesion [131]. A study [132] has revealed that the absence of cadherin in the mouse heart results in dilated cardiomyopathy and impaired cardiac function attributable to the absence of myofibril anchoring to the plasma membrane via cadherin-mediated mechanisms. These findings highlight the essential role of cadherin in the maintenance of normal cardiac function [131]. The HSP90 chaperone cycle for SHRs in the presence of ligand pathway is also upregulated in CC. The highly dynamic interactions of SHRs with HSP90 complexes regulate SHR cellular location, protein stability, steroid hormone binding ability, and transcriptional activity [133]. An upregulation could indicate an increased demand for protein folding and stabilization, potentially in response to stress or other physiological stimuli [134]. The upregulated processes observed in CC, as well as the cardioprotective effects of steroid hormone receptors, such as their anti-inflammatory and anti-apoptotic effects [134], may contribute to the absence of heart failure in CC. These mechanisms are not significantly increased in BL6.

Cluster 10, with 90 genes and 125 significantly enriched terms, shows significantly downregulated processes and pathways in BL6. One of these annotated by the Reactome database is R-MMU-2022090, the assembly of collagen fibrils and other multimeric structures. Structural proteins, i.e. fibrillar collagens and matricellular proteins, form the extracellular matrix (ECM) in the heart [135]. In most cardiovascular diseases, the ECM is severely remodeled, meaning that it provides structural support and plays a key role in maintaining the mechanical properties of the heart [136]. Downregulation of this pathway in the ECM can be a potentially important contributor to the pathophysiology of heart failure.

The Human Phenotype Ontology serves as a standardised terminology for describ-

ing phenotypic abnormalities observed in human disease. In addition, the database has mapped all human phenotypes to corresponding mouse models. It is therefore of great interest to study these terms, as this suggests that the investigated mouse genes have corresponding human orthologs that are associated with observable phenotypes in human diseases. [87] HP:0001638, described as cardiomyopathy, is one of the 75 significantly enriched terms in cluster 3, which contains 112 genes. Cardiomyopathies are a heterogeneous group of heart muscle diseases and a major cause of heart failure [137]. An increased expression of genes associated with this disease in BL6 and a rather decreased expression in CC supports the observed phenotypes in both sample groups.

Analysis of the different cluster profiles indicates that the two mouse strains have divergent metabolic and physiological responses to a high-fat diet, which may contribute to differences in heart disease susceptibility. For example, BL6 shows a marked upregulation of fatty acid metabolism in response to a high-fat diet, whereas CC does not. In addition, CC exhibits upregulation of cardioprotective mechanisms, including cadherin-mediated adhesion and steroid hormone receptor pathways, which are not significantly increased in BL6.

Accessible Regions Chromatin - ATAC-seq

For adequate interpretation of ATAC-seq samples, a sufficient number of peaks and a reasonable FRiP score are crucial. Current standards are defined by ENCODE, which can be used to assess the quality of the results. ENCODE recommends a total number of peaks within a replicate peak file of more than 150,000, while more than 100,000 may be acceptable [124]. The initial ATAC-seq samples examined in this study, with a mean of approximately 40,000 peaks in CC and BL6, are far below the specified reference value. In contrast, the new samples, with around 90,000 peaks in both strains, are still below, but very close to, this threshold. Furthermore, the FRiP score is recommended to be higher than 0.3, while values greater than 0.2 may be acceptable [124]. Both old and new samples are well below this threshold at approximately 0.055. However, the newly obtained BL6 samples perform a lot better with a score of 0.092. Other metrics to check the quality of the analysis, such as

the fragment length distribution or transcription start site enrichment, can be used for further evaluation. However, the alarming results of the DARs analysis, with no significant differential regions detected between treated and untreated sample groups (see fig. 3.16), indicate potential problems in the design of the experiment. The reason for this assumption is the measured response of the BL6 and CC samples to the high-fat diet, as shown in fig. 1.2. Furthermore, the analysis of DEGs proves a change of gene expression in the treated samples, particularly in BL6. Since TFs exert a major influence on gene expression [111] and open chromatin regions (measured by ATAC-seq) serve as binding sites for TFs [23], it is reasonable to assume the presence of some DARs, at least in BL6. This information, combined with the small number of peak regions and a low FRiP score, leads to the exclusion of these samples from further analysis.

As described above, the new ATAC-seq samples possess a higher biological significance due to the substantially higher amount of peaks. Since ATAC-seq samples are required to potentially explain gene expression changes based on TFs, this dataset will be further investigated by measuring enhancer-gene interaction with STARE [110] and performing motif enrichment analysis with HOMER [115].

Combining Differentially Expressed Genes and Highly Enriched Motifs

The output of STARE is filtered for all 1662 DEGs, resulting in 7578 regulatory regions for BL6 and 7672 for CC. The number of regulatory elements is expected to be higher than the number of genes because genes usually have multiple TFs. Each TF can mediate a variety of transcriptional regulatory instructions that can affect gene expression in response to different biological conditions [138]. Identifying TFs that bind to the regulatory regions and classifying their biological impact on cardiac function is very important. Therefore, the resulting regions are formatted in FASTA format and then loaded into HOMER for a motif enrichment analysis. A total of 608 motifs are identified as significantly enriched in BL6 and 609 in CC (FDR adjusted p-value < 0.05). The similar amount of motifs could be attributed to the comparable number of peaks in both strains, the same set of genes under consideration, and the lack of treated samples. This may also explain why 584 TFs

such as ERG (ETS) and SCL (bHLH) overlap between the two strains, with only 25 being unique in BL6 and 24 being unique in CC. However, studying the role of TFs in cardiovascular disease may explain some of the observed differences in the effects of a high-fat diet between the two species of mice.

The set of 584 shared TFs comprises a high number of TFs with a zinc finger DNA binding domain, including Sp1 and ZNF416. Zinc-finger proteins have well-established roles in various cellular processes, including transcriptional regulation, signal transduction, ubiquitin-mediated protein degradation, DNA repair, actin targeting, cell migration, and more [139]. For example, Sp1, as part of the specifity protein family, is a TF that is primarily localized in the nucleus and modifies the transcription of target genes [140]. This TF has been implicated in the development and progression of various cardiovascular diseases, including atherosclerosis, hypertension, and heart failure. Furthermore, Sp1 is known to play an essential role in LDLR gene expression, as LDLR contains Sp1-binding elements. [141] While LDLR is critical for the clearance of LDL cholesterol in plasma, a reduced expression could lead to premature cardiovascular disease [142]. This gene is found to be differentially expressed in cluster 9, which exhibits significant downregulation in BL6 samples. Consequently, it can be hypothesized that the observed downregulation of LDLR in BL6 samples may be due to a decreased impact of the TF Sp1 in treated BL6 samples compared to control samples. This in turn may contribute to the fact that the high-fat diet has a greater effect on BL6 than on CC.

Other TFs identified in both mouse strains, such as ERG and Etv2, as part of the erythroblast transformation-specific family, play important roles under physiological and pathological conditions [126]. ERG is critical in promoting endothelial homeostasis by regulating lineage-specific genes, such as CDLN5, and suppressing the expression of proinflammatory genes, like ICAM1 [143]. ICAM1 codes for a protein that is involved in cell adhesion and the immune response. It is expressed on the surface of certain cells, such as endothelial cells, and controls the recruiting of leukocytes to sites of inflammation. The overexpression of ICAM1 has been implicated in various diseases, including cardiovascular disease. [144] The strong downregulation of ICAM1 in BL6 seen in Cluster 10 may indicate reduced recruitment of immune

cells and reduced inflammation in the heart and therefore an increased activity of ERG. This may have a beneficial effect in slowing the progression of heart disease and may be one of the first responses to a high-fat diet.

