Technische Universität München
TUM School of Life Sciences

Discovering candidate regulatory networks in epigenomics and transcriptomics

Markus Daniel Hoffmann

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Mathias Wilhelm

Prüfer*innen der Dissertation:

1. IAS Fellow Lothar Hennighausen, Ph.D.
2. Prof. Dr. Jan Baumbach

Die Dissertation wurde am 18.09.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 08.11.2023 angenommen.

TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM SCHOOL OF LIFE SCIENCES WEIHENSTEPHAN

Doctoral Thesis

Discovering candidate regulatory networks in epigenomics and transcriptomics

Author:	Markus Hoffmann
First Advisor:	Prof. Dr. Lothar Hennighausen
Second Advisor:	Prof. Dr. Jan Baumbach
Mentor:	Dr. Markus List

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to Dr. Markus List, whose mentorship has shaped my journey through this expertise. Your experience, passion, and patience have been instrumental in developing my academic career.

To my esteemed supervisors, Prof. Dr. Jan Baumbach and Prof. Dr. Lothar Hennighausen, thank you for your consistent guidance and support. Your feedback and patience have been invaluable, and I am thankful for your dedication towards my doctoral journey.

Prof. Dr. Priscilla Furth, you have been an advisor in all but on paper. Your advice on research questions, career opportunities, and your influence on my research are deeply appreciated.

I extend thanks to my incredible colleagues and internal collaborators: Martina Rüttger, Dr. Olga Tsoy, Dr. Nina Wenke, Lena Hackl, Michael Hartung, Andreas Meier, Chit Tong Lio, Monica Matchado, Quirin Manz, Judith Bernett, Alexander Dietrich, Mario Picciani, Johannes Kersting, Dr. Olga Lazareva, Dr. Zakaria Louadi, Dr. Melissa Klug, Dr. Fabio Boniolo, Michael Lauber, Dr. Hye-Kyung Lee, Dr. Jakub Jankowski, Octavia Ciora, and Laura Hernández-Lorenzo. Over the past three years, your assistance and cooperation have been indispensable to the success of my work.

I want to extend my sincere appreciation to my own research group, both current and former members. Among them: Nico Trummer, Leon Schwartz, Evelyn Scheibling, Daniel Lochert, Maximilian V. Harl, Christian Hoffmann, Hunor Frey, Julian Poschenrieder, Leon Hafner, Lina-Liv Wilruth, Sylvie Baier, Kevin Yuan, Malte Weyrich, Norman Roggendorf, Shreeti Chhatrala, Johannes Veihelmann, Paul Wissenberg, and Lisa Blankenhagen and many more – your contributions have fostered a thriving intellectual environment and your individual perspectives have significantly enriched my research experience. My journey would not have been the same without the input of my excellent international collaborators - Prof. Dr. David Blumenthal, Prof. Dr. Tim Kacprowski, Prof. Dr. Mauro A. A. Castro, Dr. Bahar Tercan, Dr. A. Gordon Robertson, Prof. Dr. Ilya Shmulevich, Prof. Dr. Katherine A. Hoadley, Massimiliano Incudini, Prof. Dr. Alessandra Di Pierro, Prof. Dr. Kristel Van Steen, among many more. The expertise and unique insights you provided have given my research a global perspective, and for that, I am eternally grateful.

To Christina Trummer, thank you for your creativity and brilliance in designing the logos and web applications for our publications. Your artistic mindset has allowed us to present our work in the most engaging manner possible.

I would like to personally thank my best friends Axel Kühn and Lucy Kuhn for their friendship and support in difficult times. I want to thank Scott Keller for his support in the last months while wrapping up my doctoral thesis.

Lastly, I would like to thank my parents, Evi and Klaus Hoffmann, who allowed me to be where I am right now, doing what I love the most and supporting my decisions.

ABSTRACT

Multicellular organisms, like humans, are composed of a vast number of individual cells. Despite each cell having nearly identical DNA, they differentiate into at least 200 different types in humans, each with specific functions and forms. This differentiation requires various regulatory mechanisms on multiple molecular levels, such as genomics (i.e., the order of the nucleotides), epigenomics (i.e., the structural accessibility of the DNA), and transcriptomics (i.e., the population of RNA transcribed from the DNA). This thesis focuses on the epigenomics and transcriptomics levels.

The first publication aims to examine the role of condition-specific (e.g., disease versus healthy) associations between cis-regulatory elements, transcription factors (TFs), and their resulting target gene expression by utilizing data from the epigenetic and transcriptomic levels. Previous studies have established that the deregulation of such associations could lead to diseases. However, existing methods for studying these changes, like diffTF, are complex, demanding deep technical knowledge, and involve manual steps for validation that cannot be done by scientists with limited computational knowledge. Given the challenges of existing computational methods and the workload of performing multiple Chromatin Immuno-Precipitation DNA-sequencing (ChIP-seq) experiments to experimentally validate hypotheses of such disruptions, I introduce the TF-Prioritizer pipeline that investigates differential TF activity between conditions (e.g., differentially active TFs between health and disease) and could, hence, minimize the number of necessary follow-up TF ChIP-seq experiments. To understand deregulated mechanisms, such as in diseases, this pipeline integrates chromatin profiling data (like histone modification ChIP-seq, ATAC-seq, DNase-seq) with RNA-seq to identify differential TF activity between conditions. Ultimately, with TF-Prioritizer, I provide the first approach that is easy to use for purely wet-lab scientists, provides automated validation, and summarizes results in an interactive web application. In this publication, I show that TF-Prioritizer is capable of capturing important TFs in mammary gland development during pregnancy and lactation.

The second publication focuses on the transcriptomics layer, where I am interested in post-transcriptional gene regulation. Protein-coding transcripts (i.e., messenger RNAs) can be rendered inactive or degraded when micro RNAs (miRNAs; small RNAs consisting of 19-21 nucleotides) bind to a partially complementary miRNA binding site. Since miRNAs can bind to various miRNA binding sites on multiple transcripts, a competing endogenous RNA (ceRNA) regulatory network between all RNAs that harbor miRNA binding sites was proposed. ceRNAs include RNAs such as protein-coding RNAs, long non-coding RNAs, and circular RNAs (circRNAs). circRNAs have been shown to have a higher number of binding sites and could be key to understanding ceRNA networks. However, investigating circRNAs is complicated due to their circular nature and cannot be easily experimentally and computationally identified and investigated. To ease the process on the computational side, to elaborate on the miRNA binding between RNAs that have miRNA binding sites and especially focus on circRNA detection, I present the circRNA-sponging pipeline, which identifies and quantifies circRNAs and miRNA expression, predicts miRNA binding sites, investigates circRNA-miRNA interactions, and identifies potential circRNA biomarkers probably deregulating expression levels of protein-coding RNAs. I report on differentially expressed circRNA in various mouse brain tissues and multiple sponging events of circRNAs. With circRNA-sponging, I present the first end-to-end pipeline that automatically investigates circRNAs in the context of ceRNAs.

In this thesis, I introduce two novel accessible and comprehensive pipelines (i.e., TF-Prioritizer and circRNA-sponging) for investigating the epigenomics and transcriptomics regulatory layers, each including the analysis of experimental data. The pipelines TF-Prioritizer and circRNA-sponging could help researchers to get a more comprehensive view of their data in terms of regulation with little effort.

KURZZUSAMMENFASSUNG

Mehrzellige Organismen, wie der Mensch, bestehen aus einer Vielzahl einzelner Zellen. Obwohl jede Zelle eine nahezu identische DNA besitzt, differenzieren sie sich beim Menschen in mindestens 200 verschiedene Typen, die jeweils spezifische Funktionen und Formen haben. Diese Differenzierung erfordert verschiedene Regulierungsmechanismen auf mehreren molekularen Ebenen, wie z. B. Genomik (d. h. die Anordnung der Nukleotide), Epigenomik (d. h. die strukturelle Zugänglichkeit der DNA) und Transkriptomik (d. h. die Population der von der DNA transkribierten RNA). Diese Arbeit konzentriert sich auf die Ebenen der Epigenomik und Transkriptomik.

Die erste Veröffentlichung zielt darauf ab, die Rolle von zustandsspezifischen (z. B. Krankheit versus Gesundheit) Assoziationen zwischen cis-regulatorischen Elementen, Transkriptionsfaktoren (TFs) und der daraus resultierenden Zielgenexpression zu untersuchen, indem Daten aus der epigenetischen und transkriptomischen Ebene genutzt werden. Frühere Studien haben gezeigt, dass die Deregulierung solcher Zusammenhänge zu Krankheiten führen kann. Bestehende Methoden zur Untersuchung dieser Veränderungen, wie z. B. diffTF, sind jedoch komplex, erfordern fundierte technische Kenntnisse und beinhalten manuelle Schritte zur Validierung, die von Wissenschaftlern mit begrenzten Computerkenntnissen nicht durchgeführt werden können. Angesichts der Herausforderungen bestehender Berechnungsmethoden und des Arbeitsaufwands, der mit der Durchführung mehrerer Chromatin-Immunpräzipitations DNA-Sequenzierungsexperimente (ChIP-seq) verbunden ist, um Hypothesen über derartige Störungen experimentell zu validieren, stelle ich die TF-Prioritizer-Pipeline vor, die unterschiedliche TF-Aktivitäten unter verschiedenen Bedingungen untersucht (z. B. unterschiedlich aktive TFs zwischen Gesundheit und Krankheit) und somit die Anzahl der erforderlichen TF-ChIP-seq-Folgeexperimente minimieren könnte. Um deregulierte Mechanismen, wie z. B. bei Krankheiten, zu verstehen, integriert diese Pipeline Chromatin-Profiling Daten (wie Histon-Modifikation ChIP-seq, ATAC-seq, DNase-seq) mit RNA-seq, um unterschiedliche TF-Aktivitäten zwischen den Bedingungen zu identifizieren. Letztendlich bieten wir mit TF-Prioritizer den ersten Ansatz, der für reine Nasslabor-Wissenschaftler einfach zu handhaben ist, eine automatische Validierung ermöglicht und die Ergebnisse in einer interaktiven Webanwendung zusammenfasst. In dieser Publikation zeige ich, dass TF-Prioritizer in der Lage ist, wichtige TFs in der Brustdrüsenentwicklung während der Schwangerschaft und Laktation zu erfassen.

Die zweite Veröffentlichung konzentriert sich auf die transkriptomische Ebene, wo ich mich für die posttranskriptionelle Genregulation interessieren. Proteinkodierende Transkripte (d. h. Boten-RNAs) können inaktiviert oder abgebaut werden, wenn Mikro-RNAs (miRNAs; kleine RNAs mit 19-21 Nukleotiden) an eine teilweise komplementäre miRNA-Bindungsstelle binden. Da miRNAs an verschiedene miRNA-Bindungsstellen auf mehreren Transkripten binden können, wurde ein regulatorisches Netzwerk konkurrierender endogener RNAs (ceRNAs) zwischen allen RNAs vorgeschlagen, die miRNA-Bindungsstellen beherbergen. ceRNAs umfassen RNAs wie proteinkodierende RNAs, lange nicht-kodierende RNAs und zirkuläre RNAs (circRNAs). circRNAs haben nachweislich eine größere Anzahl von Bindungsstellen und könnten der Schlüssel zum Verständnis von ceRNA-Netzwerken sein. Die Untersuchung zirkulärer RNAs ist jedoch aufgrund ihrer zirkulären Natur kompliziert und kann nicht einfach experimentell und rechnerisch identifiziert und untersucht werden. Um den Prozess auf der rechnerischen Seite zu vereinfachen und die miRNA-Bindung zwischen RNAs, die miRNA-Bindungsstellen haben, genauer zu untersuchen, stelle ich die circRNA-Sponging-Pipeline vor, die circRNAs und miRNA-Expression identifiziert und quantifiziert, miRNA-Bindungsstellen vorhersagt, circRNA-miRNA-Interaktionen untersucht und potenzielle circRNA-Biomarker identifiziert, die vermutlich die Expressionsniveaus von proteinkodierenden RNAs deregulieren. Ich berichte über unterschiedlich exprimierte circRNA in verschiedenen Mäusegehirngeweben und über mehrere Sponging-Ereignisse von circRNAs. Mit circRNA-sponging präsentiere ich die erste End-to-End-Pipeline, die circRNAs automatisch im Kontext von ceRNAs untersucht.

In dieser Arbeit stelle ich zwei neuartige, leicht zugängliche und umfassende Pipelines (d.h. TF-Prioritizer und circRNA-sponging) für die Untersuchung der regulatorischen Ebenen der Epigenomik und Transkriptomik vor, die jeweils auch die Analyse experimenteller Daten umfassen. Die Pipelines TF-Prioritizer und circRNA-sponging könnten Forschern helfen, mit geringem Aufwand einen umfassenderen Überblick über ihre Daten im Hinblick auf die Regulierung zu erhalten.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EQUATIONS

LIST OF FIGURES

ABBREVIATIONS

Abbreviation	Term
3-D	three dimensional
A	Adenine
AAA	Lysine (template)
AAG	Lysine (template)
API	Application Programming Interface
ASO	antisense oligonucleotide
ATAC-seq	Assay for Transposase-Accessible Chromatin using Sequencing
AUG	start codon
BAM	binary alignment map
bps	base pairs
BSJ	back-splicing junction
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler Transformation
C	Cytosine
cDNA	reverse-transcribed complementary DNA originating from RNA
ceRNA	competing endogenous RNA
ChIP-seq	chromatin ImmunoPrecipitation DNA-sequencing
circRNA	circular RNA
ciRNA	circular intronic
CLR	circular-to-linear ratio
CRE	cis-regulatory elements
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
DCG	discounted cumulative gain
ddNTP	dideoxy nucleotide (N can be substituted by A, T, G, C)
DNA	deoxyribonucleic acid
dNTP	deoxy nucleotide
dsRNA	double-stranded RNA
ecirc	exonic circular
ElciRNA	exon-intron circRNA
ENCODE	Encyclopedia of DNA Elements
ER	estrogen receptor
FDR	false discovery rate

FN	false negatives
FP	false positives
FPR	false positive rate
G	Guanine
GRN	gene regulatory network
HINT-ATAC	Heuristic Identification of Nucleosome and Transcription factor footprints from ATAC-seq
HM	histone modification
IGV	integrated genome viewer
K	Lysine (one letter code)
lncRNA	long non-coding RNA
log2fc	log2 fold change
MACS2	Model-based Analysis of ChIP-seq version 2
miRNA	micro RNA
miRNA-seq	miRNA sequencing
mRNA	messenger RNA
NCA	Network Component Analysis
NGS	next-generation sequencing
ORF	open reading frame
PCR	polymerase chain reaction
poly-A	polyadenylation
pre-miRNA	precursor miRNA
pre-mRNA	precursor mRNA
pri-miRNA	primary miRNA
PSI	Percent Spliced In
PWM	position weight matrices
RBP	RNA-binding protein
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
RT-qPCR	quantitative PCR with reverse transcription
SAM	sequence alignment map
shRNA	short hairpin RNA
siRNA	small interfering RNA
SNP	single nucleotide polymorphism
STAR	spliced transcripts alignment to a reference
T	Thymine

TAD	Topologically Associating Domain
TF	transcription factor
TFA	TF activity
TFBS	transcription factor binding site
TN	true negatives
total RNA-seq	total RNA sequencing
TP	true positives
TPM	transcript per million
TPR	true positive rate
U	Uracil
UAA	stop codon
UAG	stop codon

1. GENERAL INTRODUCTION

1.1. Motivation

Every multi-cellular organism is composed of a multitude of individual and identifiable cells [1] that commonly have a cell nucleus, a protected compartment separated from the rest of the cell, where the heritable material (i.e., deoxyribonucleic acid (DNA)) is stored. The human body consists of approximately 37.2 trillion cells [2] that can be separated into at least 200 general cell types [3,4]. Each cell type has its dedicated function and morphology despite possessing a nearly identical copy of the DNA [1,5]. Hence, cells must have regulation mechanisms on various molecular levels to achieve differentiation into cell types [6,7].

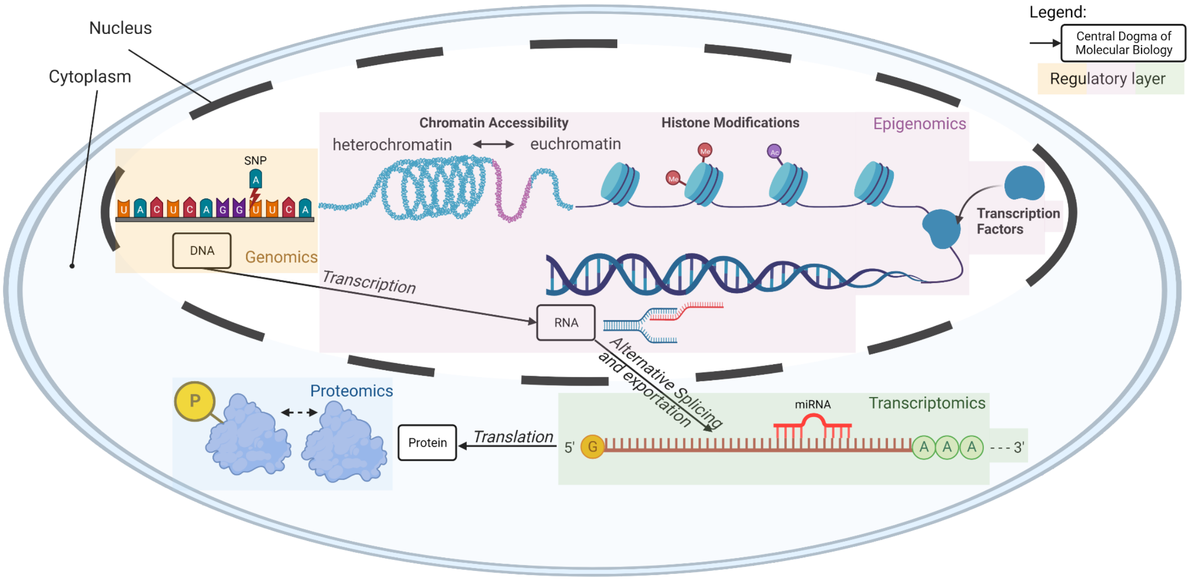


Figure 1: ***The regulatory layers of the central dogma of molecular biology for a eukaryotic cell.*** *Information from the DNA is read and transcribed into RNA (e.g., messenger RNA (mRNA) for protein-coding genes), and the mRNA is then translated into protein. Regulation can happen on the genomics level (e.g., single nucleotide polymorphisms (SNPs)), on the epigenomics level (e.g., structural alterations of the DNA), and on the transcriptomics level (e.g., transcript silencing), which are described in the central dogma of molecular biology. However, regulation can also happen in post-translational modifications (e.g., phosphorylation) that can be seen in proteomics data or on the metabolomics layer, which are both not captured by the central dogma of molecular biology. This Figure was created using Biorender.com.*

Until today, several molecular levels on which regulation can happen are known - they are partly described by the central dogma of molecular biology (see Figure 1). The definition, in its most abstract way, of the central dogma of molecular biology depicts that information from the DNA is read and transcribed into ribonucleic acid (RNA; e.g., messenger RNA (mRNA) for most protein-coding genes), and the protein-coding RNA is then translated into protein [8]. Hence, the central dogma of molecular biology covers the regulatory layers of (i) genomics (i.e., inside the cell nucleus, alterations of the DNA sequence via single nucleotide polymorphism (SNPs), sequence insertions, or deletions), (ii) epigenomics (i.e., inside the nucleus, e.g., transcriptional regulation from DNA to RNA), (iii) and transcriptomics (i.e., mostly outside of the nucleus, e.g., silencing of transcripts so they cannot be translated into proteins anymore). However, it lacks regulatory layers about possible regulations such as on the (iv) proteomics level (i.e., mostly outside of the nucleus, e.g., phosphorylation - a form of post-translational modification) and (v) metabolomics level (i.e., pathway regulation). Despite all regulatory layers being of interest, I present two bioinformatic pipelines to analyze molecular profiling data on the epigenomics and transcriptomics layers.

Epigenomics describes changes in DNA organization (e.g., DNA accessibility, chromatin organization, or histone modifications) without altering the DNA sequence itself. In eukaryotes, gene expression on an epigenomics level is, among other mechanisms, mainly controlled by cis-regulatory elements (CREs) such as promoters, enhancers, or suppressors, which are bound by transcription factors (TFs) promoting or repressing transcriptional activity depending on their accessibility and availability [9]. TFs play an important role not only in development and physiology but also in diseases (e.g., it is known that at least a third of all known human developmental disorders are associated with deregulated TF activity and mutations [10–12]). An in-depth investigation of TF regulation could help to gain deeper insight into the gene-regulatory balance found in healthy cells. Since most complex diseases involve aberrant gene regulation, a detailed understanding of this mechanism is a prerequisite to developing targeted therapies [13,14]. This is a daunting task, as multiple genes in eukaryotic genomes may affect the disease, each of which is controlled by possibly various CREs. Chromatin Immuno-Precipitation DNA-sequencing (ChIP-seq) experiments using TF-specific antibodies are the gold standard for identifying and understanding condition-specific TF-binding on a nucleotide level (see Sec. 2.2.1) [15]. However, since there are approximately 1,500 active TFs in humans [16] and about 1,000 in mice [17] and additionally considering the need to establish TF patterns separately for each tissue and each physiological condition, the application of all TFs to this approach is prohibitive. Alternatively, histone modification (HM) ChIP-seq offers a broader view of the chromatin due to its capability to, e.g., highlight open chromatin regions where gene expression can take place, hence allowing us to identify locations of condition-specific CREs [18]. Computational methods can then be used to prioritize TFs likely binding to these CREs, leading to hypotheses and, e.g., informing us which TF ChIP-seq experiments are the most promising to perform [19]; however, high-quality antibodies for TF ChIP-seq (see Sec. 2.2.1 “The protocols”) are only available for a small number of TFs. Hence, another experimental strategy that could be informed by these generated hypotheses is Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), a genome editing tool, adapted from a bacterial defense mechanism against viruses. It employs an enzyme, typically Cas9, to cut DNA, facilitating genetic modifications precisely. CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa) are specialized derivatives of CRISPR designed to regulate gene expression instead of editing it. While CRISPRi represses, CRISPRa activates genes. Both can induce epigenetic changes. This shows that computationally generated hypotheses can narrow down the scope of experiments needed to confirm working hypotheses about regulation [20–22] on the epigenomics level.

In the transcriptomics layer, several molecular mechanisms to silence transcripts without changing the genomics and epigenomics layers are known. Transcripts (e.g., mRNAs) that are translated into proteins can be controlled by prohibiting the translation. This may, for instance, be caused by the degradation of the mRNA (e.g., by the deadenylation-dependent pathway [23] or micro RNAs (miRNAs) [24]) or via silencing (e.g., by miRNAs or small interfering RNA (siRNAs)) [25]. In recent years, miRNAs, small non-coding RNAs of a length of 19-25 nucleotides [26], have attracted more and more attention in transcript regulation [27,28]. They are involved in many biological processes and human diseases [29]. miRNAs regulate their target RNA transcripts by either degrading them or preventing their translation by binding to miRNA binding sites and rendering them inactive [24,30]. miRNA binding sites can also be found on other transcripts than mRNAs, e.g., long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), transcripts of 3' untranslated regions (UTRs), and pseudogenes [31]. Hence, all of these RNAs are in competition [31], giving rise to a large regulatory network between them. In their work, Salmena et al. defined competing endogenous RNAs (ceRNAs) as RNAs that carry miRNA binding sites [32]. These RNAs compete for the available miRNAs in a cell. As a result, an overexpressed ceRNA can sponge away miRNAs (i.e., bind many miRNAs until few or none are left) required for the regulation of other RNAs, which might ultimately lead to disease. Lately, circRNAs have particularly attracted attention as key ceRNAs [33–35]. circRNAs are classified as lncRNAs despite a few being known to encode proteins [36] and could therefore function ambivalent as lncRNAs and protein-coding RNAs.

circRNAs are characterized by their loop structure [37,38]. The biogenesis of circRNAs is explained by the occurrence of a back-splicing event during the alternative splicing process of precursor messenger RNA (pre-mRNA), meaning that the 5' terminus of an upstream exon and the 3' terminus of a downstream exon are covalently joined (see Sec. 2.1.4) [37]. What differentiates them from linear RNAs is the lack of a 5' cap and a 3' polyadenylation (poly(A)) tail [39–42]. circRNAs can be made up of exonic and intronic regions of their spliced pre-mRNA and are thus found in a huge variety of sizes, ranging from under 100 to more than 4,000 nucleotides [39,43]. Some are conserved across species, and their expression is tissue- and disease-specific [38,44,45]. As a result, they could play an important role in health, acting as potential biomarkers for pathological conditions and therapeutic targets [45–47]. The enhanced stability of circRNAs might allow them to work as buffers for miRNAs by binding them until they outnumber the circRNA binding sites [45]. The regulatory function of circRNAs and their alleged association with diseases are the main reasons why identifying sponging activity between circRNAs and miRNAs is of particular interest. The presence of an interaction between miRNAs and circRNAs has been repeatedly proven, and several circRNAs (e.g., CDR1as/CiRS-7, SRY [48], and circNCX1 [49]) have been recognized as miRNA sponges. Even though individual studies confirmed the existence of circRNA sponges, the scientific community still lacks knowledge and biological understanding in this research area. From a computational point of view, the detection of circRNAs is difficult due to their circular shape and the lack of poly(A) tail, which makes observing them in poly(A)-enriched RNA sequencing (RNA-seq) libraries unlikely [43]. Hence, circRNAs can only be detected in libraries without poly(A) enrichment, such as ribosomal RNA (rRNA) depleted RNA-seq and total RNA-sequencing (totalRNA-seq), which do not cause depletion of circRNAs [43]. The key to identifying circRNAs from sequencing data lies in the reads that do not map to linear transcripts. Unmapped reads whose mapping can be explained by backsplicing junctions (BSJs) rather than by linear isoforms indicate the occurrence of back-splicing events and, consequently, potential circRNAs (see Supplementary Figure 1). The identification and estimation of circRNA abundance can only be estimated through BSJ, which is problematic as the more realistic expression of circRNAs is considered much higher [50]. circRNAs and their potential role as miRNA sponges could help us to understand regulation on the transcriptomics regulation layer.

1.2. Aim of the thesis

During the course of this thesis, I aimed to develop user-friendly computational pipelines on the epigenomics and transcriptomics regulatory layer that enable researchers with little or no computational experience to analyze their data and generate hypotheses to narrow down potential follow-up experiments.