Additional information that may explain the differences in cardiac function between BL6 and CC is provided by the study of strain-specific TFs. The tumor suppressor p53 is a TF that translates growth and survival signals into specific gene expression patterns, regulating tumor-free survival of an organism. Additionally, p53 functions as a master regulator of an intricate TF network to maintain cardiac tissue homeostasis. [145] The significant enrichment of p53 in CC, but not BL6, could indicate that the missing involvement of p53 in biological processes such as energy metabolism and the oxidative stress response with the inhibition of hypertrophic signaling and apoptosis [145], is decisive for the observed response of BL6. This hypothesis is supported by the upregulation of CDKN1A/p21 in CC, as visualized in cluster 4, which is directly regulated by p53 [146]. According to research, CDKN1A/p21 has a cardioprotective effect against ischemia-reperfusion injury by inhibiting oxidative stress. This mechanism may prevent endothelial dysfunction, inflammation, fibrosis, and hypertrophy, all of which can contribute to heart failure. [147] [148]

However, it should be mentioned that mechanisms underlying the up- or downregulation of different genes through TFs are highly controversial. For example, CDKN1A/p21 is also found to favour the development of cardiac hypertrophy, and the downregulation of CDKN1A/p21 expression could be an approach to attenuate hypertrophic growth of cardiomyocytes [149]. Thus, the interpretation of results should always be considered in the context of the current study.

In summary, there is considerable overlap of enriched motifs between BL6 and CC, including several TFs such as Sp1 and ERG. Treatment of BL6 samples results in decreased expression of LDLR due to reduced activity of Sp1 [141], while downregulation of ICAM1 may serve as a slowing factor in the progression of heart disease in BL6 through the regulation of ERG [144]. In addition, the CC-specific TF p53 plays a key role in maintaining cardiac tissue homeostasis [145]. These results may provide initial approaches to the differences in cardiac function adaptation be-

tween BL6 and CC induced by a high-fat diet. However, the role of the identified TFs in heart disease appears to be complicated and context-dependent, and further research is needed to fully understand their contribution to cardiac pathophysiology.

4.2. Outlook

This study reveals distinct gene expression profiles in BL6 and CC mouse strains in response to a high-fat diet. The identified gene clusters and their associated biological functions could provide a basis for future research into the underlying mechanisms of diet-induced heart failure and potential therapeutic targets. Moreover, analysing the specific genes and pathways involved in the upregulated and downregulated processes may reveal new targets for early detection of cardiac dysfunction or for treatment purposes. In addition, the motif enrichment analysis identified significant TFs that may be responsible for the observed up- or downregulation of genes. This may contribute to an improved understanding of how drugs can modulate the activity of these TFs and positively regulate cardiovascular disease. The differences in gene expression patterns between BL6 and CC mice suggest that these strains may be suitable models for studying the mechanisms of heart failure and developing personalised therapies. This is particularly relevant for allele-specific expression, as the study of allele-specific expression can provide valuable insights into the genetic basis of complex traits and diseases. To explore the impact of genetic differences on heart failure, it is advisable to repeat the disease model with both XB and YB samples, as quality control measurements and further investigation in this study indicate a low biological significance of the current hybrid samples. The study of the differential inheritance of the maternal and paternal alleles of a mouse strain is of fundamental importance. Allele-specific expression and imprinting phenomena can only be elucidated by hybridising two mouse strains, in which the maternal and paternal alleles of, in this case BL6 and CC, are exchanged in both directions. Future studies could also investigate the effects of other dietary treatments on gene expression profiles of these strains to identify common and unique mechanisms underlying diet-induced cardiac dysfunction. Overall, these results provide a basis for further

exploration of the biological mechanisms of heart failure and potential therapeutic strategies through the combination of mRNA-seq and ATAC-seq experiments.

4.3. Conclusion and Perspective

The study of epigenetic changes and their implications for gene expression as a consequence of a high-fat diet in different mouse samples reveals several approaches that could provide reasons for the distinct response of BL6 and CC. Conclusions about allele-specific gene expression could be drawn by comparing these two mouse strains, which differ in their cardiac function. Furthermore, this study confirms the importance of high quality sequencing data to ensure an accurate analysis. This is particularly evident in the insufficient number of peaks and the unexpected absence of statistically significant DARs of the initial ATAC-seq data and the exclusion of the hybrid XB and YB data sets.

The classification of the different gene expression clusters of treated and untreated CC samples reveals upregulation of genes involved in cardioprotective mechanisms, including cadherin-mediated adhesion and SHR pathways. In addition, associated TFs are identified to be highly enriched exclusively in CC. These include p53, which plays a key role in cardiac homeostasis and directly regulates CDKN1A/p21, that is significantly upregulated in this strain. In particular, this gene has a cardioprotective effect against ischaemia-reperfusion injury by inhibiting oxidative stress. On the other hand, BL6 exhibits an upregulation of genes involved in pathways associated with the development of heart failure, such as fatty acid metabolism. Additionally, other genes important for maintaining a healthy heart are significantly downregulated, such as LDLR, which is critical for the clearance of LDL cholesterol in the plasma. The TF Sp1 activates or enhances the expression of LDLR and has been detected to be active in untreated BL6, which may indicate a reduced activity of this regulatory element in treated samples. However, it is important to acknowledge that the regulation of gene expression by TFs can be complex and multifaceted. While some studies may show that a particular TF or gene promotes a certain disease or condition, other studies may demonstrate the opposite effect. Therefore, it

is essential to interpret the results of a study in light of previous research and to be aware of the limitations and uncertainties in the current understanding of gene regulation. Nevertheless, it appears that the varying activity of several TFs and the associated up- or downregulation of genes form the foundation for the differential response of BL6 and CC to treatment with a high-fat diet. This can also be related to allele-specific expression, as the genetic differences between the two mouse strains could lead to allele-specific expression changes in genes that contribute to the differential response. This requires further investigation as it cannot be confirmed by this analysis. Repeating the disease model of XB and YB is very suitable for this.

Many more or all of the significantly different genetic and epigenetic factors discovered in this work should be investigated in further analyses to obtain a comprehensive epigenetic understanding in terms of gene expression that explains the different cardiac functions of BL6 and CC. By including hybrid samples, allele-specific expression and imprinting phenomena can be revealed, bringing everything together into a clear picture. Ultimately, the results of this study may pave the way towards the development of novel therapeutic strategies or early diagnostic approaches for heart diseases, such as the identification of TFs as putative drug targets.

Bibliography

[1] Amy Groenewegen et al. "Epidemiology of heart failure". In: *European journal of heart failure* 22.8 (2020), pp. 1342–1356. DOI: 10.1002/ejhf.1858.

[2] Yan H et al. "Allelic variation in human gene expression". In: *Science* 297(5584) (Aug. 2002), p. 1143. DOI: 10.1126/science.1072545.

[3] AMA Ahn B et al. "Analysis of allele-specific expression using RNA-seq of the Korean native pig and Landrace reciprocal cross". In: *Asian-Australas J Anim Sci.* 32(12) (May 2019), pp. 1816–1825. DOI: 10.5713/ajas.19.0097.

[4] B. M. Cattanach and J. Jones. "Genetic imprinting in the mouse: Implications for gene regulation". In: *Journal of Inherited Metabolic Disease volume* 17 (1994), pp. 403–420. DOI: 10.1007/BF00711356.

[5] Thorvaldsen JL, Duran KL, and Bartolomei MS. "Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2". In: *Genes Dev.* 12(23) (1998), pp. 3693–3702. DOI: 10.1101/gad.12.23.3693.

[6] Amir Ben-Dor et al. "Tissue classification with gene expression profiles". In: *Reecomb 'oo* (2000), pp. 54–56. DOI: 10.1145/332306.332328.

[7] Varija N Budhavarapu, Myrriah Chavez, and Jessica K Tyler. "How is epigenetic information maintained through DNA replication?" In: *Epigenetics & chromatin* 6.1 (2013), pp. 1–7. DOI: 10.1186/1756-8935-6-32.

[8] Elizabeth C Bryda. "The Mighty Mouse: the impact of rodents on advances in biomedical research". In: *Missouri medicine* 110.3 (2013), p. 207.

[9] The Jackson Laboratory. *What is a mouse model?* 2023. URL: https://www.jax.org/why-the-mouse/model (visited on 03/07/2023).

[10] Stephan Hermann Georg Symons. "Analysis and Visualization of Gene Expression Data". PhD thesis. Universität Tübingen, 2011.

[11] National Cancer Institute. *Gene Expression Profile.* URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/gene-expression-profile (visited on 11/25/2022).

[12] Alexander F. Palazzo and Eliza S. Lee. "Non-coding RNA: what is functional and what is junk?" In: *Frontiers in Genetics* 6 (2015), p. 2. DOI: 10.3389/fgene.2015.00002.

[13] Jiannan Guo. "Transcription: the epicenter of gene expression". In: *J Zhejiang Univ Sci B.* 15 (2014), pp. 409–411. DOI: 10.1631/jzus.B1400113.

[14] Elissa J. Chesler and Ryan W. Logan. *Bioinformatics of Behavior: Part 2.* 2012.

[15] *Getting started with RNA-Seq analysis (bulk and single cell)*. URL: https://btep.ccr.cancer.gov/getting-started/ (visited on 01/26/2023).

[16] Vladimir Kiselev et al. *Analysis of single cell RNA-seq data*. URL: https://scrnaseq-course.cog.sanger.ac.uk/website/index.html (visited on 01/26/2023).

[17] Friederike Dündar. *Analysis of bulk RNA-seq data and Analysis of Next Generation Sequencing Data*. 2018. URL: https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/07_lecture.pdf (visited on 01/26/2023).

[18] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental and Molecular Medicine* 50 (2018). DOI: 10.1038/s12276-018-0071-8.

[19] *Fluorescence-activated Cell Sorting (FACS)*. URL: https://www.sinobiological.com/category/fcm-facs-facs (visited on 01/26/2023).

[20] Maria Doyle, Belinda Phipson, and Harriet Dashnow. *RNA-Seq reads to counts (Galaxy Training Materials)*. 2020. URL: https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-reads-to-counts/tutorial.html (visited on 05/17/2022).

[21] Vivek Bhardwaj et al. "snakePipes: facilitating flexible, scalable and integrative epigenomic analysis". In: *Bioinformatics* 35.22 (May 2019), pp. 4757–4759. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz436. URL: https://doi.org/10.1093/bioinformatics/btz436.

[22] Buenrostro JD et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Curr Protoc Mol Biol.* 109 (2015). DOI: 10.1002/0471142727.mb2129s109.

[23] Chao Zhou et al. "Accessible chromatin regions and their functional interrelations with gene transcription and epigenetic modifications in sorghum genome". In: *Plant Communications* 2.1 (2021). Plant Genome Biology, p. 100140. ISSN: 2590-3462. DOI: https://doi.org/10.1016/j.xplc.2020.100140. URL: https://www.sciencedirect.com/science/article/pii/S2590346220301838.

[24] Illumina. *A rapid, sensitive method for profiling accessible chromatin across the genome*. 2022. URL: https://www.illumina.com/techniques/popular-applications/epigenetics/atac-seq-chromatin-accessibility.html (visited on 10/06/2022).

[25] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome biology* 9.9 (2008), pp. 1–9. DOI: https://doi.org/10.1186/gb-2008-9-9-r137.

[26] Stefan Dillinger Ph.D. *Complete Guide to Understanding and Using ATAC-Seq*. 2022. URL: https://www.activemotif.com/blog-atac-seq (visited on 10/06/2022).

[27] Chao Zhou et al. "Accessible chromatin regions and their functional interrelations with gene transcription and epigenetic modifications in sorghum genome". In: *Plant Communications* 2.1 (2021), p. 100140. DOI: 10.1016/j.xplc.2020.100140.

[28] Maria Tsompana and Michael J Buck. "Chromatin accessibility: a window into the genome". In: *Epigenetics & chromatin* 7.1 (2014), pp. 1–16. DOI: 10.1186/1756-8935-7-33.

[29] Liesbeth Minnoye et al. "Chromatin accessibility profiling methods". In: *Nature Reviews Methods Primers* 1.1 (2021), p. 10. DOI: 10.1038/s43586-020-00008-9.

[30] Fiorella C Grandi et al. "Chromatin accessibility profiling by ATAC-seq". In: *Nature Protocols* 17.6 (2022), pp. 1518–1552. DOI: 10.1038/s41596-022-00692-9.

[31] John M Gaspar. "Improved peak-calling with MACS2". In: *BioRxiv* (2018), p. 496521. DOI: 10.1101/496521.

[32] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1 (2010), pp. 139–140. DOI: https://doi.org/10.1093/bioinformatics/btp616.

[33] M.I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". Version 15. In: *Genome Biol* (2014). DOI: 10.1186/s13059-014-0550-8.

[34] Paul Gontarz et al. "Comparison of differential accessibility analysis strategies for ATAC-seq data". In: *Scientific reports* 10.1 (2020), pp. 1–13. DOI: 10.1038/s41598-020-66998-4.

[35] R Stark and G Brown. "DiffBind: differential binding analysis of ChIP-Seq peak data". In: *Bioconductor* (2022). URL: http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf (visited on 01/27/2023).

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[37] Hassan Rajabi-Maham et al. "The south-eastern house mouse Mus musculus castaneus (Rodentia: Muridae) is a polytypic subspecies". In: *Biological Journal of the Linnean Society* 107.2 (Sept. 2012), pp. 295–306. ISSN: 0024-4066. DOI: 10.1111/j.1095-8312.2012.01957.x. eprint: https://academic.oup.com/biolinnean/article-pdf/107/2/295/16708306/bij1957.pdf. URL: https://doi.org/10.1111/j.1095-8312.2012.01957.x.

[38] Meidong Jing et al. "Phylogeography of Chinese house mice (Mus musculus musculus/castaneus): distribution, routes of colonization and geographic regions of hybridization". In: *Molecular Ecology* 23 (July 2014), pp. 4387–4405. DOI: 10.1111/mec.12873.

[39] M. Phifer-Rixey, B. Harr, and J. Hey. "Further resolution of the house mouse (Mus musculus) phylogeny by integration over isolation-with-migration histories". In: *BMC Evol Biol* 20 (2020). DOI: 10.1186/s12862-020-01666-9.

[40] Petko M Petkov et al. "An efficient SNP system for mouse genome scanning and elucidating strain relationships". In: *Genome research* 14.9 (2004), pp. 1806–1811. DOI: 10.1101/gr.2825804.

[41] Phillip Grote (PhD). *Haploid Expression of LncRNAs in the Cardiac System*. 2022. URL: https://1drv.ms/p/s!Ao_cDhG2Uzo4pttxPZ2vxCX1mO4CUw?e=qubSKl (visited on 07/06/2022).

[42] M.D Cecilia Gutierrez and M.D. Daniel G. Blanchard. "Diastolic Heart Failure: Challenges of Diagnosis and Treatment". In: *Am Fam Physician* 69 (2004), pp. 2609–2617.

[43] Grote LAB. 2023. URL: https://grotelab.com (visited on 03/07/2023).

[44] Novogene. 2023. URL: https://www.novogene.com (visited on 03/07/2023).