On the epigenomics layer, many of the previously deployed methods (e.g., diffTF [19]) necessitate significant preprocessing, a deep understanding of computation, tailoring the technique to different scenarios (like handling more than two conditions or time-series data), and hands-on assessment of the outcomes (for instance, manually searching and visualizing high-quality TF ChIP-seq data to validate predictions). To streamline this task, I introduce TF-Prioritizer [51], a tool that combines RNA-seq and open chromatin data to identify condition-specific TF activity (i.e., TFs that behave differentially in activity between, e.g., the healthy versus disease state). TF-Prioritizer is built on several existing state-of-the-art tools for peak calling, TF-affinity analysis, differential gene expression analysis, and machine learning tools. TF-Prioritizer is the first to jointly consider multiple types of modalities (e.g., different histone marks and/or time series data), provide a joint list of active TFs, and enable the user to see a visualized validation of the predictions in an interactive and feature-rich web application [51].

At the transcriptomics level, a plethora of tools for analyzing circRNA functions were published [52]. E.g., some to identify circRNA activity in ceRNA networks depicting the interplay between circRNAs and entities like miRNAs, lncRNAs, or mRNAs [52]. Other tools were designed for circRNA downstream analysis to focus on tasks such as detecting alternative splicing and predicting and visualizing circRNA structure and assembly [52]. However, in most cases, individually, each of the tools requires intense preprocessing of the data, and analysis could not be performed automatically. To the best of our knowledge, no existing pipeline offers a thorough, automated analysis of circRNA-sponging that combines both circRNA and miRNA detection and quantification, an in-depth exploration of potential circRNA–miRNA sponging interactions, and a ceRNA network examination. Hence, in the scope of this thesis, I developed circRNA-sponging [28,53,54], a nextflow pipeline integrating several state-of-the-art methods to "(1) detect circRNAs via identifying BSJ from totalRNA-seq data, (2) quantify their expression values to a realistic value, (3) perform a differential expression analysis, (4) identify and quantify miRNA expression from miRNA-sequencing (miRNA-seq) data, (5) predict miRNA binding sites on circRNAs, (6) systematically investigate potential circRNA-miRNA sponging events, (7) create a ceRNA network, and (8) identify potential circRNA biomarkers using the ceRNA network" [28,53,54].

1.3. Outline

In the background chapter, I present the essential concepts of molecular biology (Sec. 2.1), how the data is retrieved in wet-bench experiments/protocols, describe the generated data types, how the data is processed with various bioinformatics tools, and their possible interpretation (Sec. 2.2). I further elaborate on the basics in machine learning (Sec. 2.3) and network medicine (Sec. 2.4).

The discussed topics in the background chapter provide the necessary knowledge for the Methods chapter (Sec. 3), where I discuss the current state of research on regulatory mechanisms in epigenomics (Sec. 3.1.1) and circRNAs (Sec. 3.2.1). I further describe methods utilized in the algorithmic frameworks of TF-Prioritizer (Sec. 3.1.2) and the circRNA-sponging pipeline (Sec. 3.2.2).

In Sec. 4, I provide summaries of the two publications (TF-Prioritizer in Sec. 4.1 and the circRNA-sponging pipeline in Sec. 4.2), incorporated in this thesis, and I precisely describe the contributions of the author.

In Sec. 5, I first generally discuss the problems of the interplay between clinical, wet-lab, and bioinformatics. In Sec. 5.1, I then discuss the limitations of the TF-Prioritizer pipeline, what I intend to do in the future to address them, and how I plan to extend TF-Priotizer to cover three OMICS layers (genomics, epigenomics, and transcriptomics). In Sec. 5.2, I discuss the limitations of the circRNA-sponging pipeline, how I plan to address them in the future, and potential experimental methods to validate computationally generated hypotheses. Lastly, Sec. 5.3 gives a general conclusion of the thesis.

2. BACKGROUND

2.1. Gene regulation

In this chapter, I first introduce the basic molecules of biology (Sec. 2.1.1) and their roles in the concept of the central dogma of molecular biology (Sec. 2.1.2). Next, I explain the transcriptional process and its regulation in more detail (Sec. 2.1.3). Lastly, I elaborate possible regulation at the post-transcriptional state (Sec. 2.1.4).

2.1.1. The basic molecules of biology

Deoxyribonucleic acid (DNA)

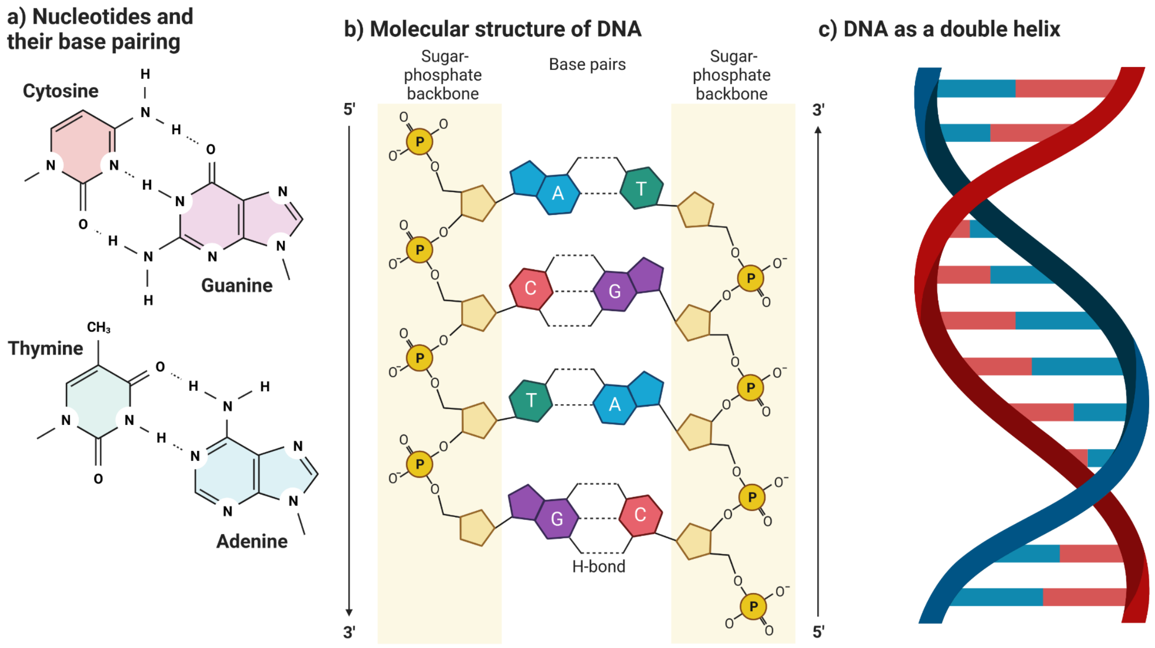


Figure 2: ***The nucleotides, their base pairing, the molecular structure of the DNA, and the form of the DNA as a double helix. (a)*** *The basic building blocks of the DNA are Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). They can form base pairs (A-T, G-C), and* ***(b)*** *are chained with the DNA backbone using sugar and phosphate bondings.* ***(c)*** *The backbone forms a double helix. This Figure was created using Biorender.com.*

In chemistry, macromolecules are larger molecules (i.e., diameter approx. from 100 to 10,000 Å - 10^{-5} to 10^{-3} mm, more than 1000 atoms) compared to ordinary molecules (i.e., a diameter of less than 10 Å - 10^{-6} mm, less than 1000 atoms) [55,56]. DNA, the carrier of the genetic information of an organism, is considered a macromolecule. DNA consists of basic building blocks that are called nucleotides. A nucleotide is composed of a sugar (i.e., monosaccharide 2-deoxyribose), a phosphate group, and a nitrogen that contains one of four bases (i.e., Adenine (A), Thymine (T), Guanine (G), and Cytosine (C), see Figure 2. a-b) [57]. Two nucleotides can form hydrogen bonds. Cytosine can form a bond with Guanine with three hydrogen bondings (stronger bond), and Adenine can form a bond with Thymine with two hydrogen bondings (weaker bond, see Figure 2. a) [57]. If two nucleotides are bonded they are called a base pair (bp) [57]. DNA consists of millions or even billions of base pairs (bps) (e.g., 3.2 billion bps in the human genome [58]). Each base pair is connected to another base pair by an alternating sequence of bonds between sugars (i.e., monosaccharides) and phosphate molecules that altogether form a chain called DNA backbone [57]. The backbone can be separated into two antiparallel DNA strands, each containing one part of the bps [59]. Due to DNA's chemical properties, it forms a double helix structure (see Figure 2. c) [59] and is organized to be readable and accessible (see Sec. 2.1.3).

Ribonucleic acid (RNA)

The macromolecule RNA is a chain of ribonucleotides (i.e., exchange of deoxyribose sugar (DNA) to ribose sugar (RNA) [60]). In most cases, it is a complementary copy of the DNA [61], where the base Thymine (T) is replaced with the base Uracil (U). Thymine differs from Uracil by one additional methyl group [62]. The replacement of Uracil in DNA is caused by the chemical instability of Cytosine (i.e., the frequent process of deamination of Cytosine leads to the mispairing of bases) [63,64]. There is a plethora of RNA subclasses that were observed, discovered, and experimentally validated between the 1960s and the early 2000s (see Table 1). In this thesis, I focus on mRNAs, circRNAs, lncRNAs, and miRNAs. They are described in detail below (mRNAs - Sec. 2.1.2, circRNAs, lncRNAs, and miRNAs - Sec. 2.1.3).

RNA subclass	Year discovered	References
mRNA	1961	[65,66]
circRNA	1976	[67]
lncRNA	1991	[68]
miRNA	1993	[69]
siRNA	1998	[70]

Table 1: ***Discovery of each RNA subclass by year.***

Proteins

Proteins are complex, multifunctional macromolecules that play various important roles in organisms by being the building blocks of basic organic components (e.g., channel proteins to import and export products of a single cell [71]). Proteins consist of long chains of amino acids, organic compounds containing carboxyl and amino groups [72]. There are 20 different amino acids in most organisms that can be incorporated into proteins (see Sec. 2.1.2), and the specific sequence of these amino acids determines the three-dimensional structure and function of the protein [72].

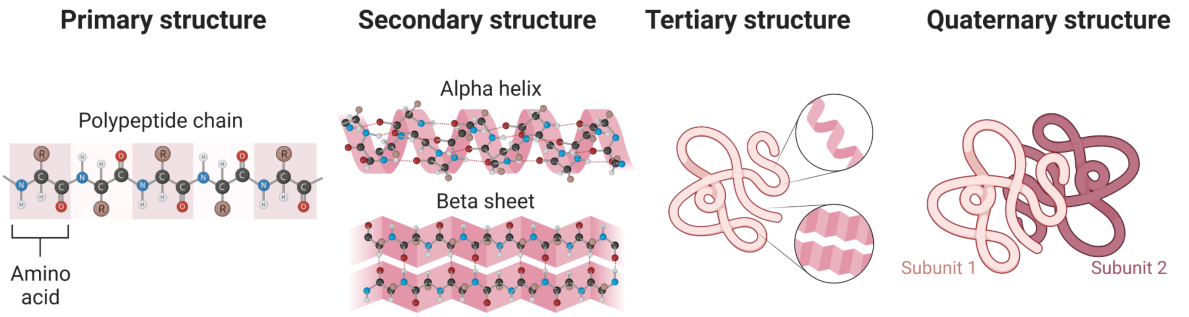


Figure 3: ***Protein structure levels.*** *The primary structure depicts the order of the amino acid sequence. The secondary structure refers to the three-dimensional (3-D) local structure (e.g., alpha helix, beta sheet). The tertiary gives information about the overall 3-D shape consisting of multiple alpha helices and beta sheets. The quaternary structure describes the subunits of the protein. Each subunit has its own primary, secondary, and tertiary structure. The figure was created with BioRender.com.*

The description of the structure of proteins can be separated into four different levels. (i) the primary structure, (ii) the secondary structure, (iii) the tertiary structure, and (iv) the quaternary structure (see Figure 3). The primary structure describes the order of the sequence of the amino acids. The secondary structure gives information about the local three-dimensional (3-D) structure of the components of the protein (e.g., alpha helix, beta sheet). Alpha helices are helical structures formed by hydrogen bonding between the backbone atoms of the protein, while beta sheets are flat structures formed by hydrogen bonding between the side chains of the protein [57]. The tertiary structure refers to the overall 3-D shape of the protein consisting of one or several alpha helices and beta sheets [73]. The quaternary structure depicts subunits of the proteins (e.g., more dense, less dense connected parts of the protein). Each of its subunits has its own primary, secondary, and tertiary structure [74]. The structure of the protein determines its function.

Proteins perform a vast collection of functions in the cell and the whole organism, including catalyzing a large number of chemical reactions (in this function, they are called enzymes [75]), replicating DNA [76], transporting molecules from one location to another (either inside the cell or between cells) [77], and building important structural components of cells and tissues [78] amongst many more [79]. Proteins are built inside cells through a process called protein synthesis (Sec. 2.1.2).

Comparison of DNA, RNA, and Proteins

This subsection is dedicated to summarizing the properties of DNA, RNA, and proteins (see Table 2). DNA and RNA both encode genetic information, while proteins are built from genetic information from RNA. However, until now, no mechanism is known where protein can be reverse-translated into RNA or DNA and, therefore, until today, primarily does not encode for reproducible genetic information [80]. DNA itself cannot perform any reaction. RNA (then called ribozymes) and proteins perform interactions with other cell components [81,82]. DNA and RNA are each composed of four different types of nucleotides, while proteins in most organisms are composed of 20 different amino acids [83]. DNA is more stable due to its repair systems (i.e., despite the repair system being prone to error [84]) and is less prone to degradation compared to RNA and proteins [83]. Overall, one could conclude that RNA is a state between DNA and protein.

	DNA	RNA	Proteins
Encodes genetic information	Yes	Yes	Currently no mechanisms known to encode and decode
Catalyzes biological reactions	No	Yes	Yes
Type of building blocks	Nucleotides	Nucleotides	Amino acids
Number of building blocks	4	4	20 (in most organisms)
Structure	Double helix	Complex	Complex
Stability to degradation	High	Variable	Variable
Repair systems	Yes	Mostly no	No

Table 2: ***Comparison of properties of DNA, RNA, and Proteins.*** *The Table was adopted from https://en.wikipedia.org/wiki/Macromolecule under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA) and the GNU Free Documentation License (GFDL).*

2.1.2. The processes of the central dogma of molecular biology

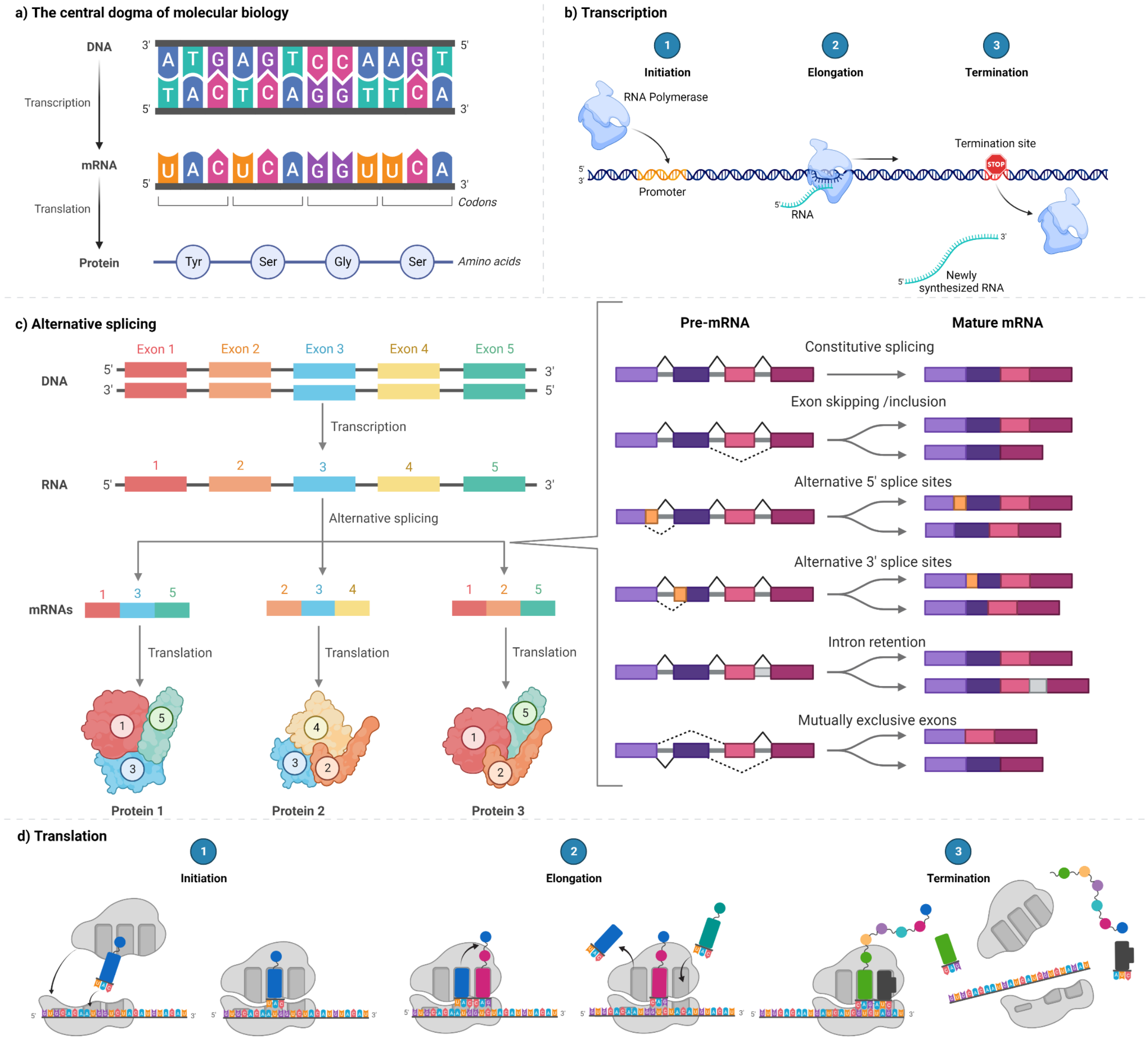
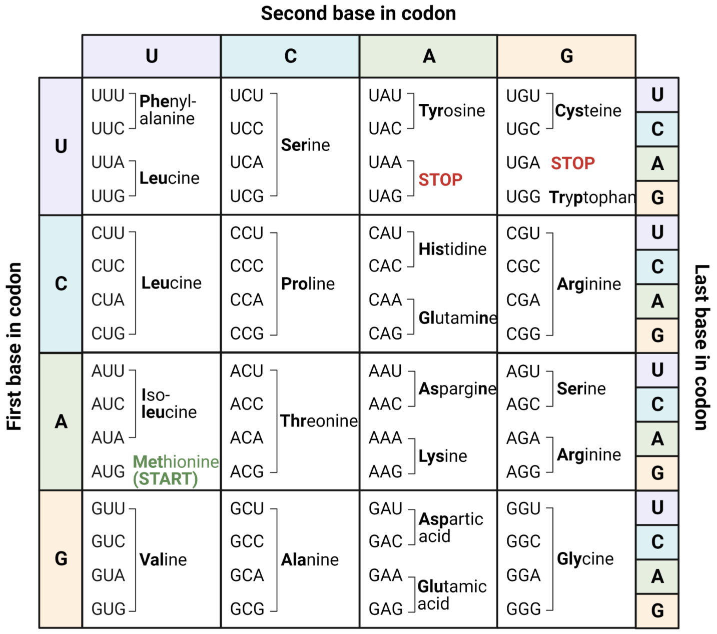


Figure 4: ***The central dogma of molecular biology and the processes of transcription and translation. (a)*** *The central dogma with the information flow from DNA to RNA (transcription) and RNA to protein (translation).* ***(b)*** *The transcriptional process with the stages of initiation, elongation, and termination.* ***(c)*** *Process and variants of alternative splicing.* ***(d)*** *The stages of the translational process. This Figure was created with Biorender.com.*

The central dogma of molecular biology is a concept first proposed by Crick in 1957 and explains the flow of genetic information within living organisms [8]. It describes how information is transferred from DNA to RNA to proteins (see Figure 4. a) [8]. The dogma can help to explain how the genetic information in DNA is used to build and maintain an organism and how changes in DNA can lead to changes in the structure and function of RNA and proteins [8]. The DNA is transcribed to RNA (transcription - see below) and then translated into proteins (translation - see below) [8].

Transcription is the information flow from DNA to RNA (see Figure 4. b). The transcriptional process can be partitioned into (i) initiation, (ii) elongation, and (iii) termination. During the initiation process, RNA-Polymerases, a class of large enzyme complexes that are made up of multiple subunits (see Figure 4. b), play the central part. RNA polymerases build RNA from DNA. Several types of RNA polymerases are known, with the most central one in human being RNA polymerase 2, which produces the mRNA during transcription. Hence, RNA polymerase 2 interacts with a regulatory element (i.e., a promotor - for an explanation, see Sec. 2.1.3) close to an open reading frame (ORF, i.e., a gene on the DNA that is about to be transcribed into RNA). During elongation, RNA polymerase 2 then transcribes the gene nucleotide by nucleotide into RNA. Lastly, the RNA-Polymerase reaches an area behind the ORF - called the untranslated region (UTR) - and terminates the transcription process [83], creating the new mRNA.



Second base in codon

First base in codon	U	C	A	G	Last base in codon
U	UUU **Phen**ylalanine	UCU **Ser**ine	UAU **Tyr**osine	UGU **Cys**teine	U
	UUC	UCC	UAC	UGC	C
	UUA **Leu**cine	UCA	UAA STOP	UGA STOP	A
	UUG	UCG	UAG	UGG **Trp**tophan	G
C	CUU **Leu**cine	CCU **Pro**line	CAU **His**tidine	CGU **Arg**inine	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA **Glu**tamine	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU **Iso**-**leu**cine	ACU **Thr**eonine	AAU **Asp**argine	AGU **Ser**ine	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA **Lys**ine	AGA **Arg**inine	A
	AUG **Met**hionine (START)	ACG	AAG	AGG	G
G	GUU **Val**ine	GCU **Ala**nine	GAU **Asp**artic acid	GGU **Gly**cine	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA **Glu**tamic acid	GGA	A
	GUG	GCG	GAG	GGG	G

Figure 5: ***Degenerated code for the codon triplets of the twenty amino acids.*** *Allocation system of amino acids to RNA triplet for most organisms. The bold letters represent the three-letter code of the amino acid. This Figure was created with Biorender.com.*

messenger RNA (mRNA) is a type of RNA that is the intermediate step between DNA and protein and carries genetic information [85,86]. mRNA exists in two forms: pre-mRNA and mature-mRNA. The pre-mRNA in most multicellular organisms is composed of an intron-exon structure. Introns are most frequently spliced from the pre-mRNA during alternative splicing (see below), and mature-mRNAs mostly lack introns. The mature-mRNA is exported from the nucleus to the cytoplasm (see Figure 1). The process of splicing out introns and sometimes also exons at splice sites (i.e., at a motif of a nucleotide position on a pre-mRNA) in several ways (see below) from the pre-mRNA is called alternative splicing (see Figure 4. c). Using alternative splicing (see below), one pre-mRNA can produce several different mature-mRNAs and, therefore, code for several different proteins then referred to as protein isoforms (see Figure 4. c, left side). For the remainder of this thesis, I refer to mature-mRNA as mRNA.

There are several alternative splicing types (see Figure 4. c, right side), including (i) constitutive splicing (i.e., splicing out all introns without any variation), (ii) exon skipping (i.e., splicing out exons from the pre-mRNA), (iii)/(iv) alternative 5’/3’ splice sites (i.e., the intron-exon structure has alternative splice sites on the 5'/3' direction that can be used to splice out introns), (v) intron retention (i.e., an intron is not spliced out of the mRNA), and (vi) mutually exclusive exons (i.e., exons that are never seen together in the same mRNA) [85,87]. circRNAs can also undergo the alternative splicing process (see Sec. 3.2.2 “Alternative splicing analysis of circRNAs using SUPPA2”).

Once an mRNA leaves the nucleus, ribosomes - the key players during translation - bind to the mRNA and start the translation process at the start codon (AUG). Ribosomes are ribozyme complexes composed of ribosomal RNA (rRNA, transcribed by RNA polymerase 1) and proteins. Ribosomes are composed of a large and a small subunit (see Figure 4. d). The small subunit reads triplets (= codons) of the mRNA that correspond to one of twenty amino acids used in most organisms (e.g., AAA and AAG codes for Lysine, see Figure 5 for full codon to amino acid translation). As codons can code for more than one amino acid, the code is called degenerated. Amino acids are bound to transfer RNAs (tRNAs) having an anticodon (a codon that is complementary to the codons in Figure 5), a complementary nucleotide sequence that matches the triplet nucleotide that was read by the ribosome. After successful bonding between codon and anti-codon, the bigger subunit disentangles the amino acid from the tRNA and adds it to the amino acid chain, which in the end, forms the protein sequence that is a translated version of the mRNA to amino acids. The ribosomes finish translating when reaching one of the three stop codons (i.e., UAA, UAG, UGA for most organisms) [85,86].

2.1.3. Transcriptional regulation

Faulty transcriptional regulation is known to be responsible for at least a third of human disorders and diseases [10–12]. Hence, a detailed understanding of this mechanism is a prerequisite to developing targeted therapies [13,14]. Transcriptional regulation can take place in many forms (see Figure 6) and is dependent on the structural organization of the DNA. The structure of the DNA can be divided into three levels: (i) the primary level (i.e., the sequence of nucleotides), (ii) the secondary level, which is in most cases a double helix, and (iii) the tertiary level which is the overall 3-D conformation of the DNA molecule. The 3-D conformation of the DNA can be influenced by various factors, such as the presence of proteins or chemical modifications to the DNA itself (e.g., chromatin environment, histone modifications - see below) [85,87]. The tertiary structure is a critical component of epigenomics and, therefore, of transcriptional regulation [88].