[45] Park JH and Marwick TH. "Use and Limitations of E/e' to Assess Left Ventricular Filling Pressure by Echocardiography". In: *J Cardiovasc Ultrasound* 19(4) (2011), pp. 169–173. DOI: 10.4250/jcu.2011.19.4.169.

[46] Soojin V Yi and Michael AD Goodisman. "The impact of epigenetic information on genome evolution". In: *Philosophical Transactions of the Royal Society B* 376.1826 (2021), p. 20200114. DOI: 10.1098/rstb.2020.0114.

[47] Kimberly R Kukurba and Stephen B Montgomery. "RNA sequencing and analysis". In: *Cold Spring Harbor Protocols* 2015.11 (2015), pdb–top084970. DOI: 10.1101/pdb.top084970.

[48] Kaumadi Wijesooriya et al. "Urgent need for consistent standards in functional enrichment analysis". In: *PLoS Computational Biology* 18.3 (2022), e1009935. DOI: 10.1371/journal.pcbi.1009935.

[49] Urmi H. Trivedi et al. "Quality control of next-generation sequencing data without a reference". In: *Frontiers in Genetics* 5 (2014). ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00111. URL: https://www.frontiersin.org/articles/10.3389/fgene.2014.00111.

[50] Simon Andrews et al. *FastQC*. Babraham Institute. Babraham, UK, 2010.

[51] Xiangfu Zhong et al. "Accurate adapter information is crucial for reproducibility and reusability in small RNA seq studies". In: *Non-coding RNA* 5.4 (2019), p. 49. DOI: 10.3390/ncrna5040049.

[52] Hongshan Jiang et al. "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads". In: *BMC bioinformatics* 15 (2014), pp. 1–12. DOI: 10.1186/1471-2105-15-182.

[53] Marc Sturm, Christopher Schroeder, and Peter Bauer. "SeqPurge: highly-sensitive adapter trimming for paired-end NGS data". In: *BMC bioinformatics* 17 (2016), pp. 1–7. DOI: 10.1186/s12859-016-1069-7.

[54] Quan Peng et al. "Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes". In: *BMC genomics* 16.1 (2015), pp. 1–12. DOI: 10.1186/s12864-015-1806-8.

[55] Yu Fu et al. "Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers". In: *Bmc Genomics* 19 (2018), pp. 1–14. DOI: 10.1186/s12864-018-4933-1.

[56] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (2016), pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354.

[57] Yang Liao, Gordon K Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (2014), pp. 923–930. DOI: 10.1093/bioinformatics/btt656.

[58] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), pp. 357–359.

[59] CSC - IT Center for Science. *Read counts and alignments -terminology and how to read to results*. 2023. URL: https://chipster.rahtiapp.fi/manual/aligners-comparison.html (visited on 02/12/2023).

[60] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21. DOI: 10.7490/f1000research.1117634.1.

[61] Vivek Bhardwaj et al. *snakepipes*. 2016. URL: https://github.com/maxplanck-ie/snakepipes/issues.

[62] Patro R et al. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nat Methods* 14(4) (2017), pp. 417–419. DOI: doi:10.1038/nmeth.4197.

[63] Avi Srivastava et al. "Alevin efficiently estimates accurate gene abundances from dscRNA-seq data". In: *Genome biology* 20 (2019), pp. 1–16. DOI: 10.1186/s13059-019-1670-y.

[64] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". In: *Cell* (2021). DOI: 10.1016/j.cell.2021.04.048. URL: https://doi.org/10.1016/j.cell.2021.04.048.

[65] Evan D Tarbell and Tao Liu. "HMMRATAC: a Hidden Markov ModeleR for ATAC-seq". In: *Nucleic Acids Research* 47.16 (June 2019), e91–e91. ISSN: 0305-1048. DOI: 10.1093/nar/gkz533. eprint: https://academic.oup.com/nar/article-pdf/47/16/e91/31234555/gkz533.pdf. URL: https://doi.org/10.1093/nar/gkz533.

[66] John Gaspar. *Genrich*. 2018. URL: https://github.com/jsh58/Genrich.

[67] Brent S. Pedersen et al. *Fast and accurate alignment of long bisulfite-seq reads*. 2014. DOI: 10.48550/ARXIV.1401.1129. URL: https://arxiv.org/abs/1401.1129.

[68] Heinig M, Colomé-Tatché M, and Taudt A et al. "histoneHMM: Differential analysis of histone modifications with broad genomic footprints". Version 16. In: *BMC Bioinformatics* (2015). DOI: 10.1186/s12859-015-0491-6.

[69] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (May 2009), pp. 1754–1760. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp324. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/14/1754/605544/btp324.pdf. URL: https://doi.org/10.1093/bioinformatics/btp324.

[70] Fidel Ramírez et al. "High-resolution TADs reveal DNA sequences underlying genome organization in flies". In: *Nature Communications* 9.1 (2018), p. 189. DOI: 10.1038/s41467-017-02525-w. URL: https://doi.org/10.1038/s41467-017-02525-w.

[71] Felix Krueger and Simon R Andrews. "SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes". In: *F1000Research* 5 (2016). DOI: 10.12688/f1000research.9037.2.

[72] A. Anjum et al. "Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach". In: *Journal of computational biology : a journal of computational molecular cell biology* 23(4) (2016), pp. 239–247. DOI: 10.1089/cmb.2015.0205.

[73] R. Rodriguez-Esteban and X. Jiang. "Differential gene expression in disease: a comparison between high-throughput studies and the literature". In: *BMC Med Genomics* 10 (2017). DOI: 10.1186/s12920-017-0293-y.

[74] Tulika Kakati et al. "DEGnext: classification of differentially expressed genes from RNA-seq data using a convolutional neural network with transfer learning". In: *BMC bioinformatics* 23.1 (2022), p. 17. DOI: 10.1186/s12859-021-04527-4.

[75] T Soni Madhulatha. "An overview on clustering methods". In: *IOSR Journal of Engineering* 2(4) (Apr. 2012), pp. 719–725. DOI: 10.48550/arXiv.1205.1117.

[76] Liis Kolberg et al. "gprofiler2– an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler". In: *F1000Research* 9 (ELIXIR).709 (2020). R package version 0.2.1. DOI: 10.12688/f1000research.24956.2.

[77] Pascale Gaudet. "The Gene Ontology". In: *Encyclopedia of Bioinformatics and Computational Biology* 2 (2019), pp. 1–7. DOI: 10.1016/B978-0-12-809633-8.20500-1.

[78] The Gene Ontology Consortium. "The Gene Ontology Resource: 20 years and still GOing strong". In: *Nucleic Acids Res.* 47 (2019). DOI: 10.1093/nar/gky1055.

[79] M Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nat Genet.* (2000). DOI: 10.1038/75556.

[80] Minoru Kanehisa et al. "KEGG as a reference resource for gene and protein annotation". In: *Nucleic acids research* 44.D1 (2016), pp. D457–D462. DOI: 10.1093/nar/gkv1070.

[81] Marc Gillespie et al. "The reactome pathway knowledgebase 2022". In: *Nucleic acids research* 50.D1 (2022), pp. D687–D692. DOI: 10.1093/nar/gkab1028.

[82] Marvin Martens et al. "WikiPathways: connecting communities". In: *Nucleic acids research* 49.D1 (2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.

[83] Hsi-Yuan Huang et al. "miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions". In: *Nucleic acids research* 50.D1 (2022), pp. D222–D230. DOI: 10.1093/nar/gkab1079.

[84] Volker Matys et al. "TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes". In: *Nucleic acids research* 34.suppl_1 (2006), pp. D108–D110. DOI: 10.1093/nar/gkj143.