One way the transcription is controlled is via the chromatin environment. Chromatin is a complex of DNA and proteins that makes up the structure of chromosomes in the nucleus of a cell. Nucleosomes are the basic structural units (i.e., the proteins) of chromatin. A nucleosome consists of a segment of DNA wrapped around a histone octamer, which includes two copies of each of the histones (i.e., proteins) H2A, H2B, H3, and H4 (see Figure 6) [89]. The DNA winds around this protein core roughly 1.65 times, with about 147 base pairs [89]. The positioning and modification of nucleosomes (e.g., acetylation or methylation, among others, see below) can influence DNA accessibility. Histone acetylation and methylation are post-translational modifications and are called HMs [90]. Acetylation, involving the addition of an acetyl group to histone lysine residues (i.e., K), typically neutralizes the positive charge, leading to a relaxed chromatin structure and, hence, to gene transcription [90]. On the other hand, histone methylation, which involves adding a methyl group to lysine or arginine residues, can either activate or repress gene transcription based on the specific location and degree of methylation [90]. When lysine residues at positions H3K9 (i.e., lysine at position 9 at histone protein 3) or H3K27 (i.e., lysine at position 27 at histone protein 3) are methylated, it typically leads to a closed chromatin conformation, resulting in gene silencing [90]. Conversely, methylation of H3K4 (i.e., lysine at position 4 at histone protein 3) is often associated with an open chromatin structure, promoting gene transcription [90]. The whole of nucleosomes and their acetylation or methylation state is called the chromatin environment [91].

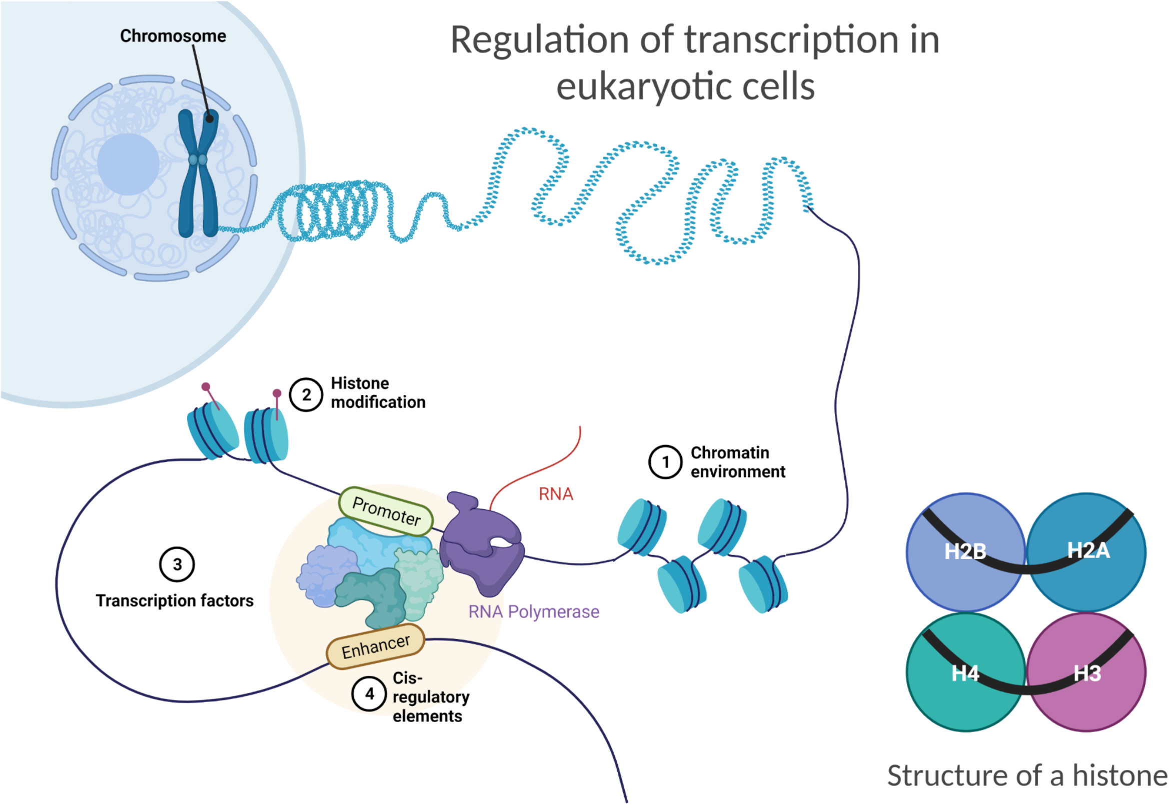


Figure 6: ***Organization of the DNA and regulatory elements affecting RNA polymerase 2.*** *Several possibilities exist to regulate the transcription of genes, e.g., chromatin environment, histone modification, transcription factors, and cis-regulatory elements. The Figure was created using Biorender.com.*

The most common active form of transcriptional regulation and also fine-tuning of regulation happens via CREs - short DNA sequences that are located in the DNA where TFs can bind. A TF that binds to a CRE (e.g., a specific promoter or (super-)enhancer - see below) can regulate (activate or repress) the expression of target genes without altering the DNA sequence itself [85,86]. A TF can also interact with other TFs (sometimes called co-factors) or with the chromatin environment to modulate gene expression [92,93]. CREs, in combination with TFs, are important for the regulation of gene expression, which is essential for the normal functioning of cells and organisms. Dysregulation of gene expression can lead to various diseases and disorders [85,86].

Some CREs can be found near or inside a gene and can regulate the expression of that gene (e.g., promoters). A promoter contains specific sequences that bind TFs and is in close proximity upstream of the target gene. In most cases, promoters, in combination with TFs, control the initiation of the transcriptional process by interacting with the RNA polymerases (see Figure 6) [85,86]. Other CREs (e.g., enhancers - see below) can be located very far (thousands of bps) from the gene or genes they control. TFs binding to (super-)enhancer or suppressor sequences can upregulate or repress the expression of the target gene. As depicted in Figure 6, an enhancer bound with a TF forms a loop with the promoter and interacts with the RNA polymerase. Superenhancers are currently poorly defined, but a super-enhancer is usually located more closely to the genes it regulates and is frequently found near genes that control cell identity. Additionally, super-enhancers often control multiple target genes and consist of multiple enhancer sequences in close proximity [94–96]. In the past, Shin et al. found that promoters, enhancers, and super-enhancers have combinatorial effects and that the highest expression can be achieved if all CREs are active and bound by activating TFs [97]. However, the impact of enhancers is not strictly additive, making it challenging to rank their relevance in comparison to one another [98]. Nevertheless, it is evident that the TFs with higher binding opportunities in these areas play a significant role in regulating gene expression, making them crucial to study [98].

2.1.4. Post-transcriptional regulation

Post-transcriptional regulation has been recognized in the past as key to understanding the underlying architecture of diseases [99]. As in most regulatory levels, post-transcriptional regulation can happen in many different manners. Especially non-coding RNAs (such as siRNAs, miRNAs, pseudogenes that can be coding or non-coding [100], lncRNAs, and circRNAs, see Figure 7. a) have been recognized to regulate mRNAs. Small RNAs like siRNAs (~20-25 nucleotides) and miRNAs (~19-25 nucleotides) can play an important part in post-transcriptional regulation [101–103]. While siRNAs are highly specific with only one mRNA target, miRNAs can have multiple (up to hundreds of) targets [104] due to, e.g., wobble-pair binding (i.e., some mismatches between miRNA and target sequence) [105]. RNAs that can either bind many miRNAs or just a few miRNAs but have many miRNA binding sites are called sponges. lncRNAs (> 200 nucleotides) can interact with mRNA directly and either enhance or repress translation [106] and can also upregulate the translation of mRNA indirectly by binding miRNAs that would otherwise prohibit its translation [30].

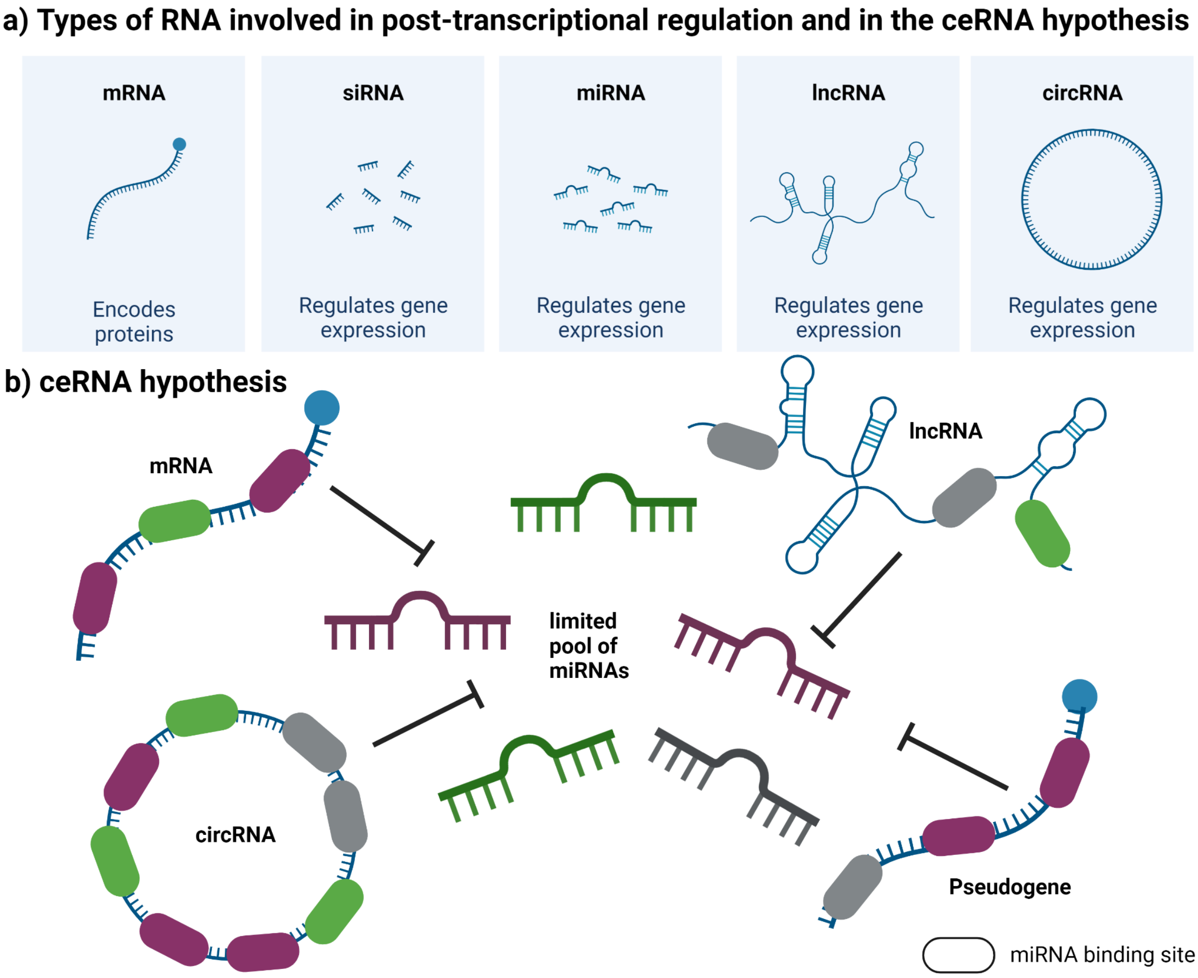

a) Types of RNA involved in post-transcriptional regulation and in the ceRNA hypothesis
mRNA
Encodes proteins
siRNA
Regulates gene expression
miRNA
Regulates gene expression
lncRNA
Regulates gene expression
circRNA
Regulates gene expression
b) ceRNA hypothesis
mRNA
lncRNA
limited pool of miRNAs
circRNA
Pseudogene
miRNA binding site

Figure 7: ***Layers of post-transcriptional regulation. (a)*** *Some of the most important types of RNA that are involved in post-transcriptional regulation are depicted.* ***(b)*** *I visualized the ceRNA hypothesis, different small molecules compete for a limited pool of miRNAs. The Figure was created with Biorender.com.*

miRNA biogenesis involves several steps (see Figure 8. a), including (i) transcription by RNA polymerase 2 to form primary miRNA transcripts (pri-miRNAs) - a double-stranded RNA, (ii) processing, during which pri-miRNAs are cleaved by the RNase 3 enzyme an endonuclease enzyme that cleaves double-stranded RNA (dsRNA) molecules [102,107,108]) and further processing by Drosha to form precursor miRNA (pre-miRNA) molecules, and (iii) maturation,i.e., pre-miRNAs are then transported out of the nucleus and cleaved by the RNase 3 enzyme Dicer to form mature miRNAs. Mature miRNAs are part of the RNA-induced silencing complex (RISC) that is essential in posttranscriptional regulation. This thesis refers to mature miRNAs as miRNAs [107–109].

circRNAs were first discovered in pathogens in 1976 by Sanger et al. as “closed circular RNA molecules” [67]. The process of circRNA biogenesis is not yet fully understood. From what is known, circRNAs are formed in the nucleus through a process called back-splicing (Figure 8. b), in which the exons (during the process of alternative splicing, Sec. 2.1.2) of a pre-mRNA molecule are spliced together in a circular fashion [110] (i.e., circularization of the exons happens during the attachment of a 5 ′ splice site to an upstream 3 ′ splice site on the same pre-mRNA [111]). The circRNA is then stabilized by base pairings of Alu elements (i.e., inverted repeats of complementary nucleotides) “in the flanking introns and/or association of specific RNA-binding proteins (RBPs)” (i.e., proteins that bind to the circRNA and stabilize the structure) [111]. Due to their circular form, circRNAs are resistant to exonuclease-mediated degradation and, therefore, more stable [112]. Even though their expression rate seems low, their overall expression can exceed the expression of their linear counterparts over time due to not being degraded [112].

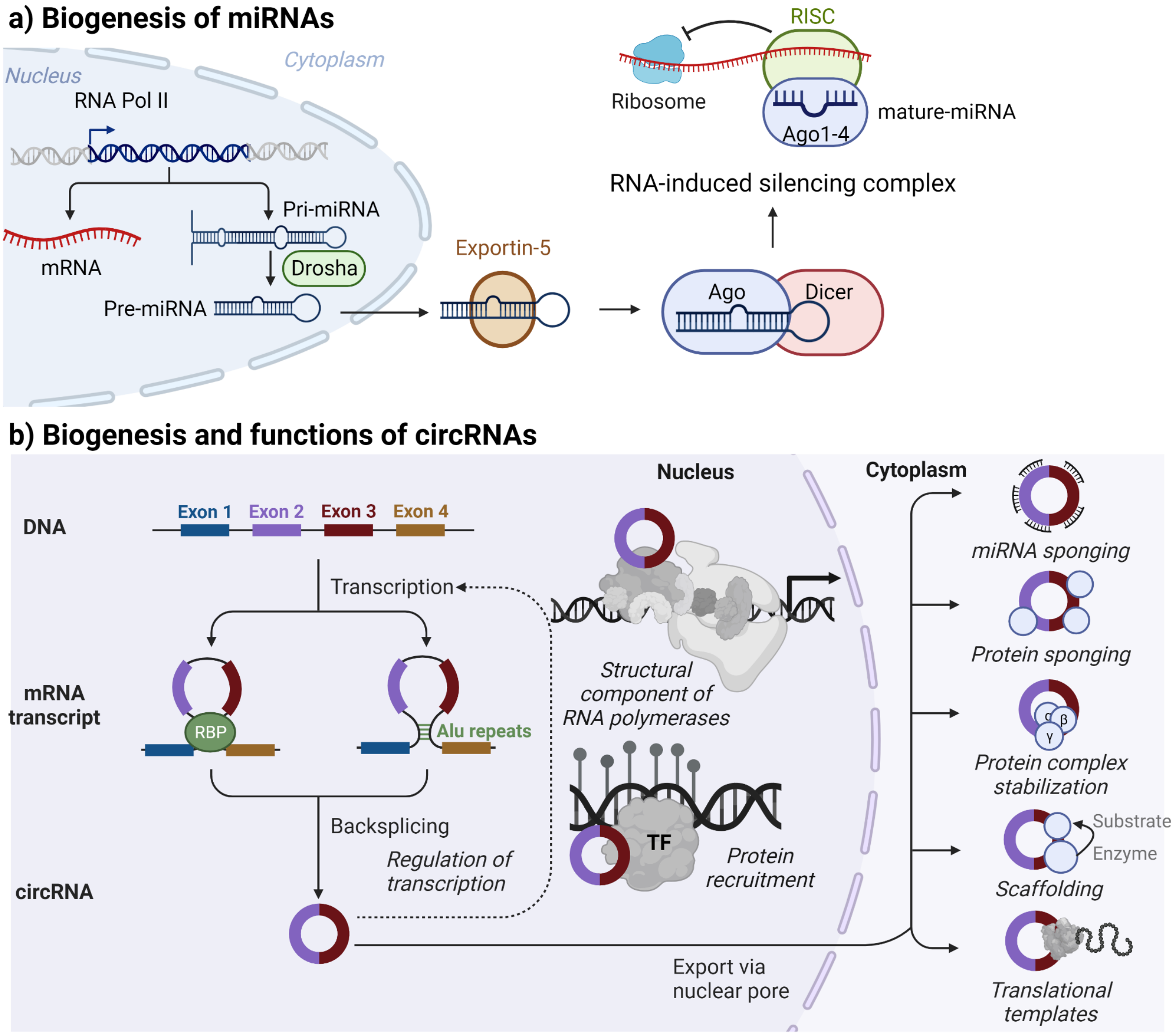
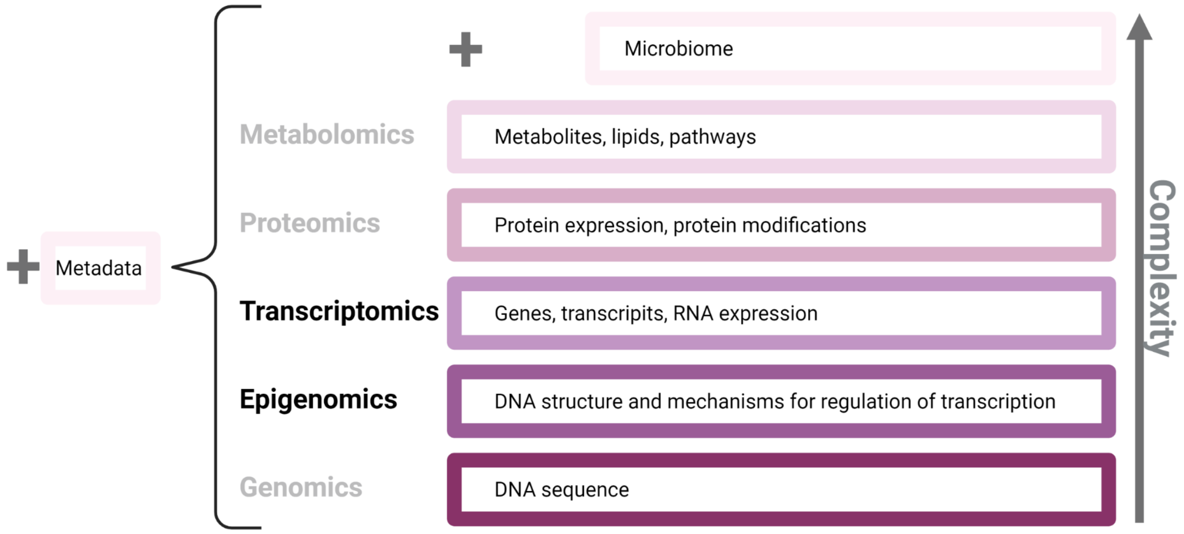


Figure 8: ***Biogenesis of (a) miRNAs and (b) circRNAs.*** *The Figure was created with Biorender.com.*

circRNAs have a wide range of functions (Figure 8. b). They can regulate at the transcriptional level inside the nucleus (e.g., regulating the RNA polymerases [113] and recruitment of TFs [114]), at the post-transcriptional level outside of the nucleus (e.g., miRNA sponging [115]), and at the post-translational level (e.g., protein sponging [116], protein complex stabilization [117], scaffolding [118]). Additionally, scientists recently detected circRNAs that code for proteins [36]. In this thesis, I focus on the potential interactions among circRNAs, miRNAs, protein-coding messenger RNAs (mRNAs), and other RNA types sharing miRNA binding sites that could constitute a potentially large regulatory network. Researchers found that most of the detected circRNAs have a higher density of miRNA binding sites than their linear counterparts [118,119]. Any RNA with miRNA binding sites can function as a miRNA sponge [120]. Sponges are important to determine as they could be involved in the regulation of mRNAs (i.e., a sponge sponges all miRNAs that would be required to regulate another mRNA) [120]. This gives rise to a complex indirect regulatory network that Salmena et al. name the ceRNA network [32] (see Figure 7. b). One RNA can have multiple miRNA binding sites, and multiple miRNAs can bind to this binding site. The ceRNA theory says that the RNAs are in competition with these miRNAs, and if one ceRNA is overexpressed, it could, therefore, indirectly regulate other ceRNAs that have binding sites of the same miRNA that is ultimately sponged by the overexpressed ceRNA [32]. With our endeavors during this thesis, I intend to provide the possibility of quick and easy detection of significant sponging events of circRNAs between conditions.

2.2. Molecular OMICS levels

OMICS is a general term and suffix for branches of scientific data produced in research fields such as biology and biomedicine. As there is a plethora of OMICS terms, I elaborate on the most common ones that are essential in bioinformatics and this thesis: (i) genomics, (ii) epigenomics, (iii) transcriptomics, (iv) proteomics, (v) metabolomics, and (vi) the metagenomics (commonly called the microbiome). Each of them usually includes metadata (e.g., clinical data, information about the sample, Figure 9) [121]. In this chapter, I discuss several methods commonly used to experimentally gather data from cells.



+
Microbiome
Metabolomics
Metabolites, lipids, pathways
Proteomics
Protein expression, protein modifications
+
Metadata
Transcriptomics
Genes, transcripts, RNA expression
Epigenomics
DNA structure and mechanisms for regulation of transcription
Genomics
DNA sequence
Complexity

Figure 9: ***Overview of the common OMICS research fields in Bioinformatics.*** *Bold OMICS are the ones described and investigated in this thesis. The Figure was created with Biorender.com.*

Data from the genomics field refers to the sequence of the DNA of an organism. There are plenty of different methods to determine the DNA sequence, from the first one in 1977 by Sanger et al. (see below) [122] and the next-generation sequencing methods to the more recent ones like real-time sequencing that are significantly faster [123]. Epigenomics refers to the DNA modifications that affect the regulation of gene expression. There are methods like Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) [124] and

DNase-seq [125] to determine the structural properties (e.g., open/closed chromatin regions) of the DNA (see Sec. 2.2.1). HM ChIP-seq [126] can also be used to determine structural properties among other use cases. Transcriptomics focuses on the expression levels of genes, including protein-coding or non-coding transcripts and other RNAs (e.g., miRNAs or circRNAs). There is a plethora of methods that perform RNA-seq to determine the RNA expression levels of transcripts longer than 100 bps [127] (see Sec. 2.2.2). There are also protocols like totalRNA-seq available that capture a bigger picture (e.g., circRNAs that lack a polyA tail and cannot be sequenced with standard polyA enrichment methods) of expressed RNAs in the cell [128] (see Sec. 2.2.2). For smaller RNAs (like miRNAs) there are protocols like miRNA-seq available (see Sec. 2.2.2). The proteomics term aims to determine protein expressions and protein modifications, typically using mass spectrometry [129]. In metabolomics, the center of attention is on metabolites and lipids, typically also using mass spectrometry. The microbiome focuses on the community of bacteria living in a given habitat (e.g., an organism's gut) [121]. There is a plethora of methods available to determine the proportion of the different bacteria in the gut and investigate its implications on the overall health of the organism (see [130]).

Multi-OMICS refers to analyzing multiple types of omics data, such as genomics, transcriptomics, proteomics, and metabolomics, in a single study. By integrating data from multiple omics disciplines, one could intend to better understand the interplay between different molecular components and their impact on cellular processes and disease [131,132]. Multi-OMICS studies often rely on machine learning and network medicine to integrate the data and unravel insights in medical research, such as personalized medicine, drug development, and disease diagnosis and treatment [133]. In this thesis, I intend to use epigenomics and transcriptomics data in combination to discover different TF activity between conditions.

One can classify molecular experiments as: in vivo, in vitro, and in silico based on the environment in which they are conducted. The term in vivo is used when experiments are conducted within a living organism. The term in vitro refers to experiments administered outside of a living organism, such as in test tubes or Petri dishes, and often involves cell lines (i.e., a population of cells derived from a single cell and cultured under controlled conditions). The term in silico indicates that experiments were conducted through computer simulation or computational models [134].

Genomics, epigenomics, transcriptomics, and the microbiome rely on the sequencing of the DNA of the genome or complementary DNA (cDNA) that was reverse-transcribed from RNA. To understand the principles of sequencing, I superficially elaborate on Sanger sequencing, the first developed sequencing method ([122], see below). However, since the development of Sanger sequencing, other faster methods such as next-generation sequencing (e.g., Illumina ([135]) and third-generation sequencing have become available (e.g., Single-Molecule Real-Time sequencing, which can also capture full transcripts by long-read sequencing) that won't be discussed in detail in this thesis [136]. In Figure 10, I illustrated the principle of Sanger sequencing.