[85] Mathias Uhlén et al. "Tissue-based map of the human proteome". In: *Science* 347.6220 (2015), p. 1260419. DOI: 10.1126/science.1260419.

[86] Madalina Giurgiu et al. "CORUM: the comprehensive resource of mammalian protein complexes—2019". In: *Nucleic acids research* 47.D1 (2019), pp. D559–D563. DOI: 10.1093/nar/gky973.

[87] Sebastian Köhler et al. "The human phenotype ontology in 2021". In: *Nucleic acids research* 49.D1 (2021), pp. D1207–D1217. DOI: 10.1093/nar/gkaa1043.

[88] Jason P Smith and Nathan C Sheffield. "Analytical approaches for ATAC-seq data analysis". In: *Current protocols in human genetics* 106.1 (2020), e101. DOI: 10.1002/cphg.101.

[89] Claire R Williams et al. "Trimming of sequence reads alters RNA-Seq gene expression estimates". In: *BMC bioinformatics* 17.1 (2016), pp. 1–13. DOI: 10.1186/s12859-016-0956-2.

[90] Felix Krueger et al. July 2021. DOI: 10.5281/zenodo.512789. URL: https://github.com/FelixKrueger/TrimGalore (visited on 01/27/2023).

[91] Knut Reinert et al. "Alignment of next-generation sequencing reads". In: *Annual review of genomics and human genetics* 16 (2015), pp. 133–151. DOI: 10.1146/annurev-genom-090413-025358.

[92] Sagar Chhangawala et al. "The impact of read length on quantification of differentially expressed genes and splice junction detection". In: *Genome biology* 16.1 (2015), pp. 1–10. DOI: 10.1186/s13059-015-0697-y.

[93] Illumina. *BAM File Format*. 2023. URL: https://support.illumina.com/help/BS_App_MDProcessor_Online_1000000007932/Content/Source/Informatics/BAM-Format.htm (visited on 01/26/2023).

[94] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2 (Feb. 2021). giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. URL: https://doi.org/10.1093/gigascience/giab008.

[95] Jianxing Feng et al. "Identifying ChIP-seq enrichment using MACS". In: *Nature protocols* 7.9 (2012), pp. 1728–1740. DOI: 10.1038/nprot.2012.101.

[96] URL: https://app.diagrams.net (visited on 01/26/2023).

[97] M. Ryan Corces et al. "The chromatin accessibility landscape of primary human cancers". In: *Science* 362.6413 (2018), eaav1898. DOI: 10.1126/science.aav1898. eprint: https://www.science.org/doi/pdf/10.1126/science.aav1898. URL: https://www.science.org/doi/abs/10.1126/science.aav1898.

[98] URL: https://www.biostars.org/p/442474/ (visited on 01/26/2023).

[99] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12. DOI: http://dx.doi.org/10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

[100] P. Ferragina and G. Manzini. "Opportunistic Data Structures with Applications". In: FOCS '00. USA: IEEE Computer Society, 2000. ISBN: 978-0769508504. DOI: 10.5555/795666.796543.

[101] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (May 2009), pp. 2078–2079. DOI: doi:10.1093/bioinformatics/btp352.

[102] Heng Li. *Manual page from samtools-1.16 (samtools view)*. 2022. URL: http://www.htslib.org/doc/samtools-view.html (visited on 02/02/2023).

[103] Heng Li. *Manual page from samtools-1.16 (samtools sort)*. 2022. URL: http://www.htslib.org/doc/samtools-sort.html (visited on 02/02/2023).

[104] Heng Li. *Manual page from samtools-1.16 (samtools fixmate)*. 2022. URL: http://www.htslib.org/doc/samtools-fixmate.html (visited on 02/02/2023).

[105] Eric C. Anderson. *Practical Computing and Bioinformatics for Conservation and Evolutionary Genomics*. Bookdown, Jan. 2023. Chap. 19.

[106] Andrew Whitwham. *Manual page from samtools-1.16 (samtools markdup)*. 2022. URL: http://www.htslib.org/doc/samtools-markdup.html (visited on 02/02/2023).

[107] Feng Yan et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis". In: *Genome biology* 21 (2020), pp. 1–16. DOI: https://doi.org/10.1186/s13059-020-1929-3.

[108] Jin Lee et al. "kundajelab/atac_dnase_pipelines: 0.3.0". In: *Zenodo* (Sept. 2016). DOI: 10.5281/zenodo.156534.

[109] Tao Liu. *ATAC-seq settings #145*. 2016. URL: https://github.com/macs3-project/MACS/issues/145 (visited on 02/22/2023).

[110] Dennis Hecker et al. "The adapted Activity-By-Contact model for enhancer–gene assignment and its application to single-cell data". In: *Bioinformatics* 39.2 (Jan. 2023). btad062. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad062. eprint: https://academic.oup.com/bioinformatics/article-pdf/39/2/btad062/49200105/btad062.pdf. URL: https://doi.org/10.1093/bioinformatics/btad062.

[111] Samuel A Lambert et al. "The human transcription factors". In: *Cell* 172.4 (2018), pp. 650–665. DOI: 10.1016/j.cell.2018.01.029.

[112] Stefan Schoenfelder and Peter Fraser. "Long-range enhancer–promoter contacts in gene expression control". In: *Nature Reviews Genetics* 20.8 (2019), pp. 437–455. DOI: 10.1038/s41576-019-0128-0.

[113] Robert C McLeay and Timothy L Bailey. "Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data". In: *BMC bioinformatics* 11.1 (2010), pp. 1–11. DOI: 10.1186/1471-2105-11-165.

[114] Modan K Das and Ho-Kwok Dai. "A survey of DNA motif finding algorithms". In: *BMC bioinformatics* 8.7 (2007), pp. 1–13. DOI: 10.1186/1471-2105-8-S7-S21.

[115] Sven Heinz et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". In: *Molecular cell* 38.4 (2010), pp. 576–589. DOI: 10.1016/j.molcel.2010.05.004.

[116] Felix Krueger and Simon R Andrews. "SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes". In: *F1000Research* 5 (2016). DOI: 10.12688/f1000research.9037.2.

[117] Dr. Renesh Bedre. *MA plot to visualize gene expression data using Python*. 2022. URL: https://www.reneshbedre.com/blog/ma.html (visited on 07/06/2022).

[118] Yusuf Khan et al. "Normalization of gene expression data revisited: the three viewpoints of the transcriptome in human skeletal muscle undergoing load-induced hypertrophy and why they matter". In: *BMC bioinformatics* 23.1 (2022), pp. 1–9. DOI: 10.1186/s12859-022-04791-y.

[119] Hugo Tavares and Georg Zeller. *Exploratory analysis of transcriptomics data in R*. 2021. URL: https://github.com/tavareshugo/data-carpentry-rnaseq (visited on 02/15/2023).

[120] "Analysis of Microarray Data Using Z Score Transformation". In: *The Journal of Molecular Diagnostics* 5.2 (2003), pp. 73–81. ISSN: 1525-1578. DOI: 10.1016/S1525-1578(10)60455-2.

[121] Zuguang Gu, Roland Eils, and Matthias Schlesner. "Complex heatmaps reveal patterns and correlations in multidimensional genomic data". In: *Bioinformatics* (2016). DOI: 10.1093/bioinformatics/btw313.

[122] Ken Phillips. *What Is CIELAB Color Space?* 2023. URL: https://www.hunterlab.com/blog/what-is-cielab-color-space/ (visited on 02/09/2023).

[123] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[124] Yunhai Luo et al. "New developments on the Encyclopedia of DNA Elements (ENCODE) data portal". In: *Nucleic acids research* 48.D1 (2020), pp. D882–D889. DOI: 10.1093/nar/gkz1062.

[125] Hui Xia, Catherine R Dufour, and Vincent Giguère. "ERRα as a bridge between transcription and function: role in liver metabolism and disease". In: *Frontiers in endocrinology* 10 (2019), p. 206. DOI: 10.3389/fendo.2019.00206.

[126] Zaki Shaikhibrahim et al. "ERG is specifically associated with ETS-2 and ETV-4, but not with ETS-1, in prostate cancer". In: *International journal of molecular medicine* 30.5 (2012), pp. 1029–1033. DOI: 10.3892/ijmm.2012.1097.

[127] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[128] Heinrich Taegtmeyer et al. "Assessing cardiac metabolism: a scientific statement from the American Heart Association". In: *Circulation research* 118.10 (2016), pp. 1659–1701. DOI: 10.1161/RES.0000000000000097.

[129] Yeda Sant'Ana Diniz et al. "Diets rich in saturated and polyunsaturated fatty acids: metabolic shifting and cardiac health". In: *Nutrition* 20.2 (2004), pp. 230–234. DOI: 10.1016/j.nut.2003.10.012.

[130] Tsunehisa Yamamoto and Motoaki Sano. "Deranged myocardial fatty acid metabolism in heart failure". In: *International journal of molecular sciences* 23.2 (2022), p. 996. DOI: 10.3390/ijms23020996.

[131] Margaret Anne Craig et al. "Dysregulation of cadherins in the intercalated disc of the spontaneously hypertensive stroke-prone rat". In: *Journal of molecular and cellular cardiology* 48.6 (2010), pp. 1121–1128. DOI: 10.1016/j.yjmcc.2010.01.017.

[132] Igor Kostetskii et al. "Induced deletion of the N-cadherin gene in the heart leads to dissolution of the intercalated disc structure". In: *Circulation research* 96.3 (2005), pp. 346–354. DOI: 10.1161/01.RES.0000156274.72390.2c.

[133] National Center for Biotechnology Information. *PubChem Pathway Summary for Pathway R-HSA-3371497, HSP90 chaperone cycle for SHRs*. 2023. URL: https://pubchem.ncbi.nlm.nih.gov/pathway/Reactome:R-HSA-3371497 (visited on 02/19/2023).

[134] Robert H Oakley and John A Cidlowski. "Glucocorticoid signaling in the heart: a cardiomyocyte perspective". In: *The Journal of steroid biochemistry and molecular biology* 153 (2015), pp. 27–34. DOI: 10.1016/j.jsbmb.2015.03.009.

[135] Sarah McCurdy et al. "Cardiac extracellular matrix remodeling: fibrillar collagens and Secreted Protein Acidic and Rich in Cysteine (SPARC)". In: *Journal of molecular and cellular cardiology* 48.3 (2010), pp. 544–549. DOI: 10.1016/j.yjmcc.2009.06.018.

[136] Ana Catarina Silva et al. "Bearing my heart: the role of extracellular matrix on cardiac development, homeostasis, and injury response". In: *Frontiers in Cell and Developmental Biology* 8 (2021), p. 621644. DOI: 10.3389/fcell.2020.621644.

[137] Petar M Seferović et al. "Heart failure in cardiomyopathies: a position paper from the Heart Failure Association of the European Society of Cardiology". In: *European journal of heart failure* 21.5 (2019), pp. 553–576. DOI: 10.1002/ejhf.1461.

[138] Ken WY Cho. "Enhancers". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 1.4 (2012), pp. 469–478. DOI: 10.1002/wdev.53.

[139] Matteo Cassandri et al. "Zinc-finger proteins in health and disease". In: *Cell death discovery* 3.1 (2017), pp. 1–12. DOI: 10.1038/cddiscovery.2017.71.

[140] Li Xu et al. "LncRNA AK045171 protects the heart from cardiac hypertrophy by regulating the SP1/MG53 signalling pathway". In: *Aging (Albany NY)* 12.4 (2020), p. 3126. DOI: 10.18632/aging.102668.

[141] Jie-Feng Jiang et al. "Role of Sp1 in atherosclerosis". In: *Molecular Biology Reports* 49.10 (2022), pp. 9893–9902. DOI: 10.1007/s11033-022-07516-9.

[142] Eythor Bjornsson et al. "Lifelong reduction in LDL (Low-Density Lipoprotein) cholesterol due to a gain-of-function mutation in LDLR". In: *Circulation: Genomic and Precision Medicine* 14.1 (2021), e003029. DOI: 10.1161/CIRCGEN.120.003029.

[143] Viktoria Kalna et al. "The transcription factor ERG regulates super-enhancers associated with an endothelial-specific gene expression program". In: *Circulation research* 124.9 (2019), pp. 1337–1349. DOI: 10.1161/CIRCRESAHA.118.313788.

[144] Qiu-Yue Lin et al. "Pharmacological blockage of ICAM-1 improves angiotensin II-induced cardiac remodeling by inhibiting adhesion of LFA-1+ monocytes". In: *American Journal of Physiology-Heart and Circulatory Physiology* 317.6 (2019), H1301–H1311. DOI: 10.1152/ajpheart.00566.2019.

[145] Tak W Mak et al. "p53 regulates the cardiac transcriptome". In: *Proceedings of the National Academy of Sciences* 114.9 (2017), pp. 2331–2336. DOI: 10.1073/pnas.1621436114.

[146] M Fischer. "Census and evaluation of p53 target genes". In: *Oncogene* 36.28 (2017), pp. 3943–3956. DOI: 10.1038/onc.2016.502.

[147] Hong Li et al. "p21 protects cardiomyocytes against ischemia-reperfusion injury by inhibiting oxidative stress". In: *Molecular medicine reports* 17.3 (2018), pp. 4665–4671. DOI: 10.3892/mmr.2018.8382.

[148] Sihui Huang et al. "Autophagy is involved in the protective effect of p21 on LPS-induced cardiac dysfunction". In: *Cell Death & Disease* 11.7 (2020), p. 554. DOI: 10.1038/s41419-020-02765-7.

[149] Yang-Fei Tong et al. "Cyclin-dependent kinase inhibitor p21WAF1/CIP1 facilitates the development of cardiac hypertrophy". In: *Cellular Physiology and Biochemistry* 42.4 (2017), pp. 1645–1656. DOI: 10.1159/000479407.

A. Commands and R-scripts to Perform mRNA-seq and ATAC-seq Analysis

A.1. Commands for Quality Control on Console Level

```
fastqc --noextract -o output_path input_path_raw_sequencing_data
```

Listing A.1: Command to Execute FastQC

```
multiqc directory_to_scan -o output_directory -n "
name_of_multiqc_file"
```

Listing A.2: Command to Execute MultiQC

A.2. Commands for snakePipes on Console Level - mRNA-seq

```
createIndices -o index_path --genomeURL http://ftp.ensembl.org/pub
/release-105/fasta/mus_musculus/dna/Mus_musculus.GRCm39.dna_sm.
primary_assembly.fa.gz --gtfURL http://ftp.ensembl.org/pub/release
-105/gtf/mus_musculus/Mus_musculus.GRCm39.105.gtf.gz --blacklist
blacklist.bed --ignoreForNormalization ignore.txt GRCm38_105 --
local
```

Listing A.3: Command to Build Genome Index - snakePipes

```
mRNA-seq -i sample_path -o result_path --local GRCm38_105 --
VCFfile  snps.vcf --strains 'CAST_EiJ,C57BL_6NJ' --mode allelic-
mapping --libraryType 1 --ext .fq.gz --reads '_1' '_2'  GRCm38_105
```