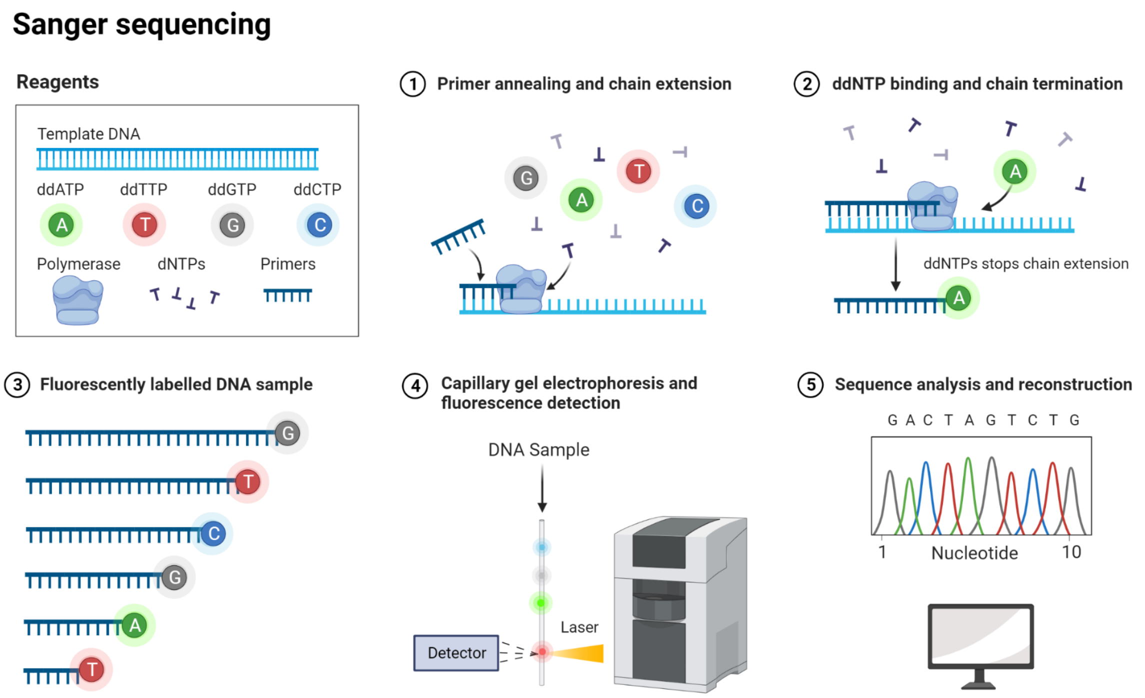

Sanger sequencing
Reagents
Template DNA
ddATP
ddTTP
ddGTP
ddCTP
A
T
G
C
Polymerase
dNTPs
Primers
1 Primer annealing and chain extension
G
T
A
C
2 ddNTP binding and chain termination
A
ddNTPs stops chain extension
A
3 Fluorescently labelled DNA sample
G
T
C
G
A
T
4 Capillary gel electrophoresis and fluorescence detection
DNA Sample
Laser
Detector
5 Sequence analysis and reconstruction
G A C T A G T C T G
1
Nucleotide
10

Figure 10: ***Illustration of the Sanger sequencing protocol.*** *The Figure was created with Biorender.com.*

In the Sanger method (Figure 10), a DNA fragment to be sequenced is replicated in vitro in four separate reactions (i.e., physically separate containers), each containing a different dideoxynucleotide (ddNTP; e.g., ddATP, ddCTP, ddGTP, or ddTTP) along with the regular deoxynucleotides (dNTPs; e.g., dATP, dCTP, dGTP, dTTP). These ddNTPs are similar to the standard nucleotides used in DNA replication but lack a 3' hydroxyl group. This absence prevents the addition of further nucleotides, hence terminating the DNA strand elongation. The replication results in a series of DNA fragments of varying lengths, each ending with the respective ddNTP. These fragments are then separated by size using capillary electrophoresis, with the smallest fragments moving the fastest. A laser excites the fluorescently labeled ddNTP at the end of each fragment, and the emitted light is detected and recorded. The color of the light identifies which base is at the end of the fragment, providing the sequence of the original DNA fragment. [122]

In the remainder of this section, I explain how one can retrieve data experimentally in the layers of epigenomics and transcriptomics using different protocols and how one can process them.

2.2.1. Epigenomics

In epigenomics (i.e., structural properties and accessibility of the DNA), one wants to determine the organization of the DNA (e.g., which part of the DNA is open and can be transcribed into RNA) or at what position a protein can bind to enhance or repress transcription (i.e., a TF binds to a CRE) by employing well-known protocols like ATAC-seq, DNase-seq, or ChIP-seq. In this thesis, I employ epigenomics in combination with transcriptomics to investigate different TF activities between conditions.

The protocols

One possibility to assess the accessibility of chromatin is to use the ATAC-seq protocol (see Figure 11. a). ATAC-seq uses Tn5 transposase, an enzyme that can move DNA sequences to another site of the genome by cutting the DNA at a specific site, known as the transposon. [137]. Hence, it can be used to insert sequencing adapters (i.e., short pieces of DNA that are used to prepare a sample DNA fragment for sequencing [138]) into open chromatin regions [137]. The general idea is that regions that are inaccessible remain unreachable to Tn5 intervention, thus excluding them from sequencing. As a result, sequencing reads predominantly correspond to areas where the chromatin is in an open configuration. The ATAC-seq protocol includes several steps: (i) Tn5 transposase carrying adaptors binds to open chromatin regions in the DNA, (ii) fragmentation of the binding sites by the Tn5 transposase, (iii) amplification using polymerase chain reaction (PCR, making copies of the fragments [139]) and purification of the DNA (e.g., using alcohol [140]), and (iv) sequencing of the DNA (e.g., Sanger sequencing) [124]. Ultimately, one can investigate the open chromatin regions and their possible implications by employing bioinformatic pipelines (see below).

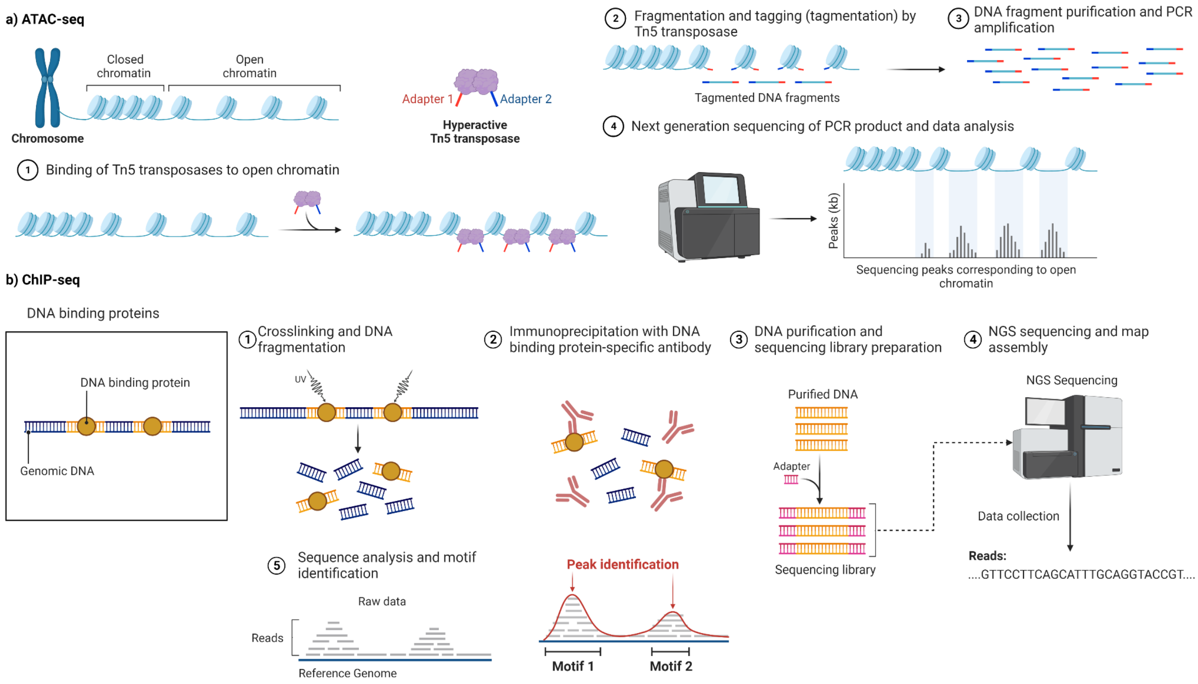
A second but older, less qualitative, and less used technique to study CREs is DNase-seq. The key principle underlying DNase-seq involves treating chromatin with DNase I, an enzyme that preferentially cuts DNA in regions where it is least condensed, i.e., where transcription factors or other DNA-binding proteins have made the chromatin accessible. After DNase I treatment, DNA is extracted and sequenced. The sequenced fragments correspond to regions in the genome where chromatin is open and can then be used for downstream analysis. [125]

Another technique to study open-chromatin regions or to identify the locations of specific DNA-binding proteins, such as transcription factors binding to CREs, within the genome is ChIP-seq (see Figure 11. b). ChIP-seq employs an antibody - a protein that recognizes a specific DNA-binding protein (e.g., a TF) or an HM - to identify proteins binding to the DNA or open-chromatin regions [141–143]. The ChIP-seq protocol uses the following steps: (i) It first crosslinks the protein with the DNA using formaldehyde, a small molecule that has the ability to covalently link proteins (i.e., antibodies) to DNA through the formation of methylene bridges; the protocol specifies then to fragment the DNA (e.g., using sonication [144]) [145], (ii) then the protocol specified to use a specific antibody to pull the protein-DNA complexes from the sample (immunoprecipitation); next the immunoprecipitated protein-DNA complexes are purified by reverse crosslinking (e.g., using heat [146]) to gain the DNA fragments without the bound protein, (iii) then it adds sequencing adaptors and multiplies the DNA fragments by PCR, (iv) afterward it sequences the DNA fragments. Bioinformatic pipelines can then be used to investigate the experimental data [126] (see below).

Since each of the protocols (ATAC-seq, DNase-seq, and ChIP-seq) have different approaches to studying important elements and structures of the DNA, one has to carefully accommodate for biases in downstream analysis (e.g., using the HINT-ATAC method, described in Section 3.1.2). The results of ATAC-seq, DNase-seq, or ChIP-seq, combined with RNA-seq data (Sec. 2.2.2), can be used to investigate how changes in chromatin or TFs can alter the expression of a target gene.

In the remainder of this section, I elaborate on the files and data formats generated by the protocols, as well as bioinformatic tools used to preprocess and analyze the data. ATAC-seq and ChIP-seq both produce FASTQ files that contain sequenced reads. I employ the nf-core pipelines [147] (i.e., a framework that is based on Nextflow [148], a pipeline management tool that allows for parallel execution of tasks on a cluster or cloud infrastructure) nf-core/atac-seq and nf-core/chip-seq [147] to perform quality checks using FastQC [149,150] and MultiQC [151] (see below) and process the files to several file formats that are often used for downstream analysis. The pipelines use the tool “Trim Galore!” (see below) [152] to trim the adaptors necessary for sequencing to prepare the data for alignment (i.e., map sequences to a reference genome). Per default, the pipelines use the Burrows-Wheeler Aligner (BWA) tool (see below) [153] to map the raw and adapter-free sequences to the reference genome. After mapping, the pipelines automatically prepare more standard files (e.g., bigWig) typically used as input in downstream analysis (e.g., bigWig for visualization in a genome browser). One of the next important steps is peak calling using the tool Model-based Analysis of ChIP-seq version 2 (MACS2) (see below) [154], as one can detect regions of the genome that have a higher-than-expected coverage of sequencing reads that could indicate a transcription factor binding site (TFBS) in TF ChIP-seq data or an open chromatin region in HM ChIP-seq data.

In the following, I explain the most important tools that I use for preprocessing FASTQ files in more detail: (i) FastQC and MultiQC, (ii) TrimGalore!, (iii) Burrows-Wheeler Aligner, and (iv) MACS2.



a) ATAC-seq
Closed chromatin
Open chromatin
Chromosome
Adapter 1
Adapter 2
Hyperactive Tn5 transposase
2 Fragmentation and tagging (tagmentation) by Tn5 transposase
3 DNA fragment purification and PCR amplification
Tagmented DNA fragments
4 Next generation sequencing of PCR product and data analysis
1 Binding of Tn5 transposases to open chromatin
Peaks (kb)
Sequencing peaks corresponding to open chromatin
b) ChIP-seq
DNA binding proteins
DNA binding protein
Genomic DNA
1 Crosslinking and DNA fragmentation
UV
2 Immunoprecipitation with DNA binding protein-specific antibody
3 DNA purification and sequencing library preparation
Purified DNA
Adapter
Sequencing library
4 NGS sequencing and map assembly
NGS Sequencing
Data collection
Reads:
....GTTCCTTCAGCATTTGCAGGTACCGT....
5 Sequence analysis and motif identification
Raw data
Reads
Reference Genome
Peak identification
Motif 1
Motif 2

Figure 11: ***Illustration of the (a) ATAC-seq and (b) ChIP-seq protocol****. The Figure was created with Biorender.com.*

Quality control using FastQC and MultiQC

FastQC is a software tool designed to provide an initial, visualized quality-check overview of raw sequencing data per sample. FastQC analyses include (i) per base sequence quality, (ii) per sequence quality scores, (iii) per base sequence content, (iv) per sequence GC content, (v) sequence length distribution, (vi) duplicate sequences, and (vii) overrepresented sequences. The (i) per base sequence quality represents the probability of an incorrect base call (i.e., an inaccurately identified nucleotide), with higher scores indicating higher confidence in the base call. Low-quality bases can introduce errors such as incorrect variant calls or misassembled transcripts. A (ii) per sequence quality score is a combined score of per base sequence quality and helps to investigate if a subset of the sequences has consistently low-quality scores. Sequences with bad quality could then be removed for further analysis. The (iii) per base sequence content measure shows the proportion of each base (A, T, C, G) at each position, which should be approximately equal for a random library. If not, it may indicate problems like adapter contamination or over-amplification of certain regions, affecting the representation of the sequences. The (iv) per sequence GC content of a genome or a set of reads should be approximately normally distributed. Deviation from this norm might suggest contamination, over-amplification of specific regions, or a systematic bias in the library preparation, which may affect the validity of variant detection or gene expression estimation. The (v) sequence length distribution should be consistent for certain sequencing methods. Variations could indicate issues with the sequencing process or data truncation, which could cause problems with the alignment and interpretability of downstream analyses. The (vi) analyses of duplicate sequences and (vii) overrepresented sequences identify sequences that appear more often than expected by chance, which could be due to technical artifacts (like PCR duplication) or biological significance (like highly transcribed RNA). Excessive duplicates may skew the interpretation of ChIP-seq or RNA-seq analyses by overrepresenting certain regions and should be removed [149,150]. As FastQC processes one sample at a time, MultiQC aggregates results from multiple FastQC analyses into a single report, providing a summary of the whole dataset [151].

Adaptor trimming with TrimGalore!

Adaptor trimming is an essential preliminary step in the analysis of sequencing data. The process involves the removal of adaptor sequences, which are artificially added during sequencing. They must be removed in the final sequencing read, as they can cause false alignments and interfere with downstream data analyses. TrimGalore! [152] automatically trims adaptor sequences using information from FastQC to detect and remove adaptor sequences by searching for partial overlaps between the adapter sequence and the reads; hence TrimGalore! can also trim the adaptor sequence even if only partly present.

Burrows-Wheeler Aligner

Burrows-Wheeler Aligner (BWA) is a tool for mapping sequences against a reference genome, such as the human genome [153]. The primary operation of BWA involves transforming the genome into an index (see below), followed by aligning the reads against this indexed genome. This indexing strategy drastically reduces the computational requirements and enables efficient alignment of sequencing reads. The Burrows-Wheeler

Transform (BWT) is used for indexing. BWT is a string transformation algorithm that reorganizes the string into sequences of similar characters. This transformation is particularly useful because it tends to produce a large number of sequences with repeated patterns that are beneficial for compression and is also reversible, meaning that the original string can be reconstructed from the transformed string. In general, the transformation works as follows: (i) creation of a suffix array (i.e., a sorted array of all suffixes of a string). It represents the positions at which each of the sorted suffixes would appear in the genome. (ii) the Burrows-Wheeler Transform is applied to the array, i.e., BWT reorganizes the reference genome such that all the genomic characters (A, C, G, T) that are alike cluster together. (iii) The transformed array is then indexed for its rank, which will allow efficient searching later. The BWT is then used in the BWA to align the sequenced read to the genome, which will result in the Sequence Alignment/Map (SAM) format, which provides detailed information on the alignment of each read. These SAM files were further converted into the more compact and smaller Binary Alignment/Map (BAM) format for downstream analysis [153].

Peak calling with MACS2

MACS2 investigates reads mapped onto the genome to identify regions or "peaks" where there is a significant enrichment of reads and, hence, potentially correspond to protein-DNA binding sites. To achieve the identification of such protein-DNA binding sites, MACS2 compares the signal in the experimental sample (e.g., ATAC-seq, DNase-seq, or ChIP-seq) against a control sample (i.e., a sample generated by sequencing randomly sheared DNA from the same cell type and, ideally also, the condition as the sample. The control sample represents the baseline DNA fragment level to account for experiment biases and to enable more accurate identification of regions with significant protein-DNA interaction enrichment, or "peaks" over the background noise) [154].

2.2.2. Transcriptomics

Transcriptomics captures the RNA levels in an organism or a cell. In this thesis, I intend to use transcriptomics data to (i) determine different TF activity between conditions and (ii) detect circRNAs and investigate their miRNA sponging capability and what their functions could be.

The protocols

In transcriptomics, one generally studies the set (i.e., the transcriptome) of RNA molecules (e.g., mRNA, lncRNA, circRNA, and miRNA) that are actively transcribed from the genome to understand the implications of their expression. The transcriptome changes over time, and gene expression can alter drastically from time point to time point [155]. To determine the transcriptome, one could use various technologies for RNA sequencing. RNA sequencing often relies on next-generation sequencing methods, which are executed in the following steps (Figure 12): (i) isolate RNA from the sample, (ii) fragment RNA in shorter segments, (iii) convert RNA into cDNA (using reverse transcriptase [156]), (iv) add sequencing adaptors to the RNA fragments, and (v) perform sequencing. Finally, one could use bioinformatic pipelines to map reads to the genome or transcriptome to quantify the expression values of genes and use this information for downstream analysis (see below). The first step (step (i)) to isolate the RNA from the sample can be modified to accommodate the type of RNA that should be sequenced, e.g., smaller RNAs (miRNAs, siRNAs,...) or longer RNAs (mRNAs, lncRNAs, circRNAs,...).

One could obtain FASTQ files that contain the raw sequences of the RNAs. Similar to the reads in epigenomics protocols, FASTQs from transcriptomics protocols are assessed for quality using FastQC and MultiQC (see Sec. 2.2.1) and adaptor-trimmed by TrimGalore! (see Sec. 2.2.1) using the nf-core/RNA-seq pipeline. As a next step, the pipeline determines which read belongs to which gene to quantify the gene expression in the sample. To map reads to the genome or transcriptome, one needs to consider alternative splicing (i.e., the spliced intron-exon structure), so exact matching is not possible. The aligning tool spliced transcripts alignment to a reference (STAR) tackles the aligning problem by using the seed-and-extend approach (i.e., the approach first identifies a small subset of the read (seed) that can be confidently aligned to the genome and then extends the alignment from the seed to cover the entire read, see below for more details) [157]. The output of STAR could be used for Salmon, a tool that quantifies gene expression by counting the number of reads that align to each exon of the gene (see below for more details [158]). Once one obtains the quantification of the genes or transcripts per sample, one could determine the changes in gene expression between conditions (e.g., health and control).

Aligning reads using STAR

STAR operates by first generating a genome-wide suffix array (i.e., a sorted array of all suffixes). This suffix array includes both the reference genome sequence and annotated splice junctions to match reads spanning across exons. STAR proceeds as follows: (i) STAR searches for seeds by breaking reads into shorter pieces and aligning these seeds to the reference genome. STAR then stitches the aligned seeds together into a coherent alignment for each read. If a seed aligns to separate exons, STAR infers the existence of a splice junction corresponding to the gap between the aligned seeds [157].

Quantifying gene expression using Salmon

Salmon can quantify transcript abundances from RNA-seq data by either using raw sequencing data or already aligned sequencing reads from other tools, such as STAR, through its alignment-based mode. In nf-core/RNA-seq, Salmon takes the alignments and uses an expectation maximization algorithm to infer the transcript abundances that are most likely to have resulted in the observed aligned reads. Salmon then optionally performs bootstrapping (i.e., performs quantification over and over again with subsampled datasets) to estimate uncertainty in the abundance estimates [158].

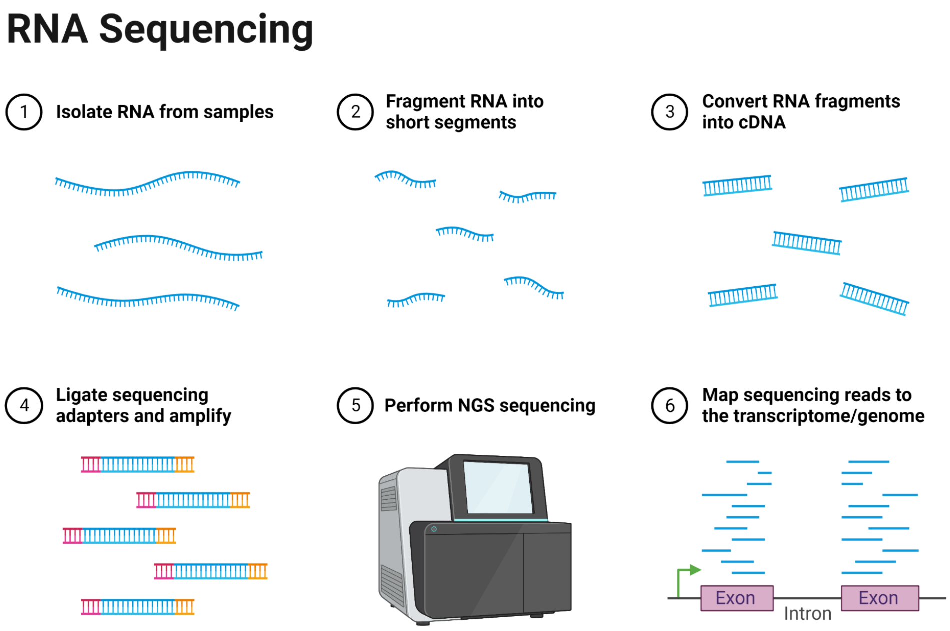

RNA Sequencing
1 Isolate RNA from samples
2 Fragment RNA into short segments
3 Convert RNA fragments into cDNA
4 Ligate sequencing adapters and amplify
5 Perform NGS sequencing
6 Map sequencing reads to the transcriptome/genome
Exon
Intron
Exon

Figure 12: ***Illustration of RNA-sequencing.*** *The Figure was created with Biorender.com.*

Batch effects in RNA-sequencing

Batch effects are non-biological variations in data that occur due to technical inconsistencies in processing samples. These inconsistencies could arise from a plethora of causes, such as different timing, personnel, or reagents used in the experimental procedures. In transcriptomics, batch effects can obscure genuine biological differences and lead to inaccurate interpretations. Several strategies exist to correct for these effects. First, a thoughtful experimental design can help minimize their impact by evenly distributing samples from different conditions across batches. In data analysis, statistical methods, such as ComBat, are often employed to adjust for batch effects. ComBat first normalizes the means of each batch to correct for additive batch effects; secondly, it adjusts the variances to account for multiplicative batch effects. By doing this, ComBat makes data more comparable across batches. However, this method can cause overcorrection, which could remove real biological signals and should, thus, be used with caution [159].

2.3. Machine learning

Machine learning methods enable computers to understand structures and logic from existing data and, based on the gained knowledge, make predictions on new data [160]. Such methods have been successfully used in biomedical research, facilitating the analysis of large-scale biological data to unveil insights into molecular mechanisms and interactions, e.g., identifying faulty regulatory mechanisms in diseases [161]. It enables researchers to make previously unattainable predictions and discern patterns; however, it is essential to note that predictions must be experimentally validated. In general, machine learning methodologies can be divided into two primary categories: supervised learning and unsupervised learning (see below, Figure 13, [162]). In this thesis, I employ a regression-based approach to determine different TF activity between conditions and a random forest approach to extract potentially important circRNAs between conditions.

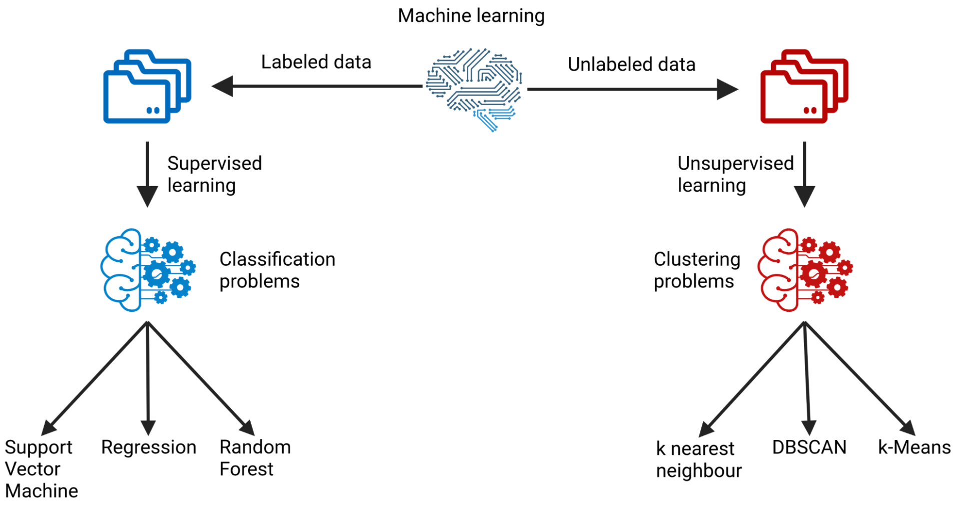


Figure 13: ***Overview of machine learning models.*** *The Figure was created with Biorender.com.*

2.3.1. Supervised learning

Supervised learning is often employed for classification problems (e.g., distinguishing between healthy and unhealthy individuals). In this approach, an algorithm is trained using a labeled dataset (i.e., samples labeled as healthy and unhealthy). The main objective of supervised learning is to create a model capable of generalizing from the training data to accurately predict labels for previously unseen, independent samples. Commonly used machine learning models are, e.g., support vector machines, regression (see Sec. 2.3.5), and random forest (see Sec. 2.3.6) [160,163]. A real-world example of supervised machine learning could be disease classification based on gene expression profiles in transcriptomics. For instance, with patient gene expression data labeled for leukemia presence or absence, one could train models like a regression or random forest to predict leukemia likelihood in new patients.

2.3.2. Unsupervised learning

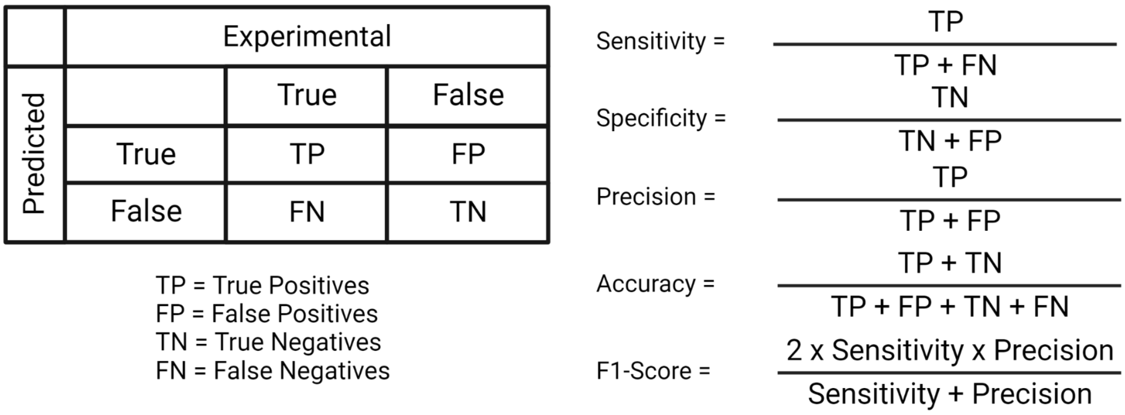
Unsupervised learning algorithms are often used for clustering, where similar data points are grouped together. In unsupervised learning, the algorithm is provided with an unlabeled dataset (e.g., the data does not contain information if the sample is healthy or unhealthy) and must discover patterns or underlying structures in the data without any guidance on the desired output. The goal is to find relationships among the input features, which can then be used for further analysis. Commonly used unsupervised machine learning models are, e.g., k nearest neighbor, principal component analysis, hierarchical clustering in heatmaps, DBSCAN, and k-Means [160,163]. An applied instance of unsupervised machine learning in transcriptomics might involve elucidating molecular disease subtypes. For instance, with tumor gene expression data that includes known subtype information, unsupervised techniques like hierarchical clustering can be employed. These methods can group tumors based on gene expression patterns without utilizing the provided subtype labels. When the unsupervised grouping aligns with the pre-established labels, it allows us to identify novel biomarkers for these subtypes.

2.3.3. Overfitting

Overfitting is a common problem in machine learning, where a model learns the training data too well without generalizing it to unseen data. It captures the noise along with the underlying pattern in the data. When a model is overfitted, it performs well on the training data but poorly on new data. Methods like k-fold cross-validation (i.e., systematically leaving parts of the data out of training to test the model on it) aim to prevent overfitting [163].

2.3.4. Statistical evaluation metrics of machine learning models

The performance of machine learning models is typically assessed using various metrics (i.e., derived from the model's predictions on a test dataset where the outcome is known). Key metrics include measures calculated from the confusion matrix, a fundamental tool for evaluating the performance of a classification model (Figure 14).



		Experimental	
Predicted		True	False
	True	TP	FP
	False	FN	TN

TP = True Positives
FP = False Positives
TN = True Negatives
FN = False Negatives

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{F1-Score} = \frac{2 \text{ x Sensitivity x Precision}}{\text{Sensitivity} + \text{Precision}}$$

Figure 14: ***Visualization of a confusion matrix.*** *The Figure was created with Biorender.com.*

The matrix consists of four measures: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). True positives and true negatives are the instances correctly predicted by the model, while false positives and false negatives represent the errors made by the model. After determining TP, TN, FP, and FN, one generally calculates the following measures: (i) sensitivity (also known as recall), (ii) specificity, (iii) precision, (iv) accuracy, and (v) F1-score (for formulas see Figure 14. a). (i) The sensitivity measures the proportion of actual positive cases that the model correctly identified, reflecting its ability to avoid false negatives. (ii) The specificity measures the proportion of actual negative cases that the model correctly identified, representing its ability to avoid false positives. (iii) The precision calculates the proportion of predicted positive cases that are actually positive, indicating the model's ability to return only relevant instances and minimize false positives. (iv) The accuracy measures the proportion of total predictions that the model got right, providing an overall performance. (v) The F1-score is the harmonic mean of precision and sensitivity, balancing the two metrics to provide a single measure of model performance. Precision is preferred over specificity if false positives come with a high cost (e.g., while testing for a highly contagious and deadly infection).

2.3.5. Regression

Regression analysis belongs to the category of supervised machine learning [160]. There are various types of regression methods available, such as linear regression and logistic regression, each suited to different kinds of data and different types of prediction problems [164]. In general, with regression analysis, one could intend to model the relationship between a dependent variable and one or more independent variables (also known as "features") [163]. The aim of regression is to estimate the values of the dependent variable based on the values of the independent variable. In linear regression, one tries to fit a straight line that best approximates the relationship between the dependent and independent variables. This line is represented mathematically as: $Y = aX + b$, where Y is the dependent variable, X is the independent variable, $'a'$ is the slope of the line (showing how much Y changes for each unit change in X), and $'b'$ is the intercept (showing the value of Y when X is zero). While training, one intends to find the best values for $'a'$ and $'b'$ that minimize the difference between the predicted Y values (from the $aX + b$ equation) and the actual Y values in the training data. For example, by investigating the effect of a specific TF on the expression levels of a set of target genes, the hypothesis is that the presence (or the quantity) of this TF in a cell regulates the expression of these genes. In this example, the independent variable would be the concentration or activity level of the TF in the cell. The dependent variable would be the mRNA expression levels of the target genes. Regression analysis models how changes in the TF level (independent variable) influence changes in the mRNA expression level of the target genes (dependent variable).

2.3.6. Random forest

Random forest is a supervised machine learning model [160]. It works by constructing a multitude of decision trees and combining their predictions in a majority vote (i.e., the most frequent prediction determines the final prediction). To build a decision tree, one utilizes a dataset where each sample has a known condition. At each step, the data is split based on a feature that maximizes the separation of the conditions. This process is repeated recursively, resulting in a tree of decisions. A random forest works as follows: (i) random forest starts with bootstrap sampling (i.e., drawing a random subset of the data with replacement, hence the same data point can be sampled more than once). This sampling process results in different subsets of data, each used to train a separate decision tree. (ii) Next, one generates trees out of the bootstrap samples. At each node of the tree, a random subset of features is selected, and the best split among these features is chosen according to a defined criterion. This randomness in feature selection injects diversity into the model and helps prevent overfitting. (iii) Once the forest of trees is built, predictions are made by running the new input data through all the trees in the forest. The most chosen prediction among all trees is considered the final prediction (majority voting) [163].

2.4. Network and systems medicine

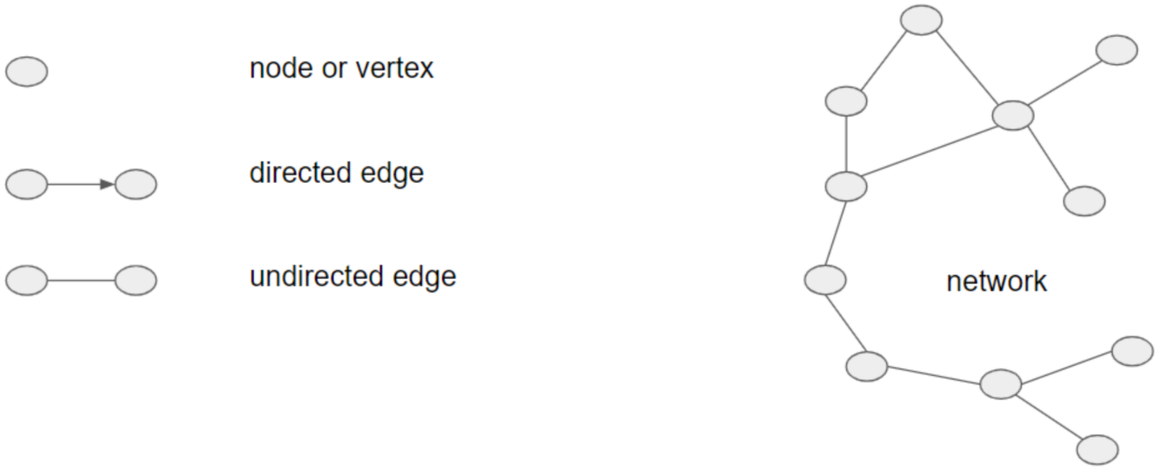


Figure 15: ***General overview of network components.*** *The Figure was created with Biorender.com.*

In general, networks can be used to visualize and order relationships between entities. In this thesis, I use networks to visualize and order ceRNAs into a regulatory network.

Network theory

Regulatory mechanisms (like other complex relationships) can be visually represented as networks due to their element-relationship-element nature. A network generally consists of two main components (see Figure 15): nodes and edges. Nodes embody the individual elements within the network (e.g., proteins, genes, CREs, TFs). Edges depict the relationships or interactions between node pairs in the network (e.g., regulation or interaction). Edges can be either directed or undirected, based on the interaction type (e.g., directed when one TF represses or enhances the expression of a gene but undirected when a protein interacts with a protein). Together, nodes and edges define the structure and connectivity of a network, enabling the analysis of complex systems. One can identify critical components such as hub nodes (i.e., nodes with high connectivity) with network analysis methods. In biological networks, hubs are usually important for a regulatory system, as a defective hub (such as a TF or CRE) usually has a major impact and thus may lead to disease [165,166].

Systems medicine

Systems medicine is the name of an interdisciplinary field that integrates various disciplines, such as the OMICS, bioinformatics, and network theory, to investigate complex biological systems and their role in health and disease. In systems medicine, one focuses on the interactions and relationships (e.g., regulatory mechanisms) between the components (e.g., regulatory elements) of a system rather than studying individual elements in isolation to investigate the underlying mechanisms [167,168].

3. GENERAL METHODS

3.1. TF-Prioritizer framework

3.1.1. The current state of condition-relevant transcription factor identification

TFs are proteins that modulate gene expression by binding to CREs [169,170] and, often referred to as TF activity (TFA), can influence cellular functions like development, differentiation, response to environmental stimuli, and also diseases [171,172]. To analyze TFA, one can employ data retrieved from RNA-seq that measure TF transcript levels, giving insights into TFs present in a cell. Moreover, methods such as ChIP-seq, ATAC-seq, and DNase-seq can provide insights into TF binding sites and their potential function [141]. Enhancing the accuracy of TF binding site predictions often includes combining motif-based predictions with epigenetic markers [173,174]. Many computational tools have been developed for TFA analysis and can be mainly divided into two types: those based on gene regulatory networks (GRNs; e.g., a network of TFs regulating gene expression of target genes) and those that focus on genome occupancy (i.e., assess the presence and location of TFs on the DNA, indicating where these molecules are active on the genome) [175].

Investigating transcription factor activity by utilizing gene regulatory networks

In the early 2000s, initial methods to approximate TFA utilized linear regression between TF expression levels and target gene expression levels. Given the limited data available at that time, these methods were often applied to yeast data due to their simpler genetic composition compared to mammalian cells [176]. Dujon examined sequences preceding differentially expressed genes to identify recurring motifs, then attempted to interpret expression variations based on these motifs [176]. The motif coefficients, which could implicitly signify a TF, served as an indirect measure of a TF's activity [177–180].

In 2003, Network Component Analysis (NCA) was introduced as a novel method to estimate TFA using gene expression data coupled with a priori knowledge of regulatory interactions between TFs and target genes [181]. NCA then breaks down the gene expression data matrix into two separate matrices: (i) the control strength matrix (describing the possible influence of TFs on target genes) and (ii) the TFA Matrix (capturing the activity of each TF across varying samples or conditions). Next, NCA leverages the control strength matrix; any interactions that don't have evidence backing them from the connectivity matrix are set to zero (e.g., experimentally validated interactions found in literature). Finally, NCA outputs the TFA matrix, which presents the activity of each transcription factor across samples or conditions from the initial expression data. However, NCA had its limitations, such as: (i) It relies heavily on gene expression data and prior knowledge about the regulatory network, meaning its effectiveness can be compromised without well-defined connectivity between transcription factors and their target genes. (ii) The method assumes a linear relationship between TFs and gene expression, potentially missing the intricate non-linear interactions in biological systems. (iii) Scalability is a concern for larger datasets due to increasing computational complexity. (iv) Additionally, the static connectivity matrix used by NCA does not adjust to the dynamic nature of biological systems, where connections might change based on different conditions [181]. This led to the development of enhanced versions like FastNCA, gNCA, ROBNCA, gfNCA, sparseNCA, and LNCA, each addressing specific challenges or extending its functionalities [182–187].

Other approaches tried to model activation and repression effects in the NCA approach [188]. Challenges particularly arise when examining heterogeneous samples like those from cancer patients, where genetic variations can hinder interpretations. To address this, tools like RACER [189] and RABIT [190] incorporate additional genetic factors into their modeling, such as: (i) RACER trying to model post-transcriptional regulation by integrating information about the mRNA degradation rates. The degradation rates provide insights into how long mRNAs remain stable in the cell after being transcribed. By considering these rates, RACER can differentiate between changes in mRNA levels due to transcriptional events (i.e., TF activities) and changes due to post-transcriptional processes (like mRNA degradation) [189]. (ii) RABIT uses a background model trying to capture additional genetic and epigenetic effects by randomizing gene expression data, disrupting any relationships between genes and their potential regulators. Using this randomized dataset, RABIT then establishes a baseline picture of what TFA might resemble in an environment devoid of actual TF-gene interactions. This baseline serves as a null model to which real data can be adjacent. Incorporating the insights from this background model, RABIT contrasts the regulatory activity inferred from real data with the expectations set by the background model, RABIT determines which regulatory connections are statistically significant, limiting the report of random occurrences [190].

Investigating transcription factor activity by utilizing genome occupancy

Some tools don't utilize gene regulatory networks and instead measure a TFs' genome-wide binding behaviors, aiming to identify significant TFs for specific samples or highlight those causing differences between conditions. A particular focus within this category is on the accessibility profile shape termed "footprints" (i.e., pinpoints of precise TF binding sites) [191].

One tool that tries to capture footprints is TOBIAS which offers a wide range of analyses for TF binding dynamics based on footprint data from ATAC-seq. TOBIAS corrects for ATAC-seq-specific biases since ATAC-seq has preferences for certain DNA sequences, leading it to insert sequencing adapters more frequently at these preferred sites, which ultimately might appear artificially more accessible than they truly are in vivo [192]. TOBIAS then employs a scoring function that considers two primary factors: the accessibility and the depth of the local footprint and correlates them with the presence or absence of transcription factor binding sites, ultimately enabling it to distinguish between regions where transcription factors are bound (active sites) and not bound (inactive or unbound sites) [192]. TOBIAS not only identifies footprints but also compares them across conditions (e.g., health and disease) and can predict potential interactions between them [192]. BaGFoot [193] is another tool leveraging chromatin accessibility data for TF footprinting. The primary focus of BaGFoot is on the depth of the footprint and the accessibility of the surrounding region at each motif occurrence of all given transcription factors [193]. A tool named HINT [194–196], which is described in Sec. 3.2. - “ATAC-seq or DNase-seq data processing using HINT-ATAC” - also belongs to this category. diffTF [19] focuses on estimating TFA changes between conditions using accessibility shifts at potential TF binding sites. It scans the genome for overlapping TF binding motifs with a consensus peak set from ATAC-seq data. The tool then calculates fold changes for all peaks of a TF, comparing them against a background distribution. If provided with additional expression data, diffTF can further classify TFs as activators or repressors [19].

However, most of these proposed methods to investigate TFA demand significant preprocessing, computational skills, method adjustment for novel use cases (for instance, more than two conditions and/or time-series data), manual results evaluation (such as a manual search and visualization for TF ChIP-seq data to evaluate the predictions), and often investigate only parts of the whole picture. Therefore, to make this process more efficient and comprehensive, I introduce TF-Prioritizer [51], a Java pipeline that prioritizes TFs displaying condition-specific changes in their activity. TF-Prioritizer automates several labor-intensive steps, such as data processing, TF affinity analysis, machine learning predictions of CREs relationships to target genes, prioritization of relevant TFs, data visualization, and visual experimental verification of the results using publicly available TF ChIP-seq data from ChIP-Atlas (see Sec. 3.1.2, [197]).

3.1.2. Integrated tools and techniques in TF-Prioritizer

In this section, I generally elaborate on methods that are integrated into TF-Prioritizer, including (i) general preprocessing [147]; (ii) RNA-seq data filtering, normalization, and processing using DESeq2 [198]; (iii) ATAC-seq or DNase-seq data processing using HINT-ATAC [194–196]; (iv) blacklisted regions and combining samples [199]; (v) preselection of condition-relevant transcription factors using TEPIC [200,201], the peak-valley-peak model [202], TRAP [203], and DYNAMITE [201]; (vi) final prioritization using the Mann-Whitney U Test [204–206] and a discounted cumulative gain; and (vii) ChIP-Atlas [197] and the Integrative Genome Browser [207–209] as the source for automatic experimental validation and visualization.

General preprocessing

TF-Prioritizer requires nf-core ChIP-seq / ATAC-seq [147] and nf-core RNA-seq [147] preprocessed data as input files (see Sec. 2.2.1/2.2.2 for information on the protocols and how this data is generated and processed and the official TF-Prioritizer GitHub repository for detailed formatting instructions) - more specifically, broad peaks (by MACS2) and gene counts (by Salmon). Once started, the pipeline downloads necessary additional data (e.g., gene lengths, gene symbols, and short descriptions of the genes) from biomaRt [210].

RNA-seq data filtering, normalization, and processing using DESeq2

TF-Prioritizer employs DESeq2 [198], a widely used method for determining differential gene expression from RNA-seq data. It starts by normalizing the raw read counts to account for differences in library size or sequencing depth across samples since, uncorrected, these differences can confound the differential expression analysis. DESeq2 uses a so-called "median of ratios" normalization that (i) calculates the ratio of each gene’s read count to the mean (i.e., average) of the gene's read counts across all samples generating a scaling factor for each gene in each sample, (ii) takes the median of these scaling factors from all genes as the size factor for that sample, and (iii) divide the raw counts by the size factor in each sample resulting in normalized counts which are comparable across all samples [211]. The user can decide to use a transcript per million (TPM; default set to 1.0; see below) filter or a gene count filter to filter noise (i.e., uninformative small changes or underexpressed genes, which could affect downstream analyses). The TPM is calculated by (i) determining the raw read count for each gene, (ii) dividing each gene's raw read count by its length in kilobases = "reads per kilobase / RPK", (iii) sum RPKs across all genes, (iv) divide each RPK by the sum. Next, DESeq2 determines the log2 fold change (log2fc), which represents the effect size, i.e., the magnitude of differential expression for each unfiltered gene. The log2fc is a measure of the proportional change in expression levels, where a value of 1 would indicate a doubling in expression, and -1 would indicate a halving. TF-Prioritizer also allows for batch effect correction in DESeq2 (see Sec. 2.2.2) [198].

ATAC-seq or DNase-seq data processing using HINT-ATAC

If ATAC-seq or DNase-seq is used as an input, TF-Prioritizer employs the method HINT-ATAC (see below) [194–196] to process the peaks to account for the different biochemical nature of the protocols (see Sec. 2.2.1). HINT-ATAC incorporates various aspects of data generation procedures, such as nucleosome positioning, transcription factor footprinting, and intrinsic sequencing biases. The tool rectifies these biases by utilizing a heuristic scoring system that calculates the probability of a read originating from a particular genomic feature like a nucleosome or a transcription factor binding site. The scoring system takes into consideration factors such as the strength of the signal, the distribution of fragment length, and the relative positioning of reads surrounding potential binding sites [194–196].

Blacklisted regions and combining samples

TF-Prioritizer preprocesses the broad peaks (originating either from ChIP-seq or protocol bias-corrected from ATAC-seq or DNase-seq) by filtering blacklisted regions identified by ENCODE (Encyclopedia of DNA Elements) [199]. Blacklisted regions frequently correspond to areas known for genome assembly complications, such as satellite DNA (i.e., repeat-rich regions ranging from a few bps to hundreds of bps in length, occurring in tandem arrays spanning up to millions of bps; [212]). This includes repeat-rich regions, such as centromeres (i.e., active during cell division [213]) and telomeres (i.e., protective ends of chromosomes [213]), and regions with abnormal GC content or sequence-specific biases, which can skew the distribution of sequencing reads. Additionally, blacklisted regions also contain whole large DNA segment duplications [199]. After filtering the blacklisted regions, I recommend using the sample combination option to combine broad peak samples of the same group into one peak file (i.e., overlapping regions will be joint), as the total runtime of the pipeline is reduced significantly without losing information.

Preselection of condition-relevant transcription factors using TEPIC, the peak-valley-peak model, TRAP, and DYNAMITE

TF-Prioritizer utilizes the tools and models (i) TEPIC to find TF binding sites using (ii) the peak-valley-peak model to guide the TF binding site search of (iii) TRAP, and (iv) DYNAMITE for preselection of condition-relevant TFs. (i) TEPIC [200,201] requires two primary inputs - genome-wide chromatin accessibility data (i.e., sourced from bias-corrected and blacklist-filtered ATAC-seq/DNase-seq data or blacklist-filtered ChIP-seq data) and gene expression data from RNA-seq. The first step of the workflow involves TEPIC processing the chromatin accessibility data and forming a list of accessible genomic regions, which manifest in peaks representing open chromatin regions and are thus accessible for TFs to bind and are possible locations for CREs. Then, TEPIC filters the ChIP-seq data for open chromatin regions that satisfy the (ii) peak-valley-peak model (i.e., multiple consecutive peaks that include maximum 500 bps span valleys where only a few reads could be mapped) to find potential CREs since CREs are especially enriched between HM peaks [202]. In the case of ATAC-seq or DNase-seq data, the peak-valley-peak model filter is skipped since these protocols cannot capture such short valleys and will instead use the whole open chromatin region as a search space. Next, TEPIC uses the proximity of CREs to target genes combined with an exponential decay model (i.e., estimating the influence of CREs on target genes, which decays exponentially as the distance between the CREs and the target gene increases; see below for a more detailed description) to associate potential CREs to target genes. At this stage, TEPIC employs (iii) TRAP to calculate the affinity of all known TFs (that pass the TPM filter similarly to the target genes that was employed in the DESeq2 step) for the identified open chromatin regions [203]. TRAP uses a combination of position weight matrices (PWMs) and a biophysical model, representing the DNA-binding preference of each TF, to compute TF affinities to positions inside the potential CREs. The PWM is retrieved from experimentally identified TF binding sites (e.g., TF ChIP-seq, see Sec. 2.2.1) and known TF sequence binding motifs (i.e., recurring patterns in DNA that are recognized and bound by a specific TF). Each entry in the PWM signifies the contribution of a specific base at a distinct position within the TF binding site to the overall binding energy. Consequently, a higher PWM value for a certain base at a specific location implies stronger, more favorable TF-DNA binding [203]. Then TEPIC calculates a TF-Gene score using the weighted sum of the TF affinities for all accessible chromatin regions associated with a given gene. The weight in this context is derived from an exponential decay function, where the decay factor is the genomic distance between the gene and the chromatin region. The decay function ensures that regions closer to the gene are given a higher weight, reflecting their potential higher influence on gene regulation. The final step of TEPIC's workflow involves the application of (iv) DYNAMITE [201], a regression-based method used to compute regressions between chromatin accessibility and TF affinity data with the gene expression data. DYNAMITE first constructs a supervised machine learning regression model (see Sec. 2.3.5) to capture the relationships between chromatin accessibility and TF affinities for each gene and their corresponding gene expression levels. DYNAMITE uses as independent variables the open chromatin regions, transcription factor affinities, and the TF-gene score. The TF-Gene score $a_w(g,t)$ for a gene g and a TF t in window size w is according to the description by Schmidt et al. [214], calculated as in Equation 1.

Equation 1: Calculation of the TF-Gene score

$$a_w(g,t) = \sum_{p \in P_{g,w}} \frac{a_{p,t}}{|p| - l} e^{-\frac{d_{p,g}}{d_0}}$$

“In Equation 1, $a_{p,t}$ is the affinity of TF t in peak p. The set of peaks $P_{g,w}$ contains all open-chromatin peaks in a window of size w around the gene g. $d_{p,g}$ is the distance from the center of the peak p to the transcription start site of the gene g, and d_0 is a constant fixed at 50,000 bp [215]. The affinities are normalized by peak and motif length, where $|p|$ is the length of the peak p and l is the total length of the motif of TF t (see Schmidt et al. for more specific information on how the TF-Gene score is calculated [200,201,214])” [51]. The dependent variable is the gene expression. DYNAMITE then attempts to predict the gene expression based on the independent variables. DYNAMITE automatically adjusts the model's parameters during the training process to minimize the difference between the predicted and actual gene expression levels. This allows the model to 'learn' the relationship between chromatin state, TF binding, and gene expression. DYNAMITE uses cross-validation to prevent overfitting (see Sec. 2.3.3). After establishing a robust model, DYNAMITE ranks the TFs based on their impact on gene expression. This ranking is performed by evaluating the trained regression coefficients (the slope in the regression model, see Sec. 2.3.5) of the TFs in the regression model. TFs with larger coefficients (closer to -1 or +1) are ranked higher, as they have a stronger impact on gene expression, and hence, are likely to be key players in the regulatory networks [200,201].

Final prioritization using the Mann-Whitney U Test and a discounted cumulative gain function

In this section, I first introduce some preliminary general statistical concepts such as the concept of populations (i.e., a set encompassing all items of interest) and samples (i.e., a subset of a population). A distribution illustrates the spread of values for a particular population (i.e., how frequently each value occurs). In general, statistical tests assume that data follow a certain distribution, most commonly the 'normal' or bell-curve distribution (Figure 16). The normal distribution is characterized by displaying data in a symmetrical pattern around the mean value (i.e., calculated by adding all values in a population and then dividing by the total number of samples). The central point of the normal distribution, where it reaches its highest point, represents the mean, median (i.e., sort all values and pick the middle value), and mode (i.e., most frequently occurring value) values of the dataset (i.e., meaning that the most frequently occurring values cluster around the average). From the center in either direction, the curve begins to drop, i.e., these values are less likely to occur.

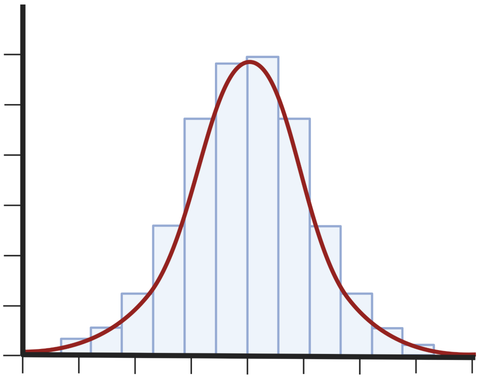


Figure 16: ***Schematic illustration of a bell curve, also known as a normal distribution.*** *The Figure was created with Biorender.com.*

The assumption of the distribution often informs the selection of an appropriate test. Parametric tests, for example, mandate data to be normally distributed. However, in real-world scenarios, data might not always conform to this normal distribution pattern. This lack of fit could be due to various reasons, such as outliers (i.e., samples that behave differently than the rest of the population). In these instances, where the data do not satisfy the requirements for a normal distribution, non-parametric tests like the Mann-Whitney U test are used since they do not require the data to be normally distributed [216,217].
The Mann-Whitney U test is a non-parametric statistical test that is used to determine if there are differences between two groups that are not normally distributed. The test begins by pooling the data from both groups, then ranking all values from both groups in ascending (i.e., lower value to the higher value) order. Once all data points are ranked, the ranks are then separated back into their original groups. Next, the sum of the ranks for each group separately is calculated. The sum of the ranks of both groups are then compared, and if they differ significantly (see below). In general, I use the default value of 0.05 and Equation 2 to determine if they differ significantly.