Listing A.4: Command to Execute mRNA-seq Workflow - snakePipes

A.3. R-script Extract for Analysis of Differentially Expressed Genes - mRNA-seq

```
# get matrix from featureCounts and metadata
dds_YB <- DESeqDataSetFromMatrix(countData = featureCounts,
colData <- metadata, design = ~ Genome)
# create DESeq object
dds_YB <- DESeq(dds_YB)
# get results from object
res_YB <- results(dds_YB)
# LFC shrinkage
res_LFC_YB <- lfcShrink(dds_YB,
coef="Genome_genome2_vs_genome1", type="apeglm")
# FDR correction > 0.05
subs_res_YB <- subset(res_YB, res_YB$padj < 0.05)
# write .csv with DEGs
write.csv(subs_res_YB, resultpath)
```

Listing A.5: R-Script: Obtain Differentially Expressed Genes using DESeq2

A.4. Commands for Manual Analysis of ATAC-seq Data on Console Level

```
trim_galore --output_dir path_output --paired sample_1.fq.gz
sample_2.fq.gz > trim_log.out 2> trim_log.err;
bowtie2 -X 1000 -x reference_genome -1 trimmed_sample_1.fq.gz -2
trimmed_sample_2.fq.gz --fr -p 5 2> Bowtie2_summary.txt | samtools
 view -Sb - | samtools sort -n -O bam - | samtools fixmate -O bam
-m - - 2> fixmate_log.log | samtools sort -O bam - | samtools
markdup -O bam -r -s - output_sorted_filtered.bam 2> filtered_log.
log;
macs2 callpeak -t sorted_filtered.bam --outdir path_output -f
BAMPE --keep-dup all -q 0.05 -g mm> macs_log.out 2> macs_log.err
```

Listing A.6: Command to Execute ATAC-seq Workflow

A.5. R-script for Analysis of Differential Accessibility Regions - ATAC-seq

```
library(DiffBind)
# Load sample sheet
samplesheet <- read.csv("path_sample_sheet", header=T,
stringsAsFactors=F, sep=";")
# Create a DBA object
dba_obj <- dba(sampleSheet = samplesheet, peakFormat="narrow")
# Count the number of reads in each peak for each sample
dba_count <- dba.count(dba_obj, summits=75)
# contrast
dba_contrast <- dba.contrast(dba_count, categories=DBA_CONDITION,
minMembers=2)
# Perform differential peak calling
dba_res <- dba.analyze(dba_contrast, method=DBA_DESEQ2)
# extract regions FDR < 0.05
dba_db_significant <- dba.report(dba_res)
# extract all regions
```

```
dba_db <- dba.report(dba_res, th=1, method=DBA_DESEQ2, bCounts=
TRUE)
# Write results to file
write.table(dba_db, file="path_output", sep="\t", row.names=T, col
.names=T)
```

Listing A.7: R-Script: Obtain Differential Accessibility Regions using DiffBind

A.6. Command for Obtaining Enhancer-Gene Interactions and Motif Enrichment Analysis on Console Level

```
STARE_ABCpp -b MACS2.narrowPeak -a Genes.gtf -o output_directory -
n 7 -u DEGs.txt -f chromatin_contact_file
```

Listing A.8: Command to Execute STARE

```
findMotifs.pl enhancer_regions.fa fasta output_dir
```

Listing A.9: Command to Perform Motif Enrichment Analysis - HOMER

A.7. Session Info - R

- R version 4.2.2 (2022-10-31), `x86_64-apple-darwin17.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Running under: `macOS Big Sur 11.0.1`
- Matrix products: default
- LAPACK: `/Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, stats4, utils

- Other packages: apeglm 1.20.0, Biobase 2.58.0, BiocGenerics 0.44.0, biomaRt 2.54.0, ComplexHeatmap 2.14.0, dendextend 1.16.0, DESeq2 1.38.3, DiffBind 3.8.4, dplyr 1.1.0, forcats 1.0.0, GenomeInfoDb 1.34.9, GenomicRanges 1.50.2, ggfortify 0.4.15, ggplot2 3.4.1, gprofiler2 0.2.1, IRanges 2.32.0, MatrixGenerics 1.10.0, matrixStats 0.63.0, purrr 1.0.1, RColorBrewer 1.1-3, readr 2.1.4, S4Vectors 0.36.1, stringr 1.5.0, SummarizedExperiment 1.28.0, tibble 3.1.8, tidyr 1.3.0, tidyverse 2.0.0, VennDetail 1.14.0

- Loaded via a namespace (and not attached): amap 0.8-19, annotate 1.76.0, AnnotationDbi 1.60.0, ashr 2.2-54, assertthat 0.2.1, backports 1.4.1, bbmle 1.0.25, bdsmatrix 1.3-6, BiocFileCache 2.6.1, BiocIO 1.8.0, BiocManager 1.30.19, BiocParallel 1.32.5, Biostrings 2.66.0, bit 4.0.5, bit64 4.0.5, bitops 1.0-7, blob 1.2.3, broom 1.0.3, BSgenome 1.66.3, cachem 1.0.6, caTools 1.18.2, cellranger 1.1.0, circlize 0.4.15, cli 3.6.0, clue 0.3-64, cluster 2.1.4, coda 0.19-4, codetools 0.2-19, colorspace 2.1-0, compiler 4.2.2, crayon 1.5.2, crosstalk 1.2.0, curl 5.0.0, data.table 1.14.8, DBI 1.1.3, dbplyr 2.3.0, DelayedArray 0.24.0, deldir 1.0-6, digest 0.6.31, doParallel 1.0.17, ellipsis 0.3.2, emdbook 1.3.12, fansi 1.0.4, farver 2.1.1, fastmap 1.1.0, filelock 1.0.2, foreach 1.5.2, formatR 1.14, fs 1.6.1, futile.logger 1.4.3, futile.options 1.0.1, gargle 1.3.0, geneplotter 1.76.0, generics 0.1.3, GenomeInfoDbData 1.2.9, GenomicAlignments 1.34.0, GetoptLong 1.0.5, ggrepel 0.9.3, GlobalOptions 0.1.2, glue 1.6.2, googledrive 2.0.0, googlesheets4 1.0.1, gplots 3.1.3, GreyListChIP 1.30.0, gridExtra 2.3, gtable 0.3.1, gtools 3.9.4, haven 2.5.1, hms 1.1.2, htmltools 0.5.4, htmlwidgets 1.6.1, httr 1.4.4, hwriter 1.3.2.1, interp 1.1-3, invgamma 1.1, irlba 2.3.5.1, iterators 1.0.14, jpeg 0.1-10, jsonlite 1.8.4, KEGGREST 1.38.0, KernSmooth 2.23-20, labeling 0.4.2, lambda.r 1.2.4, lattice 0.20-45, latticeExtra 0.6-30, lazyeval 0.2.2, lifecycle 1.0.3, limma 3.54.1, locfit 1.5-9.7, lubridate 1.9.2, magrittr 2.0.3, MASS 7.3-58.2, Matrix 1.5-3, memoise 2.0.1, mixsqp 0.3-48, modelr 0.1.10, munsell 0.5.0, mvtnorm 1.1-3, numDeriv 2016.8-1.1, parallel 4.2.2, pillar 1.8.1,

pkgconfig 2.0.3, plotly 4.10.1, plyr 1.8.8, png 0.1-8, prettyunits 1.1.1, progress 1.2.2, R6 2.5.1, rappdirs 0.3.3, Rcpp 1.0.10, RCurl 1.98-1.10, readxl 1.4.2, reprex 2.0.2, restfulr 0.0.15, rjson 0.2.21, rlang 1.0.6, Rsamtools 2.14.0, RSQLite 2.3.0, rstudioapi 0.14, rtracklayer 1.58.0, rvest 1.0.3, scales 1.2.1, shape 1.4.6, ShortRead 1.56.1, SQUAREM 2021.1, stringi 1.7.12, systemPipeR 2.4.0, tidyselect 1.2.0, timechange 0.2.0, tools 4.2.2, truncnorm 1.0-8, tzdb 0.3.0, UpSetR 1.4.0, utf8 1.2.3, vctrs 0.5.2, VennDiagram 1.7.3, viridis 0.6.2, viridisLite 0.4.1, withr 2.5.0, XML 3.99-0.13, xml2 1.3.3, xtable 1.8-4, XVector 0.38.0, yaml 2.3.7, zlibbioc 1.44.0

B. Readme - Novogene Rawdata Explanation mRNA-seq and ATAC-seq

Raw Sequencing Data

The original fluorescence images obtained from high throughput sequencing platforms are transformed to short reads by base calling. These short reads are recorded in FASTQ format, which contains base information (reads) and corresponding sequencing quality information.