Equation 2: Mann-Whitney U test

$$U = n_1 * n_2 + \frac{(n_1(n_1+1))}{2} - R$$

In Equation 2, I depict n_1 as the size of the first group, n_2 as the size of the second group, and R as the sum of ranks in the first group. If U is then smaller than 0.05, I accept that the values do not come from the same population [204,205].

Generally, one could use the Mann-Whitney U test to assess the significance of a TF to the condition, e.g., on the influence on the expression of the target gene. Since such information can come from multiple sources (e.g., various HM ChIP-seq data), a systematic ranking algorithm, such as the discounted cumulative gain (DCG), is necessary. The main idea behind DCG is to assign higher importance to relevant observations appearing at the top of the list from multiple sources. DCG models the importance of observations based on their position in the lists from several sources under the assumption that items at the top of the list are more valuable [206]. In this thesis, I used a slightly modified DCG approach as follows. “Let $S(m)$ be the set of transcriptions factors t of an HM m such that the Mann-Whitney U test between the foreground distribution $FG(t, m)$ and the background distribution $BG(m)$

(see Hoffmann et al. integrated in this thesis [51]) yields a significant P-value (i.e., < 0.05). For a fixed TF $t \in S(m)$ and target genes TG, let $rank_m(t) = \sum_{t' \in S(m)} [\text{mean}_{g \in TG(t')} \omega_w(g, t') \leq \text{mean}_{g \in TG(t)} \omega_w(g, t)]$ be the rank of t in $S(m)$ w.r.t. the mean TF-TG scores (see Hoffmann et al. integrated in this thesis [51]) across all target genes, where $[\cdot]$ is the Iverson bracket, i.e., $[\texttt{true}] = 1$ and $[\texttt{false}] = 0$. I now compute an overall TF score $f(t)$ by aggregating the HM-specific ranks as in Equation 3.

Equation 3: Discounted cumulative gain

$$f(t) = \sum_{m \in HM(t)} 1 - \frac{rank_m(t)}{|S(m)|}$$

In Equation 3, $HM(t)$ denotes the set of histone modifications on strands of the DNA where the TF t can bind. Note that if $t \notin S(m)$, $\text{rank}_m(t)$ is not defined. In this case, I set $\text{rank}_m(t) = |S(m)|$ such that the summand for t equals 0. Lastly, I sort TFs in ascending order according to the scores $f(t)$" [51].

ChIP-Atlas and the Integrative Genome Browser as the source for automatic experimental validation and visualization

Computational predictions need to be validated experimentally. Since experiments in mice can take several years to conduct, already available condition-specific TF ChIP-seq data can speed up the process of validation. ChIP-ATLAS [197] is a comprehensive database that aggregates ChIP-seq data from various public sources and contains data for numerous transcription factors, histone modifications, and chromatin-associated proteins across different species. ChIP-ATLAS allows the automatic download and utilization of its data via an openly accessible API (Application Programming Interface, i.e., an algorithmic framework offering specific and easy-to-use ways a program can interact or communicate with another program). Automatically downloaded TF ChIP-seq data can then be visualized via the integrated genome viewer (IGV) [207–209]. TF ChIP-seq peaks that are visualized in close proximity to predicted TF binding sites can then validate the prediction and enhance the confidence.

3.2. circRNA-sponging framework

3.2.1. The current state of circRNA research

The study of circRNAs began in 1976 with the identification of circular RNA genomes in plant viroids [67]. In 1979, circRNA was discovered in the cytoplasm of eukaryotic cells [218], and by 1986, it was detected in the hepatitis delta virus [219]. The first human circRNA was identified in 1991 by Nigro et al. [220]. By 1995, research showed that circRNA had the ability to participate in protein synthesis in vitro [221]. Still, most functions were unknown, but in 2006, when RNase R treatment was discovered to enrich most circRNAs by degrading preferably linear RNAs, new opportunities to study circRNA became feasible [41]. The landscape of circRNA research expanded in 2012 with the advent of genome-wide profiling using RNA-Seq [222]. In 2013, circRNAs were first associated with miRNA sponging with the circRNAs CDR1as and Sry [48,223]. In 2015, researchers proposed circRNAs as potential cancer biomarkers and established their presence in exosomes [224]. Additionally, findings indicated that circRNAs exhibited long lifespans within cells due to their circular structure, meaning they could accumulate over time [40]. In general, detection and analysis of circRNAs is challenging on the (i) experimental and (ii) computational side due to the linear nature of data and the circular structure of circRNAs.

Capturing circRNAs from an experimental point of view during sequencing is challenging due to their unique covalently closed-loop structure without a 5' cap and without a 3' poly(A) tail, making sequencing them with traditional methods such as poly-A enrichment impossible (see Sec. 2.2.2; [225]). The currently predominantly used way to capture circRNA sequences is by using totalRNA-seq data or rRNA depleted data (see Sec. 2.2.2). Using reads from these sequencing protocols allows for detection of the BSJs occurring during circRNA synthesis (see Sec. 2.1.4). Since this BSJ is a very short region, long-read sequencing methods could play a crucial role in recent circRNA research. Long-read sequencing methods (see Sec. 2.2) can sequence entire RNA molecules (i.e., the end-to-end sequences of circRNAs), including their BSJs, providing a clearer picture of the expression of circRNAs [226]. For further investigation of circRNAs, RNase R treatment is an option since RNase R is an exonuclease that degrades linear RNAs, but it is ineffective against circular RNAs due to their unique structure. By treating a total RNA sample with RNase R, linear RNAs can be depleted, leaving circular RNAs for greater clarity. [227]

On the computational front, currently, the only possibility to detect circRNAs is through the detection of a BSJ in unmappable reads during the alignment process (see Sec. 2.2.2 “Aligning reads using STAR”). However, by focusing on the back-splicing junction alone, the expression of circRNAs in relation to their linear counterparts is typically underestimated [50]. The introduction of psirc-quant (see Sec. 3.2.2; [50]) has enabled researchers to quantify estimated genuine circRNA expression levels in relation to their linear counterparts, providing a more accurate representation of circRNA levels in biological samples.

Detection of circRNAs

A plethora of computational methods that can detect circRNAs have emerged. One could subdivide the methods into three core types: those based on BSJs, ones driven by machine learning, and integrated tools that use more than one prediction tool to minimize false positive circRNAs.

The BSJ-based tools like, but not limited to, Find_circ [223], CIRI [228–230], and CIRCexplorer (see Sec. 3.2.2 “Blacksplicing junction identification and circRNA detection using a combination of STAR and CIRCexplorer2”) [231,232] primarily identify circRNAs by recognizing the molecular signature known as the BSJ read. Each of these methods first identifies reads that cannot be aligned to the genome or transcriptome. Find_circ uses the bowtie approach (an approach similar to the BWA; see Sec. 2.2.1 “Burrows-Wheeler Aligner”) to identify reads that cannot be aligned. It then splits each of these unmapped reads into two so-called anchor segments. This segmentation is based on the understanding that if a read comes from a BSJ, one anchor should align upstream, while the other should align downstream, contrary to the standard genomic order. Following this, the anchor segments undergo another round of alignment to the genome. At this stage, Find_circ retains only those reads where both segments align uniquely to the genome but in reverse orientation (i.e., back-splicing). Find_circ then counts these circRNA and reports them [223]. Initially, CIRI [228–230] relies, in contrast to Find_circ, on the BWA-MEM algorithm (see Sec. 2.2.1 “Burrows-Wheeler Aligner”) to align reads to the genome or transcriptome. CIRI also searches for BSJ by identifying signals where the end of a read aligns to a downstream genomic location while the other end aligns upstream. In difference to Find_circ, CIRI additionally checks for the presence of reverse complementary matches around the junction sites that typically exist in circRNA structures. Finally, CIRI reports the counts of the circRNA as a result [228–230].

With the emergence of artificial intelligence and machine learning, tools for circRNA detection via these techniques became available. Machine learning tools for circRNA detection often use features like ALU repeats, structural motifs, and sequence motifs [44,233]. Examples of machine learning-based circRNA detection tools include PredcircRNA [233], WebCircRNA [234], PredicircRNATool [235], and DeepCirCode [236].

Integrated tools for identifying circRNAs have emerged as promising solutions by combining the features of multiple existing stable tools. This integration has been shown to minimize false-positive identifications of circRNAs [52,237–239]. A wide range of circRNA detection tools incorporating multiple BSJ-based detection tools have emerged, examples are CirComPara [240], circ_battle [238], RAISE [241], PcircRNA_finder [242], and CircRNAwrap [243].

While the BSJ-based computational detection cannot fully capture circRNA expression values, many detection tools use a method based on the ratio of BSJ read to regular splicing junction read to quantify circRNA expression. This method, known as the circular-to-linear ratio (CLR), helps in approximating the overall expression value of circRNA in comparison to linear RNA [244] (e.g., psirc-quant (see Sec. 3.2.2; [50])) [52].

circRNA annotation databases

Several databases are tailored for circRNA. For instance, circBase [245] features animal circRNAs along with their sequences and genomic coordinates (see Sec. 3.2.2 “Annotation of circRNAs using circBase”). CircFunBase [246] offers manually curated circRNAs, while CIRCpedia [232,247] stands out due to its extensive collection of circRNA annotations and expression profiles from six species, spanning various cell types and tissues. CircRNADb [248] is another inclusive resource focusing on human circRNA, particularly those with protein-coding potential, extracting annotations directly from published literature. Other notable databases include CircBank [249], which emphasizes human circRNA with a proposed new standard nomenclature, PigcirNet [250] with pig circRNA data, and AtCircDB [251], which focuses on tissue-specific Arabidopsis circRNAs. Plant-focused databases, such as PlantcircBase [252], PlantCircNet [253], and CropCircDB [254], offer a range of data from circRNA locations to interactions and associations with stress conditions in crops [52].

However, to the best of our knowledge, no end-to-end pipeline is known to investigate the sponging effects of circRNA in an automated way. Hence, I introduce circRNA-sponging [53], a pipeline that utilizes a combination of tools to elaborate on the potential sponging function of circRNAs.

3.2.2. Integrated tools and techniques

In this section, I generally elaborate on methods that are integrated into the circRNA-sponging framework: (i) blacksplicing junction identification and circRNA detection using a combination of STAR [157] and CIRCexplorer2 [231,232]; (ii) Quantification of raw circRNA expression levels retrieved from backsplicing junctions using psirc-quant [50] and kallisto [255]; (iii) Normalization of quantified circRNA expression, (iv) annotation of circRNAs using circBase; (v) Majority vote using three circRNA-miRNA target site prediction tools (miRanda [256], PITA [257], and TarPmiR [258]); (vi) Alternative splicing analysis of circRNAs using SUPPA2 [259]; (vii) miRDeep2 [260] for miRNA detection; (viii) normalization of miRNA expression [198]; and (ix) construction and analysis of a ceRNA network using SPONGE and spongEffects [27,28,54].

Blacksplicing junction identification and circRNA detection using a combination of STAR and CIRCexplorer2

STAR can be used to align reads to the genome (see Sec. 2.2.2, “Aligning reads using STAR”). However, not all reads, such as chimeric reads (i.e., reads that span a splicing junction, connecting two exons that are not consecutively located on the reference genome), can be mapped. These unmapped reads can represent circRNAs, and the method CIRCexplorer2 [231,232] can be used to identify if they represent circRNAs. CIRCexplorer2 utilizes the chimeric junction reads and then distinguishes backsplicing junctions by a 'head-to-tail' alignment, where the downstream (3') end of an exon is connected to the upstream (5') end of either the same exon or an upstream exon. The identified circRNAs are then annotated based on the reference genome (i.e., providing the chromosomal start and end positions) [231,232].

Quantification of raw circRNA expression levels retrieved from backsplicing junctions using psirc-quant and kallisto

Since CIRCexplorer2 can estimate circRNA abundance on back-splicing junction solely, the expression of circRNAs in relation to their linear counterparts is typically underestimated and needs to be quantified [50]. For this quantification, psirc-quant [50] was developed. psirc-quant utilizes kallisto [255], a tool similar to Salmon (see Sec. 2.2.2; [158]), with a slightly different algorithmic approach for quantifying gene expression from linear RNA-seq data to quantify circRNA expression. psirc-quant uses the backsplicing junctions identified by CIRCexplorer2 as pseudo transcripts, which are then used to construct a pseudo-transcriptome (i.e., each pseudo-transcript corresponds to a unique backsplicing junction by extracting the sequences around the backsplicing junction and concatenating them in a 'head-to-tail' arrangement to simulate the circular structure). psirc-quant uses the pseudo-transcriptome and the original RNA-seq sequencing reads for kallisto to quantify the expression of the circRNA. [50]

Normalization of quantified circRNA expression

circRNA expression can be normalized across samples using DESeq2 [198] (see. Sec. 3.1.2 "RNA-seq data filtering, normalization, and processing using DESeq2").

Annotation of circRNAs using circBase

One can annotate the names of circRNAs identified by CIRCexplorer2 by using the circBase [245] database, which provides experimentally validated and annotated circRNAs. A successful annotation of computationally predicted circRNA to a previously experimentally validated circRNA can boost confidence in the existence of the circRNA. circBase provides detailed annotation information for each circRNA, including its genomic coordinates, the gene it originates from, the exonic composition, and the tissues or cell types in which it has been observed. For successful annotation, the circRNA-sponging pipeline matches the genomic coordinates of the identified backsplicing junctions from CIRCexplorer2 with the coordinates of circRNAs in the circBase database [245].

Majority vote using three circRNA-miRNA target site prediction tools (miRanda, PITA, and TarPmiR)

Generally speaking, a miRNA is able to bind to a target if there exists a certain complementarity of base pairing between the miRNA seed region (the 2-8 nucleotides at the 5' terminus) and the target. Past research on miRNA binding site prediction showed that this is a complex problem. Most prediction methods rely on (i) seed matching, (ii) free energy, (iii) site accessibility, (iv) target-site abundance, (v) conservation, and (vi) machine learning. (i) Seed matching is based on the complementarity of the miRNA seed region to the target RNA, which typically allows for Wobble pairing (i.e., some connections between target and seed sequence with lower binding affinity such as G-U) [105]. (ii) The free energy refers to the hybridization energy of the relevant binding site and is used for scoring the reliability of miRNA binding sites. (iii) Site accessibility ranks the effort of the miRNA-RISC duplex to bind to the target RNA in terms of energy. (iv) Target-site abundance refers to the lower bound of the number of binding sites found on the target RNA for miRNA. (v) Conservation is often used as a criterion because miRNA binding sites are known to be located in conserved regions of RNAs [261]. Lastly, recent algorithms incorporate (vi) machine learning (e.g., random forest) in an effort to enhance their prediction effectiveness based on experimentally known miRNA binding sites. In general, machine learning approaches provide pre-trained models that enable to find miRNA binding sites in other data sets.

To boost reliability, I predict circRNA-miRNA binding sites using a majority voting between miRanda [256], PITA [257], and TarPmiR [258] since each method has a distinct approach for predicting miRNA binding sites. I consider a circRNA-miRNA binding site as relevant if it is predicted by at least two out of the three methods. miRanda is arguably the simplest approach of the three, as it only considers seed matching, conservation, and free energy. The seed-matching part of miRanda can be fine-tuned to allow for wobble pairs. To increase confidence in our results, I restrict our analysis to the binding site predictions assigned with a score above the 25 percent quantile. PITA uses the approaches of seed matching, conservation, free energy, and additionally considers site accessibility and target-site abundance. TarPmiR combines the approaches of seed matching, conservation, free energy, site accessibility, target-site abundance, and further integrates machine learning, among other small differences (see Ding et al., [258]). According to Ding et al., the machine learning approach vastly improves the predictions of TarPmiR but limits this approach to species that the model has been trained on. For the rest of the ceRNAs (excluding the circRNAs), circRNA-sponging incorporates experimentally validated target sites from DIANA-LncBase v3 [262], miRTarBase [263,264], and a mix of predicted and experimental target sites from miRWalk3.0 [265].

Alternative splicing analysis of circRNAs using SUPPA2

Since circRNAs can be alternatively spliced, circRNA-sponging utilizes SUPPA2 [259], which can detect alternative splicing events in linear RNAs from RNA-seq data that was already quantified by Salmon (see Sec. 2.2.2; [158]) or kallisto (see Sec. 3.2.2 “Quantification of raw circRNA expression levels retrieved from backsplicing junctions using psirc-quant and kallisto”; [255]). SUPPA2 calculates the relative abundance of different transcript isoforms through Percent Spliced In (PSI) values (i.e., the abundance of the inclusion transcripts - a particular exon or segment is included as a part of the final processed mRNA - divided by the total abundance of both the inclusion and exclusion transcripts - a particular exon or segment is excluded as part of the final process mRNA [266]). PSI values range from 0 to 1. A PSI value of 0 indicates that all transcripts correspond to the exclusion isoform, implying that the exon or intron involved in the splicing event is always excluded. Conversely, a PSI value of 1 suggests that all transcripts correspond to the inclusion isoform, denoting that the exon or intron is always included. A PSI value somewhere between 0 and 1 indicates that both inclusion and exclusion isoforms are present, with the value providing an estimate of their relative proportions [266]. The initial step in SUPPA2 involves the generation of a set of potential AS events based on a transcriptome (e.g., circRNAs can be included in a pseudo-transcriptome - see Sec. 3.2.2 “Quantification of raw circRNA expression levels retrieved from backsplicing junctions using psirc-quant and kallisto”). The events considered include exon skipping, alternative 5’ or 3’ splice sites, intron retention, and mutually exclusive exons (see Sec. 2.1.2) for which SUPPA2 calculated the PSI value.

miRDeep2 for miRNA detection

miRDeep2 [260] is an experimentally validated tool that can identify known and novel miRNAs from small RNA sequencing (miRNA-seq; see Sec. 2.2.3) data. Initially, the adaptors of the raw sequencing reads are removed to ensure adapter-free reads are used for miRNA prediction (similar to Sec. 2.2.1 “Adaptor trimming with TrimGalore!”). Following this, the cleaned reads are mapped onto a reference genome using an alignment tool called bowtie [267], which has similar functionality as STAR (see Sec. 2.2.2) but is optimized for smaller reads due to utilization of the BWT (see Sec. 2.2.1 “Burrows-Wheeler Aligner”) to locate the genomic origin of the sequenced reads. Once mapped, the reads that align to the reference genome are categorized by miRDeep2 into distinct genomic regions. The reads are segregated into those that align to known miRNAs, other non-coding RNAs (such as siRNAs; Figure 7), or unannotated regions of the genome to filter out sequences that originate from other types of small RNAs and not miRNAs.

Normalization of miRNA expression

miRNA expression can be normalized across samples using DESeq2 [198] (see. Sec. 3.1.2 “RNA-seq data filtering, normalization, and processing using DESeq2”).

Construction and analysis of ceRNA networks with SPONGE and spongEffects

SPONGE is a method for constructing genome-wide ceRNA networks (i.e., a competing endogenous RNA network, where a ceRNA is any RNA that carries miRNA binding sites, e.g., mRNAs, circRNAs, pseudogenes, transcripts of 3’ untranslated regions, and lncRNAs [27,32]). SPONGE uses a ’multiple sensitivity correlation’, a partial correlation-derived measure that quantifies the extent of ceRNA competition with respect to one or several shared miRNAs. In the case of ceRNA interactions, the random variables would be the expression levels of the ceRNAs, and the controlling variables would be the expression levels of the miRNAs (for more details, see List et al., [27]). SPONGE can estimate the distribution of this correlation value under the null hypothesis that no miRNA competition takes place, and this allows inferring p-values for the statistical significance of ceRNA interactions. From this, SPONGE then constructs a ceRNA network, where ceRNAs are represented as nodes and miRNAs as edges between the ceRNAs. This leads to an enormous network that is difficult to grasp. SPONGE has been applied to TCGA cancer types and offers insights into the ceRNA regulatory landscape in various cancers [28]. To gain insights into the global ceRNA network created by SPONGE, circRNA-sponging applies spongEffects [54]. Global ceRNA networks give an overview of ceRNA interactions but do not offer sample- or patient-specific insights. spongEffects extracts ceRNA modules (i.e., a significant central ceRNA with all its directly connected ceRNA neighbors) using a random forest (see Sec. 2.3.6) approach from previously inferred ceRNA networks that are predictive between the groups. spongEffects is capable of identifying biomarkers between conditions (for more details, see Boniolo and Hoffmann et al., [54]).

3.3. Further tools

I used the text editor google docs (https://docs.google.com/) with the plugins Paperpile (citation manager, https://paperpile.com/), Auto-LaTeX Equations (equation manager, https://www.autolatex.com/), and Cross References (cross reference manager, https://github.com/davidrthorn/cross_reference) to produce this text document. Furthermore, figures were mainly created using Biorender (https://www.biorender.com/); in which some of the figures include icons from Flaticon (https://www.flaticon.com/) under the paid premium license. While preparing this thesis, I utilized several tools to improve the quality of the text. I used the paid version of ChatGPT by OpenAI (https://chat.openai.com/) as well as the google docs add-on Grammarly (https://app.grammarly.com/) with its paid version. ChatGPT, an AI model based on GPT-4 architecture, served as a resource in refining the language and enhancing the readability of the thesis. Grammarly, a digital writing assistant, was relied upon for its proficiency in detecting and correcting spelling, comma, and grammatical errors. The AI DeepL (https://www.deepl.com/) was used for the translation of the abstract from English to German. These tools and artificial intelligence methods, however, were supplementary to my intellectual contributions. These tools were employed to refine the expression of ideas, not to replace the analytical and critical thinking involved in the execution of the analyses or the formulation of this thesis.

4. PUBLICATIONS

4.1. Publication 1: TF-Prioritizer: a java pipeline to prioritize condition-specific transcription factors

Citation

The article titled "TF-Prioritizer: a java pipeline to prioritize condition-specific transcription factors" was published online at Oxford University Press GigaScience on 03 May 2023.

Full citation:

Hoffmann, M., Trummer, N., Schwartz, L., Jankowski, J., Lee, H. K., Willruth, L.-L., Lazareva, O., Yuan, K., Baumgarten, N., Schmidt, F., Baumbach, J., Schulz, M. H., Blumenthal, D. B., Hennighausen, L., & List, M. (2023). TF-Prioritizer: a Java pipeline to prioritize condition-specific transcription factors. GigaScience, 12, giad026. PMCID: PMC10155229

Summary

This research focuses on understanding the regulation of eukaryotic gene expression, which is majorly controlled by cis-regulatory elements (CREs), such as promoters and enhancers, bound by transcription factors (TFs). The differential expression of TFs and their binding affinity at putative CREs determines tissue- and developmental-specific transcriptional activity. Consolidating genomic data sets can offer further insights into CRE accessibility, TF activity, and gene regulation. However, integrating and analyzing multi-modal data sets, including chromatin state data (ChIP-seq, ATAC-seq, or DNase-seq) and RNA-seq data, faces significant technical challenges. Existing methods that highlight differential TF activity from combined chromatin state and RNA-seq data suffer from limited usability, inadequate support for large-scale data processing, and minimal functionality for visually interpreting results. To overcome these limitations, I developed TF-Prioritizer, an automated pipeline that prioritizes condition-specific TFs from multi-modal data and generates an interactive web report. I demonstrated the potential of TF-Prioritizer by identifying known TFs and their target genes, as well as previously unreported TFs active in lactating mouse mammary glands. Furthermore, I analyzed a variety of ENCODE data sets for cell lines K562 and MCF-7, including twelve histone modification ChIP-seq as well as ATAC-seq and DNase-seq datasets. Through this analysis, I observed and discussed assay-specific differences. In conclusion, TF-Prioritizer accepts ATAC-seq, DNase-seq, or ChIP-seq and RNA-seq data as input and identifies TFs with differential activity. This offers an improved understanding of genome-wide gene regulation, potential pathogenesis, and therapeutic targets in biomedical research. Our pipeline addresses the limitations of existing methods, providing enhanced usability, support for large-scale data processing, and visually interpretable results.

Availability

TF-Prioritizer is maintained and available as a dockerized Java application at GitHub: https://github.com/biomedbigdata/TF-Prioritizer.

The data used in this analysis is freely available at GEO under the IDs GSE161620, GSE82275, GSE84115, GSE37646, and https://www.encodeproject.org.

Contribution

I was actively involved in the planning and development phase of TF-Prioritizer. Furthermore, I implemented the first prototype myself. My responsibilities included supervising Nico Trummer and Leon Schwartz during the development of the web application and the integration of ATAC-seq and DNase-seq data. I played a crucial role in ensuring that the project proceeded smoothly and according to the proposed timeline. Additionally, I took charge of designing the computational analyses and drafting the manuscript, focusing on presenting the research and findings in an understandable and appropriate manner. Following the submission of the manuscript, I addressed the reviewers' comments and revised the text and analyses accordingly, ensuring that the final version satisfied the reviewers' demands. Lastly, I managed the entire submission process, navigating the various requirements and deadlines associated with academic publishing.

Rights and permissions

The original article is embedded under the open-access license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/) by Oxford Academic Press. All rights belong to Oxford Academic Press.

Additional supplementary material

Supplementary data are available at GigaScience online and GigaScience db: Markus, H., Nico, T., Leon, S., Jakub, J., Hye, L. K., Lina-Liv, W., Olga, L., Kevin, Y., Nina, B., Florian, S., Jan, B., Marcel, S. H., David, B. B., Lothar, H., & Markus, L. (2023). Supporting data for “TF-Prioritizer: a java pipeline to prioritize condition-specific transcription factors”. GigaScience Database. https://doi.org/10.5524/102379

OXFORD
(GIGA)^n SCIENCE

GigaScience, 2023, **12**, 1–15
DOI: 10.1093/gigascience/giad026
TECH NOTE

TF-Prioritizer: a Java pipeline to prioritize condition-specific transcription factors

Markus Hoffmann[1,2,3,*,†], Nico Trummer[1,†], Leon Schwartz[1], Jakub Jankowski[3], Hye Kyung Lee[3], Lina-Liv Willruth[1], Olga Lazareva[4,5,6], Kevin Yuan[7], Nina Baumgarten[8,9,10], Florian Schmidt[11], Jan Baumbach[12,13], Marcel H. Schulz[8,9,10], David B. Blumenthal[14], Lothar Hennighausen[2,3,‡] and Markus List[1,*,‡]

[1]Big Data in BioMedicine Group, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising D-85354, Germany
[2]Institute for Advanced Study, Technical University of Munich, Garching D-85748, Germany
[3]National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
[4]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[5]Junior Clinical Cooperation Unit, Multiparametric Methods for Early Detection of Prostate Cancer, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[6]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany
[7]Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
[8]Institute of Cardiovascular Regeneration, Goethe University, 60590 Frankfurt am Main, Germany
[9]German Center for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany
[10]Cardio-Pulmonary Institute, Goethe University Hospital, 60590 Frankfurt am Main, Germany
[11]Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore
[12]Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany
[13]Computational BioMedicine Lab, University of Southern Denmark, Odense, Denmark
[14]Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
*__Correspondence address__. Markus Hoffmann, Lehrstuhl für Experimentelle Bioinformatik Maximus-von-Imhof-Forum 3 85354 Freising Germany. E-mail: markus.daniel.hoffmann@tum.de; Markus List, Lehrstuhl für Experimentelle Bioinformatik Maximus-von-Imhof-Forum 3 85354 Freising Germany. E-mail: markus.list@tum.de.
†The first two authors should be considered shared first authors.
‡The last two authors should be considered shared last authors.

Abstract

Background: Eukaryotic gene expression is controlled by *cis*-regulatory elements (CREs), including promoters and enhancers, which are bound by transcription factors (TFs). Differential expression of TFs and their binding affinity at putative CREs determine tissue- and developmental-specific transcriptional activity. Consolidating genomic datasets can offer further insights into the accessibility of CREs, TF activity, and, thus, gene regulation. However, the integration and analysis of multimodal datasets are hampered by considerable technical challenges. While methods for highlighting differential TF activity from combined chromatin state data (e.g., chromatin immunoprecipitation [ChIP], ATAC, or DNase sequencing) and RNA sequencing data exist, they do not offer convenient usability, have limited support for large-scale data processing, and provide only minimal functionality for visually interpreting results.

Results: We developed TF-Prioritizer, an automated pipeline that prioritizes condition-specific TFs from multimodal data and generates an interactive web report. We demonstrated its potential by identifying known TFs along with their target genes, as well as previously unreported TFs active in lactating mouse mammary glands. Additionally, we studied a variety of ENCODE datasets for cell lines K562 and MCF-7, including 12 histone modification ChIP sequencing as well as ATAC and DNase sequencing datasets, where we observe and discuss assay-specific differences.

Conclusion: TF-Prioritizer accepts ATAC, DNase, or ChIP sequencing and RNA sequencing data as input and identifies TFs with differential activity, thus offering an understanding of genome-wide gene regulation, potential pathogenesis, and therapeutic targets in biomedical research.

Introduction

Understanding how genes are regulated remains a major research focus of molecular biology and genetics [1]. In eukaryotes, gene expression is controlled by *cis*-regulatory elements (CREs) such as promoters, enhancers, or suppressors, which are bound by transcription factors (TFs) promoting or repressing transcriptional activity depending on their accessibility [2]. TFs play an important role not only in development and physiology but also in diseases; for example, it is known that at least a third of all known human developmental disorders are associated with deregulated TF activity and mutations [3–5]. An in-depth investigation of TF regulation could help to gain deeper insights into the gene-regulatory balance found in normal physiology. Since most complex diseases involve aberrant gene regulation, a detailed understanding of this mechanism is a prerequisite to developing targeted therapies [6, 7]. This is a daunting task, as multiple genes in eukaryotic genomes may affect the disease, each of which is possibly controlled by candidate CREs.

TF chromatin immunoprecipitation sequencing (ChIP-seq) experiments are the gold standard for identifying and understand-

Received: October 25, 2022. **Revised:** February 23, 2023. **Accepted:** April 5, 2023

ing condition-specific TF binding at a nucleotide level. However, since there are approximately 1,500 active TFs in humans [8] and about 1,000 in mice [9], and additionally considering the need to establish TF patterns separately for each tissue and physiological condition, this approach is logistically prohibitive. Alternatively, histone modification (HM) ChIP-seq but also ATAC sequencing (ATAC-seq) and DNAse sequencing (DNAse-seq) offer a broader view of the chromatin state due to their individual capability (i.e., ChIP-seq identifies protein–DNA interactions, ATAC-seq detects open chromatin regions via Tn5 transposase cuttings, and DNAse-seq maps accessible chromatin sites by digesting chromatin with DNase I) to highlight open chromatin regions aligned with active genes, hence allowing the identification of condition-specific CREs [10]. Computational methods can then be used to prioritize TFs likely binding to these CREs, leading to hypotheses and defining the most promising TF ChIP-seq experiments. This narrows the scope of TF ChIP-seq experiments needed to confirm working hypotheses about gene regulation [11–13].

Several general approaches have been proposed to identify key TFs that are responsible for gene regulation. Among them, for example, is (i) a basic coexpression or mutual information analysis of TFs and their target genes combined with computational binding site predictions [14]. (ii) Some tools use a combination of TF ChIP-seq data—providing genome-wide information about the exact locations of TF binding—with predicted target genes that can enhance coexpression analyses [15]. (iii) Other tools employ a combination of genome-wide chromatin accessibility (e.g., HM ChIP-seq data) or activity information, putative TF binding sites, and gene expression data. This combination can be powerful in determining key TF players and is used by the state-of-the-art tool diffTF [16]. Most of the proposed approaches require substantial preprocessing, computational knowledge, adjustment of the method to a new use case (e.g., more than 2 conditions and/or time-series data), and manual evaluation of the results (e.g., manual search and visualization for TF ChIP-seq data to provide experimental evidence for the predictions). Hence, to streamline this process, we present TF-Prioritizer, a Java pipeline to prioritize TFs that show condition-specific changes in their activity. TF-Prioritizer falls into the third category of the previously described approaches and automates several time-consuming steps, including data processing, TF affinity analysis, machine learning predicting relationships of CREs to target genes, prioritization of relevant TFs, data visualization, and visual experimental validation of the findings using public TF ChIP-seq data (i.e., ChIP-Atlas [17]).

Figure 1 depicts a general overview of the pipeline. TF-Prioritizer accepts 2 types of input data: (i) histone modification peak ChIP-seq/ATAC-seq/DNase-seq data indicating accessible regulatory regions showing differential activity (peak data are typically generated by MACS2 [18]) and (ii) gene expression data from RNA-seq, which allows the identification of differentially expressed genes that are potentially regulated by TFs at specific time points or physiological condition. If peaks from ATAC-seq or DNase-seq were provided, we generate footprints (i.e., specific regions of the peaks within hypersensitive sites that could indicate the regulatory region of genes [19]) by employing HINT (i.e., HINT uses hidden Markov models to identify footprints by using strand-specific, nucleosome-sized signals with corrections for ATAC-seq and DNase-seq protocol-specific biases to successfully target CREs) for further processing [20–22]. Our pipeline searches for TF binding sites using TRAP [23] within CREs around accessible genes and calculates an affinity score for each known TF to bind at these particular loci using TEPIC [24,25]. TEPIC uses an exponential decay model that was built under the assumption that regulatory elements close to a gene are more likely important than more distal elements and weighs this relationship accordingly. This allows us to assess TF binding site specific probabilities by using TF binding affinities calculated by TRAP, which uses a biophysical model to assess the strength of the binding energy of a TF to a CRE's total sequence [23]. Beginning with these CRE candidates, we search for links to possible regulated putative target genes that are differentially expressed between given conditions (e.g., disease and healthy). Approaching the task of linking CREs to target genes, we employ the framework of TEPIC2 [25] and DYNAMITE [25] (feature comparison Supplementary Table S1), which uses a logistic regression model predicting differentially expressed genes across time points and conditions based on TF binding site information to score different TFs according to their contribution to the model and their expression (for a more technical description, see "Technical workflow" section). In general, TF-Prioritizer uses TEPIC and DYNAMITE pairwise of the provided data (i.e., pairwise for each condition and each time point). Based on a background distribution of the scores (combination of differential expression, TEPIC, and DYNAMITE—see "Discovering *cis*-regulatory elements using a biophysical model" section), TF-Prioritizer computes an empirical *P* value reflecting the significance of the results (see "An aggregated score to quantify the contribution of a TF to gene regulation" section). TF-Prioritizer offers automated access to complementary ChIP-seq data of the prioritized TFs in ChIP-Atlas [17] for validation and shows predicted regulatory regions of target genes using the Integrative Genomics Viewer (IGV) [27–29]. Then, TF-Prioritizer automatically generates a user-friendly and feature-rich web application that could also be used to publish the results as an online interactive report.

To demonstrate the potential and usability of TF-Prioritizer, we use genomic data describing mammary glands in pregnant and lactating mice and compare our analysis to established knowledge [30]. Employing the web application generated by TF-Prioritizer, we found well-studied TFs involved in the mammary gland development process, and we identified additional TFs, which are candidate key factors in mammary gland physiology. Additionally, we use ENCODE cell line data (K562 and MCF-7) to demonstrate the potential and usability of TF-Prioritizer using ATAC-seq, DNase-seq, and HM ChIP-seq data.

Materials and Methods

Implementation

The main pipeline protocol is implemented in Java version 11.0.14 on a Linux system (Ubuntu 20.04.3 LTS). The pipeline uses subprograms written in Python version 3.8.5, R version 4.1.2, C++ version 9.4.0, and CMAKE (RRID:SCR_015875) version 3.16 or higher. External software that needs to be installed before using TF-Prioritizer can be found on GitHub (see Availability Section). We also provide a bash script "install.sh," that automatically downloads and installs necessary third-party software and R/Python packages. The web application uses Angular CLI version 14.0.1 and Node.js version 16.10.0. We also provide a dockerized version of the pipeline; it uses Docker version 20.10.12 and Docker-Compose version 1.29.2 (Availability Section). TF-Prioritizer is available as a docker that can be pulled from docker hub and GitHub packages ("Availability of source code and requirements" section).

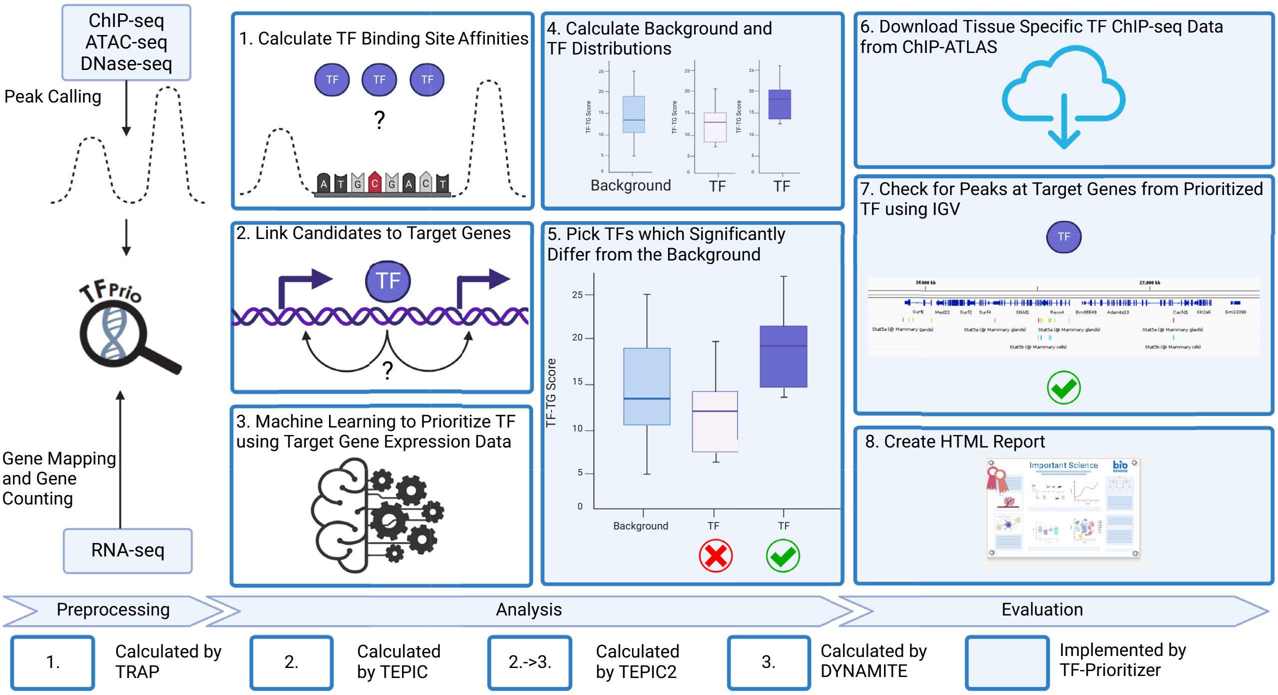

ChIP-seq
ATAC-seq
DNase-seq
Peak Calling
TFPrio
Gene Mapping and Gene Counting
RNA-seq
1. Calculate TF Binding Site Affinities
TF
?
A T G C G A C T
2. Link Candidates to Target Genes
TF
?
3. Machine Learning to Prioritize TF using Target Gene Expression Data
4. Calculate Background and TF Distributions
TF-TG Score
Background
TF
TF
5. Pick TFs which Significantly Differ from the Background
TF-TG Score
25
20
15
10
5
0
Background
TF
TF
6. Download Tissue Specific TF ChIP-seq Data from ChIP-ATLAS
7. Check for Peaks at Target Genes from Prioritized TF using IGV
TF
8. Create HTML Report
Important Science
Preprocessing
Analysis
Evaluation
1.
Calculated by TRAP
2.
Calculated by TEPIC
2.->3.
Calculated by TEPIC2
3.
Calculated by DYNAMITE
Implemented by TF-Prioritizer

Figure 1: General overview of the TF-Prioritizer pipeline. TF-Prioritizer uses peaks from ChIP-seq or ATAC-seq/DNase-seq and gene counts from RNA-seq. If peaks from the protocols ATAC-seq or DNase-seq were provided, we treat them by using the footprinting method HINT and use the footprints for further processing [20–22]. It then (1) calculates TF binding site affinities using the tool TRAP [23], (2) links candidate regions to potential target genes by employing TEPIC [24], (3) performs machine learning (by using the framework of TEPIC2 [25] and DYNAMITE) to find relationships between TFs and their target genes, (4) calculates background and TF distributions, (5) picks TFs that significantly differ from the background using the Mann–Whitney *U* test [26] and a comparison between the mean and the median of the background and TF distribution, (6) searches for tissue-specific TF ChIP-seq evaluation data in ChIP-ATLAS [17], (7) creates screenshots using the Integrative Genomics Viewer from predicted regions of interest [27–29], and (8) creates a feature-rich web application for researchers to share and evaluate their results.

Table 1: Overview of datasets covering mammary gland development from pregnancy to lactation

	p6	p13	L1	L10	Sum
ChIP-seq *H3K27ac*	3	1	8	4	16
ChIP-seq *H3K4me3*	2	3	5	0	10
ChIP-seq *Pol2*	2	0	5	4	11
RNA-seq	6	8	3	4	21

Data processing

Mammary gland development and lactation in mice

Datasets (GEO accession ID: GSE161620) are processed with the nf-core/RNA-seq [31] and nf-core/ChIP-seq pipelines in their default settings, respectively [32, 33]. The FASTQ files of pregnant and lactating mice are processed by Salmon (RRID:SCR_017036) [34] and MACS2 (RRID:SCR_013291) [35] to retrieve raw gene counts and broad peak files.

The dataset spans several time points in mammary gland development from pregnancy to lactation. For each stage, 2 distinct time points are available: pregnancy day 6 (p6), pregnancy day 13 (p13), lactation day 1 (L1), and lactation day 10 (L10). For each time point, the dataset contains RNA-seq data and ChIP-seq data for histone modifications H3K27ac and H3K4me3, as well as Pol2 ChIP-seq data (Table 1). We used H3K27ac, H3K4me3, and Pol2 data for creating the model.

Table 2: Overview of the dataset covering several HM ChIP-seq, ATAC-seq, DNase-seq, and RNA-seq for the cell lines K562 and MCF-7

Protocol		K562	MCF-7	Sum
ATAC-seq		4	1	5
DNase-seq		4	4	8
ChIP-seq	H3K27ac	1	2	3
	H3K27me3	2	2	4
	H3K36me3	2	2	4
	H3K4me3	4	2	6
	H3K9me3	1	2	3
	H2AFZ	1	1	2
	H3K4me1	2	1	3
	H3K4me2	1	1	2
	H3K79me2	1	1	2
	H3K9ac	2	1	3
	H4K20me1	1	1	2
RNA-seq		15	4	19

ENCODE cell lines

ATAC-seq, DNase-seq, ChIP-seq, and RNA-seq data are downloaded from the ENCODE project for the cell lines K562 (human chronic myelogenous leukemia cell line) and MCF-7 (human breast adenocarcinoma cell line), which are both often used to study cancer biology and have been subjected to a large number of different experimental protocols and assays (Table 2, file identifiers in Supplementary Material S1) [90].

Technical workflow

Preprocessing

TF-Prioritizer uses peak data from ChIP-seq, ATAC-seq, or DNase-seq and a gene count matrix from RNA-seq as input files (see GitHub repository for detailed formatting instructions). Initially, the pipeline downloads necessary data (gene lengths, gene symbols, and short descriptions of the genes) from BioMart (RRID: SCR_019214) [36]. Optionally, genes with low expression can be removed. TF-Prioritizer uses a transcripts per million (TPM) filter of 1 as default to remove TFs that show very low expression and are most probably not relevant. Subsequently, we use DESeq2 to normalize read counts and calculate the $\log_2$-fold change ($\log_2$fc) [37]. In parallel, TF-Prioritizer preprocesses the peaks by first employing HINT if the provided peak data are labeled as ATAC-seq or DNase-seq to perform footprinting to correct for the biases (i.e., by analyzing chromatin accessibility data in terms of histone modification state, enabling more accurate comparison between the 2 data types) between the ChIP-seq, ATAC-seq, and DNase-seq protocols [20, 38]. TF-Prioritizer then filters blacklisted regions that would likely lead to false positives [39]. Peak files from the same sample group can be merged to significantly reduce the total runtime of the pipeline without affecting the ability of the TF-Prioritizer to identify candidate CREs.

Discovering cis-regulatory elements using a biophysical model

TEPIC links CREs to target genes using a window-based approach (default: 50,000 bp) [24, 25] using TRAP, a biophysical model to quantify transcription factor affinity [23]. The window-based approach can be enhanced by providing Hi-C loop data, where the prediction window is extended or limited to a chromatin loop around potential CREs and target genes. TEPIC interprets ChIP-seq signal intensity as a quantitative measure of TF binding strength, which also helps in recovering low-affinity binding sites that would be missed in a classical presence/absence model [24]. The default TEPIC framework searches for dips on top of peaks. However, numerous studies have shown that CREs are often enriched between histone peaks (peak–dip–peak or peak–valley–peak model) [40]. To better accommodate histone modification of ChIP-seq data, we thus extended the TEPIC framework to search for transcription factor binding sites (TFBSs) between 2 peaks that have close (default 500 bp) genomic positions. TEPIC aggregates individual TF affinities into a TF-Gene score, which is the sum of the individual affinities normalized by the length of the considered CREs.

According to the description in Schmidt et al. [41], the TF-Gene score $a_w(g,t)$ for a gene g and a TF t in window size$w$ is calculated as in Equation 1:

Equation 1: calculation of the TF-Gene score

$$a_w(g,t) = \sum_{p \in P_{g,w}} \frac{a_{p,t}}{|p| - l} e^{-\frac{d_{p,g}}{d_0}} \tag{1}$$

In Equation 1, $a_{p,t}$ is the affinity of TF t in peak p. The set of peaks $P_{g,w}$ contains all open-chromatin peaks in a window of size w around the gene $g.d_{p,g}$ is the distance from the center of the peak p to the transcription start site of the gene g, and d_0 is a constant fixed at 50,000 bp [42]. The affinities are normalized by peak and motif length, where $|p|$ is the length of the peak p and l is the total length of the motif of TF t (see Schmidt et al. [24, 25, 41] for more specific information on how the TF-Gene score is calculated). Since proximal CREs are expected to have a larger influence on gene expression compared to distal ones, these contributions are weighted following an exponential decay function of genomic distance [25].

We want to point out that the biophysical model calculated by TRAP only returns the center of a potentially large area of high binding energy. The TF is supposed to bind somewhere in this area. In our IGV screenshot, the center of the high binding energy area can appear at a distance up to the window defined by TEPIC. We consider predicted TF peaks as matching if we find TF ChIP-seq peaks inside this window. Following this, we do not expect the predicted TF bindings to overlap exactly with the TF ChIP-seq peaks.

An aggregated score to quantify the contribution of a TF to gene regulation

To determine which TFs have a significant contribution to a condition-specific change between 2 sample groups, we want to consider multiple lines of evidence in an aggregated score. We introduce TF–target gene (TG) scores (Fig. 2) which combine (i) the absolute $\log_2$-fold change of differentially expressed genes since genes showing large expression differences are more likely affected through TF regulation than genes showing only minor expression differences and (ii) the TF-Gene scores from TEPIC indicating which TFs likely influence a gene. To further quantify this link, we also consider the total coefficients of a logistic regression model computed with DYNAMITE [25]. DYNAMITE predicts (high/low) expression of a gene based on the fold changes of TF-Gene scores reported by TEPIC and thus helps to prioritize among multiple potential TFs regulating a gene. We calculate TF-TG scores (ω) for each time point and each type of ChIP-seq data (e.g., different histone modifications) as in Equation 2:

Equation 2: Calculation of the TF-TG score ω for each time point and each type of ChIP-seq data:

$$\omega_w(g,t) = |\log_2(fc(g))| \cdot a_w(g,t) \cdot |\eta(g,t)|, \tag{2}$$

where $fc(g)$ represents the fold change of the target gene g between the 2 conditions, $a_w(g,t)$ is the TF-Gene score retrieved by TEPIC as detailed above, and $\eta(g,t)$ is the total regression coefficient of DYNAMITE's linear model of the expression of the target gene g as a function of the expression of the TF t.

A random background distribution allows TF-Prioritzier to exclude spurious results

The ultimate goal of TF-Prioritizer is to identify those TFs that are most likely involved in regulating condition-specific genes. To judge if a specific TF-TG score is meaningful, we generate a background distribution under the hypothesis that most TFs will not be condition specific. Therefore, we generate 2 different kinds of distributions (see Fig. 2): (i) for each HM m, a background distribution containing all positive TF-TG scores associated with m: $BG(m) = \{\omega_w(g,t) \mid t \in TF(m), g \in TG(t), \omega_w(g,t) > 0\}$. Here, $TF(m)$ denotes the set of TFs that can bind to strands of the DNA modified by m, and $TG(t)$ is the set of target genes of the TF t. (ii) For each HM-TF pair (m,t) with $t \in TF(m)$, a foreground distribution containing all positive TF-TG scores associated with (m,t): $FG(t,m) = \{\omega_w(g,t) \mid g \in TG(t), \omega_w(g,t) > 0\}$. Note that $FG(t,m) \subseteq BG(m)$ holds for all HM-TF pairs (m,t). We then test each TF distribution of each ChIP-seq against the global distribution matching the ChIP-seq data type. If the P value of a Mann–Whitney U (MWU) test [43] is below the threshold (default: 0.05) and the median and mean of the TF are higher than the background distribution, the TF is recognized as a potential candidate. In the last step, we sort the TFs based on the mean of the TF-TG scores and report the ranks.

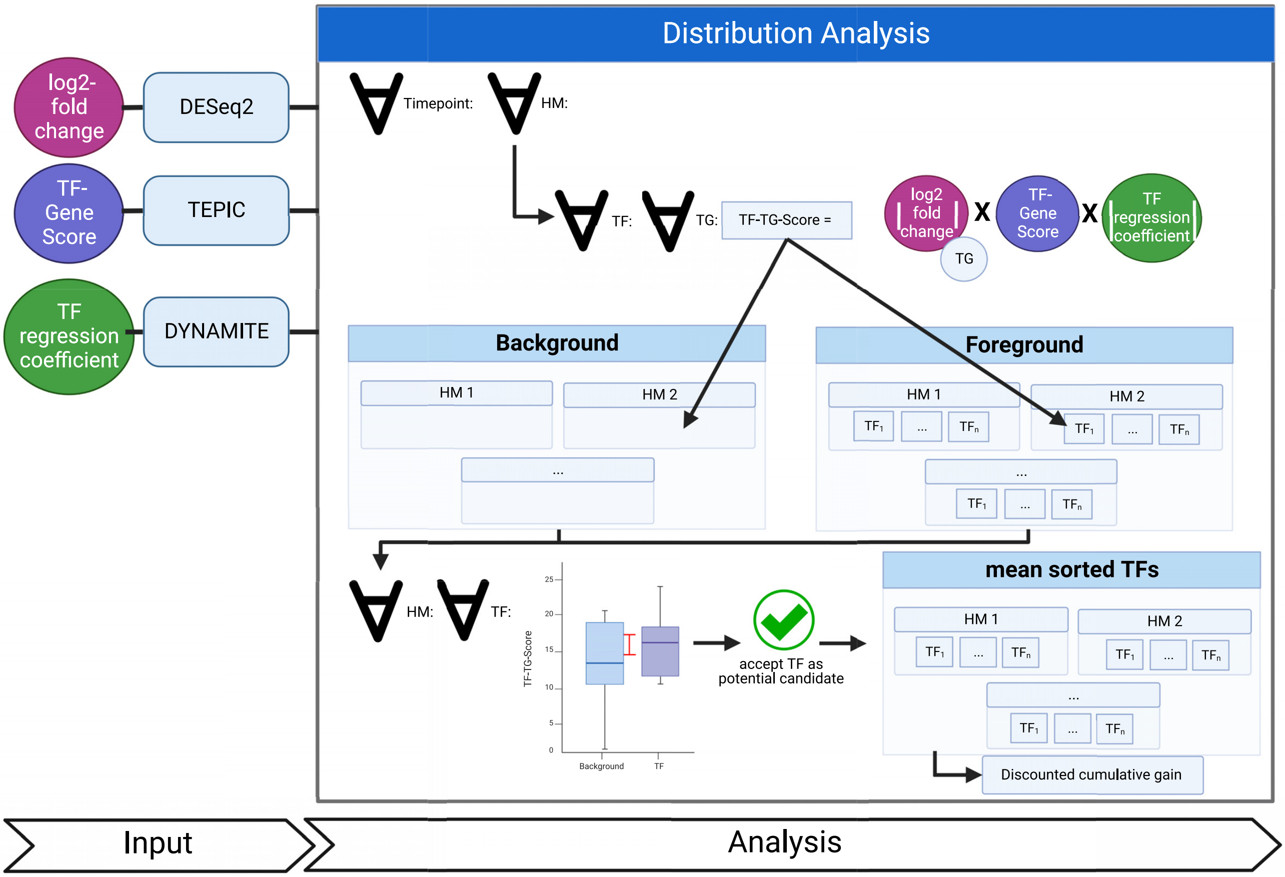


Figure 2: Workflow of the distribution analysis to prioritize TFs in a global context by using TF-TG scores. We use several scores conducted by previously performed analysis (see Supplementary Fig. S1), specifically the total $\log_2$-fold change (DESeq2), the TF-Gene score (TEPIC), and the total TF regression coefficient (DYNAMITE). We then calculate the TF-TG score for each time point for each TF on each of the TF predicted target genes (TG) and save it to separate files for the background of each histone modification and for each TF in each histone modification. In the next step, we perform a Mann–Whitney *U* [43] test between the distribution of the background of the histone modification and the distinct TF distribution of the same histone modification. If the TF passes the Mann–Whitney *U* test and the median and mean of the TF are higher than the background median and mean, we consider this TF as prioritized for the histone modification. We perform a discounted cumulative gain to receive one list with all prioritized TFs and overall histone modifications.