1 Summary of result files

1.1 Files in the folder 'RawData'

- ***.fq.gz ('*' denotes sample ids)**

  Compressed FASTQ file created by gzip. Note: _1.fq.gz and _2.fq.gz contain read1 and read2 for paired-end sequencing, respectively or .fq.gz for single-end sequencing.

- **MD5.txt**

  MD5 hash for the compressed FASTQ files. The MD5 hash can be used to verify the integrity of files. If a file has been changed as a result of a faulty file transfer, its MD5 hash would be changed.

2 Explanation on FASTQ Format

Every read is stored in four lines in FASTQ format as follows:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT
+
@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF
```

Line 1 begins with a '@' character which is followed by a sequence identifier and an optional description.

Line 2 shows the sequenced bases.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the sequencing quality for each base in line 2, and must contain the same number of characters as bases in line 2. The ASCII value of every character minus 33 equals to the phred-scaled quality score of the sequenced base.

Table Illumina sequence identifier datails

EAS139	The unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
2104	Tile number within the flowcell lane
15343	‘x’-coordinate of the cluster within the tile
197393	‘y’-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence

C. Blacklist File - Mus musculus Index

```
chr10 3110060 3110270
chr10 22142530 22142880
chr10 22142830 22143070
chr10 58223870 58224100
chr10 58225260 58225500
chr10 58228320 58228520
chr11 3148660 3148860
chr11 3154960 3155170
chr11 3158530 3158750
chr11 3161780 3161990
chr11 3167020 3167250
chr11 3169390 3169620
chr11 3172450 3172670
chr11 3172950 3173190
chr11 3184190 3185750
chr11 3185700 3186360
chr11 3186330 3189230
chr11 3189190 3190740
chr11 3190750 3191000
chr11 3190960 3194430
chr11 3194400 3195310
chr11 3195240 3197220
```

```
chr11 3197340 3197950
chr11 3197890 3198700
chr11 3198630 3199440
chr11 3199350 3200120
chr11 54139940 54140230
chr11 54140470 54140740
chr11 88967720 88969600
chr11 88969850 88970350
chr11 109011550 109012090
chr12 3109920 3110150
chr12 105436040 105436270
chr13 3372960 3373380
chr13 3373410 3373630
chr13 77438870 77439090
chr13 97190460 97190690
chr13 99790830 99791090
chr13 119488570 119489320
chr13 119597600 119598320
chr13 119599860 119600050
chr13 119601360 119601600
chr13 119601800 119602210
chr13 119602360 119602580
chr13 119609430 119611430
chr13 119612760 119613370
chr13 119613360 119617690
chr14 19415650 19417330
chr14 19417240 19417660
chr14 19417570 19418920
chr14 19418830 19419720
chr14 47454330 47454510
chr15 75085430 75085920
```

```
chr15 75085990 75086240
chr15 75086150 75086550
chr15 75086540 75087110
chr16 11143960 11144170
chr16 57391420 57391740
chr17 13305860 13306280
chr17 13590820 13591650
chr17 13654880 13655120
chr17 36231170 36231390
chr17 39842910 39846780
chr17 39846920 39847160
chr17 39847090 39847310
chr17 39847400 39847720
chr17 39847630 39848880
chr18 3005550 3005770
chr18 3005700 3006050
chr18 12949190 12949400
chr18 40307970 40308340
chr18 68691990 68692230
chr19 45650030 45650310
chr19 61199640 61199880
chr19 61224310 61224530
chr19 61266550 61266760
chr19 61266920 61267210
chr1 24612620 24612850
chr1 48881430 48881690
chr1 58613870 58614090
chr1 78573920 78574140
chr1 88217960 88221950
chr1 88223300 88224760
chr1 133595120 133595340
```

```
chr1 183299040 183299660
chr1 195241610 195241820
chr2 3050030 3050410
chr2 5379200 5379420
chr2 22743580 22743780
chr2 22744760 22744980
chr2 90395030 90395240
chr2 98662130 98663060
chr2 98663540 98664150
chr2 98664780 98665020
chr2 98664970 98665250
chr2 98666140 98667390
chr2 181917260 181917590
chr2 181917550 181917990
chr2 181918970 181919260
chr2 181928340 181928570
chr2 181928950 181929170
chr2 181929220 181929430
chr2 181930800 181931020
chr3 5860530 5860830
chr3 8245690 8245930
chr3 8246280 8246640
chr4 34935690 34935910
chr4 70378040 70378320
chr4 118548460 118548700
chr5 14914900 14915120
chr5 15006590 15006820
chr5 15462500 15462730
chr5 15463060 15463290
chr5 15486990 15487190
chr5 134378920 134379160
```

```
chr5 137152130 137152510
chr5 146260900 146261410
chr6 3201380 3201610
chr6 103648970 103649310
chr7 12010340 12010870
chr8 14306800 14307040
chr8 15519790 15520030
chr8 19711890 19712070
chr9 2999900 3000320
chr9 3000270 3000570
chr9 3000900 3001100
chr9 3001300 3001520
chr9 3004390 3004680
chr9 3004690 3004900
chr9 3005000 3005220
chr9 3005800 3006030
chr9 3006960 3007180
chr9 3008880 3009040
chr9 3015170 3015420
chr9 3015590 3015830
chr9 3016770 3016980
chr9 3017410 3017650
chr9 3018240 3018540
chr9 3018650 3018870
chr9 3019220 3019450
chr9 3021160 3021370
chr9 3021990 3022300
chr9 3024660 3024880
chr9 3025350 3025690
chr9 3026530 3026860
chr9 3027010 3027250
```

```
chr9 3027660 3027880
chr9 3028670 3028880
chr9 3030040 3030330
chr9 3031910 3032130
chr9 3032250 3032560
chr9 3032570 3032790
chr9 3034090 3034300
chr9 3034950 3035160
chr9 3035610 3036180
chr9 3036200 3036480
chr9 3036420 3036660
chr9 3037250 3037460
chr9 3037910 3038120
chr9 3038050 3038300
chr9 24541940 24542200
chr9 35305120 35305620
chr9 110281190 110281400
chr9 123872950 123873160
```

D. Ignore for Normalization File - Mus musculus Index

MT

X

Y

JH584299.1

GL456233.1

JH584301.1

GL456211.1

GL456350.1

JH584293.1

GL456221.1

JH584297.1

JH584296.1

GL456354.1

JH584294.1

JH584298.1

JH584300.1

GL456219.1

GL456210.1

JH584303.1

JH584302.1

GL456212.1

JH584304.1

GL456379.1

GL456216.1

GL456393.1

GL456366.1

GL456367.1

GL456239.1

GL456213.1

GL456383.1

GL456385.1

GL456360.1

GL456378.1

GL456389.1

GL456372.1

GL456370.1

GL456381.1

GL456387.1

GL456390.1

GL456394.1

GL456392.1

GL456382.1

GL456359.1

GL456396.1

GL456368.1

JH584292.1

JH584295.1