We obtain a global list of prioritized TFs across several ChIP-seq data types (e.g., different histone modifications) as follows:

Let $S(m)$ be the set of transcriptions factors t such that the 1-sided MWU test between the foreground distribution $FG(t, m)$ and the background distribution $BG(m)$ yields a significant P value. For a fixed TF $t \in S(m)$, let $rank_m(t) = \sum_{t' \in S(m)} [\text{mean}_{g \in TG(t')} \omega_w(g, t') \leq \text{mean}_{g \in TG(t)} \omega_w(g, t)]$ be the rank of t in $S(m)$ with respect to the mean TF-TG scores across all target genes, where $[\cdot]$ is the Iverson bracket (i.e., $[\text{true}] = 1$ and $[\text{false}] = 0$). We now compute an overall TF score $f(t)$ by aggregating the HM-specific ranks as follows in Equation 3:

$$f(t) = \sum_{m \in HM(t)} 1 - \frac{rank_m(t)}{|S(m)|}, \qquad (3)$$

where $HM(t)$ denotes the set of histone modifications on strands of the DNA where the TF t can bind. Note that if $t \notin S(m)$, $\text{rank}_m(t)$ is not defined. In this case, we set $\text{rank}_m(t) = |S(m)|$ such that the summand for t equals 0. Last, we sort TFs in ascending order according to the scores $f(t)$.

Discovering each score's contribution to the global score

To analyze the impact of the different parts of the TF-TG score, we permute its components (TF score from TEPIC, regression coefficient of DYNAMITE, $\log_2$fc of DESeq2). We execute TF-Prioritizer with the exact same configuration but with all possible combinations of the components and compare the prioritized TFs (e.g., solely TF score from TEPIC, a combination of TF score from TEPIC with the regression coefficient of DYNAMITE).

Validation using independent data from ChIP-Atlas

TF-Prioritizer is able to download and visualize experimental tissue-specific TF ChIP-seq data for prioritized TFs from ChIP-Atlas [17], a public database for ChIP-seq, ATAC-seq, DNase-seq, and Bisulfite-seq data. ChIP-Atlas provides more than 362,121 datasets for 6 model organisms (i.e., human, mouse, rat, fruit fly, nematode, and budding yeast) [44]. TF-Prioritizer automatically visualizes TF ChIP-seq peaks on predicted target sites of prioritized TFs to experimentally validate our predictions. TF-Prioritizer also visualizes experimentally known enhancers and super-enhancers from the manually curated database ENdb [45]. Additionally, experimental data from other databases or experimental data retrieved by own experiments can be supplied and processed by TF-Prioritizer.

By employing TF ChIP-seq data from ChIP-Atlas, TF-Prioritizer is capable of performing a TF co-occurrence analysis of prioritized TFs by systematically comparing the experimentally validated peaks of pairs of prioritized TFs. In a co-occurrence analysis, it is checked what percentage of available peaks of one TF is also found in another TF. TF-Prioritizer returns the percentage of similar peaks between prioritized TFs to discover the coregulation of TFs. We investigate the co-occurrence of TFs t_1 and t_2 in terms of statistical significance by calculating a log-likelihood score. Let B be the set of all TF binding sites and $\Pi(t)$ be the set of peaks for TF t. For TF t, let $count(t)$ be the number of binding sites $b \in B$ such that there is a peak $\pi \in \Pi(t)$ within b. For a TF-TF pair (t_1, t_2), let $count(t_1, t_2)$ be the number of binding sites $b \in B$ such that there is a peak $\pi_1 \in \Pi(t_1)$ and a peak $\pi_2 \in \Pi(t_2)$ within b, and then the log-likelihood score G^2 is calculated for the 4 observations: (i) $count(t_1, t_2)$ (i.e., t_1 and t_2 are co-occurring), (ii) $count(t_1) - count(t_1, t_2)$ (i.e., t_1 is occurring but t_2 is not), (iii) $count(t_2) - count(t_1, t_2)$ (i.e., t_2 is occurring but $t_1$is not), and (iv) $count(t_1, t_2) - count(t_1) - count(t_2) + |B|$ (i.e., neither t_1 nor t_2 is occurring), with their corresponding expectation values (i) $count(t_1) \cdot count(t_2)$, (ii) $count(t_1) * (|B| - count(t_2))$, (iii) $(|B| - count(t_1)) * count(t_2)$, and (iv) $(|B| - count(t_1)) * (|B| - count(t_2))$ as follows [46–48]:

$$G^2 = 2 \cdot \sum_{i\in\{a,b,c,d\}} observation_i \cdot \log\left(\frac{observation_i}{expectation_i}\right).$$

Note that when interpreting, each log-likelihood score needs to be brought into relation with the number of peaks found in the respective TFs and also set in relation with the other number of peaks determined in the entire log-likelihood table, as the log-likelihood score may differ from TF pair to TF pair. A high log-likelihood score, in combination with a high number of peaks, with respect to the entire log-likelihood table, generally indicates that the co-occurrence relationship is statistically significant and that the 2 TFs could be functionally related. For further details and explanation of the formula and interpretation, consult [46–48].

Explorative analysis of differentially expressed genes

TF-Prioritizer allows users to manually investigate the ChIP-seq signal in the identified CREs of differentially expressed genes. To this end, TF-Prioritizer generates a compendium of screenshots of the top 30 upregulated or downregulated loci (sorted by their total $\log_2$-fold change) between 2 sample groups. Additionally, we allow the user to specify loci that are of special interest (e.g., the CSN family or the *Socs2* locus in lactating mice). TF-Prioritizer then produces screenshots using the TF ChIP-seq data from ChIP-Atlas and visualizes them in a dynamically generated web application. Screenshots are produced using the IGV standalone application [27–29]. TF-Prioritizer also automatically saves the IGV session so the user can further research the shown tracks.

Handling missing data

In some cases, not all assay types are available for all samples, or the data do not have the same high quality as the rest of the samples. TF-Prioritizer then skips the grouping of missing data points and can still find meaningful results in the rest of the data. For example, the data for 3 time points for 1 histone modification are available, but 1 time point is missing or discarded. TF-Prioritizer then uses only the 3 available time points for grouping and downstream processing and analysis.

Using TF-prioritizer to investigate gene regulation

We use 3 approaches to evaluate the biological relevance and statistical certainty of our results: (i) literature research to validate whether the reported TFs are associated with the phenotype of interest, (ii) considering the top 30 target genes with highest affinity values and determining if their expression cluster by condition (note: we do not preselect differentially expressed genes for this analysis but focus on affinities to avoid a circular line of reasoning; we also review the literature and report whether these genes are known to be involved in either pregnancy or mammary gland development/lactation), and (iii) validation using independent TF ChIP-seq data from ChIP-Atlas. To conduct the third evaluation, we built region search trees, a balanced binary search tree where the leaves of the tree have a start and end position, and the tree returns all leaves that overlap with a searched region for all chromosomes of the tissue-specific ChIP-Atlas peaks for each available prioritized TF [49]. We then iterate over all predicted regions within the window size defined in TEPIC and determine if we can find any overlapping peaks inside the ChIP-Atlas peaks. If we can find an overlap with a peak defined by the ChIP-Atlas data, we count the predicted peak as a true positive (TP) or a false positive (FP). Next, we randomly sample the same number of predicted peaks in random length-matched regions not predicted to be relevant for a TF. If we find an overlap in the experimental ChIP-Atlas data, we consider this region as a false negative (FN) or a true negative (TN). Notably, we expect the FN count to be inflated since we considered condition-specific peaks of active CREs. Inactive CREs may very well have TFBSs that are not active. Nevertheless, we expect to find more such TFBSa in active regions compared to random samples, allowing us to compute sensitivity, specificity, precision, accuracy, and the harmonic mean between precision and sensitivity (F1-score) (see Supplementary Material S2).

Choice of parameters

In a pipeline like TF-Prioritizer, the choice of parameters is crucial to retrieve meaningful results. In this section, we explain our choice of parameters. We filter the RNA-seq data by a mean DESeq2 normalized gene count of 50 and a TPM of 1 to exclude noise of very weakly expressed target genes and TFs that are probably not important for the condition but would negatively impact the predictive models. We use the default configurations of TEPIC with the exception of the TF binding site search—that is, in the histone modification ChIP-seq data, it is important to search for TF binding sites between 2 peaks that are in close proximity (max. 500 bp) to each other (peak–dip–peak or peak–valley–peak model) [40]). The TEPIC2 framework and DYNAMITE were executed in default configurations as provided by the authors. We provide all default parameters in our configuration file.

Results and Discussion

We present TF-Prioritizer, which combines data to identify candidate CREs (e.g., ChIP-seq, ATAC-seq, DNase-seq) and RNA-seq to identify condition-specific TF activity. TF-Prioritizer is built on several existing state-of-the-art tools for peak calling, TF-affinity analysis, differential gene expression analysis, and machine learning tools. TF-Prioritizer is the first to jointly consider multiple types of modalities (e.g., different histone marks and/or time-series data), provide a joint list of active TFs, and enable the user to see a visualized validation of the predictions in an interactive and feature-rich web application.

Exploring TFs in mammary tissue during pregnancy and lactation in mice

We used TF-Prioritizer to identify TFs that are known to control mammary gland development and lactation. The tool also identifies TFs that are important in pregnancy, as well as new candidate TFs that have not yet been widely studied. TF-Prioritizer reported 104 TFs, many of which control Rho family GTPase-associated target genes and Casein family genes. TF-Prioritizer was evaluated using experimental TF ChIP-seq data where it showed high sensitivity, specificity, precision, and accuracy (Supplementary Fig. S2, Supplementary Material S2).

Prioritized TFs are known to play a role in mammary gland development and lactation

TF-Prioritizer prioritized STAT5, a transcription factor that plays an important role in mammary gland development [30, 50, 51]. *Stat5* messenger RNA (mRNA) levels are highly upregulated during the last days of pregnancy and at the beginning of lactation, supporting experimental findings that STAT5 is a key driver of mammary gland development. The predicted target genes of STAT5 show a clear expression separation between pregnancy and lactation (Fig. 3A, B). Peaks were predicted with a sensitivity of 57.8%, a specificity of 66.3%, a precision of 78.1%, an accuracy of 60.6%, and an F1-score of 66.5% (Supplementary Fig. S2). Additionally, STAT5 is known to activate the expression of the *Socs2* gene during mammary gland development [52, 53]. We can observe predicted peaks of STAT5 near *Socs2*, which could explain the regulation of its expression by STAT5 (Fig. 3C). STAT5 is further known to regulate the expression of the Casein gene family. *Csn2*, *Csn1s2a*, and *Csn1s2b* [54] mRNA levels are strongly upregulated during lactation, which could be explained by an activator role of STAT5 at the predicted peaks in their close proximity [55–57] (Fig. 3D, Supplementary Fig. S3, Supplementary Material S3, sec. STAT5).

Additionally, ELF5, another transcription factor that plays an important role in mammary gland development, was predicted to be relevant by TF-Prioritizer. *Elf5* mRNA levels increase at the end of pregnancy and the beginning of lactation, hence supporting ELF5's role in mammary gland development. Peaks were predicted with a sensitivity of 77.5%, a specificity of 80.5%, a precision of 81.6%, an accuracy of 79%, and an F1-score of 79.5% (Supplementary Fig. S2). TF-Prioritizer predicts ELF5 binding sites near *Gli1*. *Gli1* mRNA levels are downregulated during lactation, and ELF5 is thus probably acting as a suppressor for *Gli1*. Fiaschi et al. [58] showed experimentally that *Gli1*-expressing females were unable to lactate, and milk protein gene expression was essentially absent (Supplementary Figs. S4 and S5, Supplementary Material S3, sec. ELF5).

TF-Prioritizer further prioritized ESR1 [59] and NFIB [30], both known for their essential function in mammary gland development and lactation (Supplementary Material S3, sec. ESR1 and NFIB). Our results suggest that the mechanisms of pregnancy, mammary gland development, and lactation could be dependent on Rho GTPase [60, 61] and its regulation by several TFs reported here. Experimental validation is needed to elucidate those complex processes further (see Supplementary Material S3, sec. Rho GTPase's role in pregnancy, mammary gland development, and lactation) [62].

Prioritized novel TFs with a predicted role in pregnancy, mammary gland development, and lactation

We predict 2 TFs, CREB1 and ARNT, suggesting a role in the processes of pregnancy, mammary gland development, and lactation.

CREB1 binding sites show considerable overlap with binding sites of other TFs known to be involved in mammary gland development and lactation, such as ELF5 (22% of binding sites overlap, log-likelihood score 6,914 with a sample size of 16,531), NFIB (29% binding sites overlap, log-likelihood score 15,793 with a sample size of 23,923), and STAT5A (21% binding sites overlap, log-likelihood score 5,902 with a sample size of 15,180) (see Supplementary Fig. S6A–C). The co-occurrences could be significant due to the high log-likelihood values with a high sample size in comparison to the whole co-occurrence table. We hypothesize that a correlation of association strength may offer additional evidence for a functional association between TFs. Indeed, CREB1 shows a moderate correlation of binding site affinities with NFIB, STAT5A, STAT5B, and ELF5 (Supplementary Fig. S7). Our results suggest that CREB1 regulates a member of the Rho GTPase gene family and a member of the Casein gene family. Since CREB1 has not yet been recognized to contribute to aspects of mammary development and physiology, further experimental validation of our findings is needed (Supplementary Material S3, sec. CREB1).

Furthermore, the TF ARNT is prioritized along with 2 cofactors and predicted to be more involved in mammary gland development but less involved in lactation due to its high expression levels during the last state of pregnancy and lower expression during lactation. However, experimental mouse genetics demonstrated that ARNT is not required for mammary development and function [63], suggesting the presence of alternative and compensatory pathways (Supplementary Material S3, sec. ARNT).

Comparing TF-Prioritizer and diffTF

We compared TF-Prioritizer against the state-of-the-art tool diffTF that prioritizes and classifies TFs into repressors and activators given conditions (e.g., health and disease) [16]. diffTF does not allow multiple conditions or time-series data and distinct analysis of histone modification peak data in a single run and does not consider external data for validation. We point out that diffTF cannot use different sample sizes between ChIP-seq and RNA-seq data (i.e., diffTF requires that for each ChIP-seq sample, there is an RNA-seq sample and vice versa). diffTF does not use a biophysical model to predict TFBS but uses general, not tissue-specific, peaks of TF ChIP-seq data and considers all consensus peaks as TFBS [16]. For a comparison of features and technical details, see Supplementary Table S2 and Supplementary Table S3, respectively. Since the diffTF tool does not provide an aggregation approach to different conditions, we aggregate the prioritized TFs the same way as TF-Prioritizer does (i.e., the union of all prioritized TFs overall runs using diffTF's default q value cutoff of 0.1) to enhance the comparability of the overall conditions in the final results. In summary, diffTF prioritized 300 TFs compared to the 104 TFs (including combined TFs like Stat5a..Stat5b that count as 1 TF in TF-Prioritizer) that TF-Prioritizer reported (Fig. 4A). It thus seems that diffTF is less specific than TF-Prioritizer (see Supplementary Table S4 for a comparison of prioritized TFs). diffTF also finds known TFs that TF-Prioritizer captures (e.g., STAT5A, STAT5B, ELF5, and ESR1) but does not capture the well-known TF NFIB. diffTF also prioritizes CREB1 and ARNT, which, in our opinion, are strong candidates for experimental validation. By deploying 20 cores on a general computing cluster, TF-Prioritizer took roughly 7.5 hours

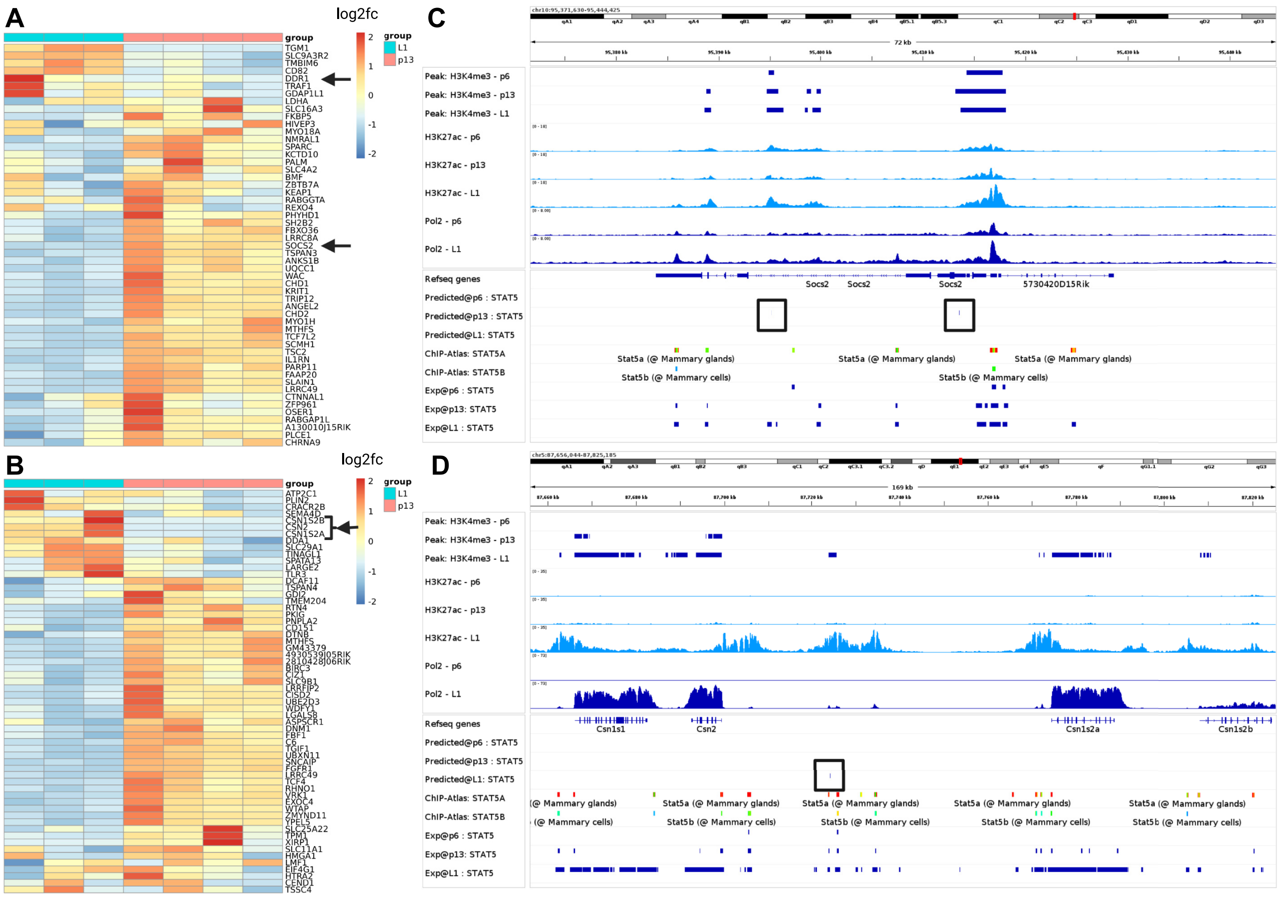


Figure 3: Validation of selected STAT5 target genes. (A, B) Heatmaps of predicted target genes. We select *Socs2* and *Csn* family genes (black arrows) as they are known to be crucial in either mammary gland development or lactation. In the heatmaps, we can observe a clear separation of these target genes between the time points p13 and L1. (C, D) IGV screenshots of the loci of *Socs2* and the *Csn* family. We included a predicted track in the IGV screenshot that indicates high-affinity binding regions for the TF that are represented by a tick and a black box surrounding it. In (C), we see that we predict peaks in p13 near *Socs2*. From these data, we suggest that *Socs2* mRNA expression is controlled by STAT5 [52, 53]. In (D), we can observe Pol2 tracks that show a distinct change in the expression of *Csn* family proteins between pregnancy and lactation. This indicates that STAT5 controls the expression of milk proteins.

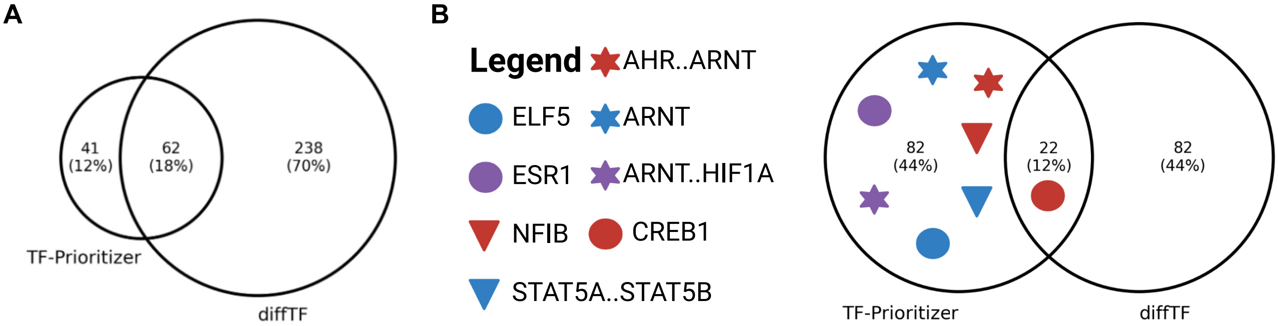


Figure 4: Venn diagram of prioritized TFs by TF-Prioritizer and diffTF. (A) diffTF and TF-Prioritizer found 62 (18.2%) common TFs. diffTF and TF-Prioritizer found known TFs (e.g., STAT5A, STAT5B, ELF5, and ESR1), but diffTF did not capture the well-known TF NFIB. diffTF and TF-Prioritizer both prioritized CREB1 and ARNT as candidates for experimental validation. (B) We ranked the diffTF results by *P* value and consider the top 104 (the same amount of TFs that the TF-Prioritizer predicted). Here only CREB1 is still predicted to be important by diffTF—other TFs such as STAT5A..STAT5B, ELF5, and NFIB drop out.

to be fully executed, and diffTF took approximately 41 hours to be fully executed. Due to the high number of TFs that are prioritized by diffTF, we ranked the TFs after their *P* value (where a low *P* value indicates higher evidence that a TF is involved in the processes) provided by diffTF and cut off the exact same amount of TFs (104 TFs) that are prioritized by TF-Prioritizer to make the benchmarking more comparable and interpretable. We observe that the known TFs drop out (e.g., STAT5A, STAT5B, ELF5, NFIB, ESR1) (Fig. 4B). CREB1, which we suggest to be a good candidate for experimental validation, can still be found in diffTF's prediction. Notably, only 22 TFs are prioritized by both TF-Prioritizer and diffTF by using this cutoff.

Limitations and considerations

TF-Prioritizer has several limitations. TF-Prioritizer is heavily dependent on the parameters of the state-of-the-art tools it is using (e.g., providing Hi-C data to TEPIC could have a significant impact on the search window while linking potential CREs to target genes). We also point out that we neither have any experimental evidence nor existing literature as proof that the default length of 500 bps of the dip model used in the extended TEPIC framework is the ideal cutoff.

We want to highlight the main disadvantage of using the TF-TG score as we significantly center the surveillance of TF-Prioritizer on genes showing a high fold change or high expression, which does not necessarily mean that those genes are the most relevant for a condition. Also, note that TF binding behavior is regulated by factors we do not observe here, such as phosphorylation. The results of the discounted cumulative gain ranking should be considered with care since the biologically most relevant TFs may manifest in only a subset of ChIP-seq data types.

The calculation of TP, TN, FP, and FN is only an approximation, as to the best of our knowledge, there is no known approach to determine if a CRE or TFBS is active in a condition or not. Sensitivity, specificity, precision, accuracy, and the harmonic mean of precision and sensitivity (F1) differ from TF to TF. We believe this is correlated with the prevalence of the binding sites or the motif specificity. We can also see a decline in the metrics if we look at cofactor regulation (Fig. 5A, AHR..ARNT, ARNT, and ARNT..HIF1A). We experience the highest performance of TF-Prioritizer by looking at TFs where no cofactor regulation is currently known or widely accepted (e.g., CREB1, ELF5, ESR1).

We further investigated the contribution of every single part of the TF-TG score to the number and quality of the prioritized TFs. To achieve this, we ran every combination of the components of the score (i.e., $\log_2$fc, TEPIC, DYNAMITE) with TF-Prioritizer. In Supplementary Table S5, we can see that the distribution analysis filters out about half of the TFs and only returns the most promising TFs. In Fig. 5B, we can see that ELF5, AHR..ARNT, and ARNT..HIF1A manifest in each of the scores independent of any combination. NFIB, CREB1, and ARNT manifest in any score that is related to TEPIC or DYNAMITE. ESR1 manifests in any score that is related to the LOG2FC. STAT5A..STAT5B only manifests in certain combinations of the scores or in the TF-TG score. The LOG2FC alone yields the most prioritized TFs, but at a closer look, the LOG2FC alone would miss NFIB, which is highly relevant in mammary gland development. Looking at these data, we believe that the TF-TG score that combines TEPIC, DYNAMITE, and LOG2FC results in the most promising TFs that are relevant.

In Figure 5C, we can see that STAT5A..STAT5B and ARNT only manifest in the HM H3K4me3. ELF5, CREB1, and NFIB only manifest in H3K27ac. ESR1, AHR..ARNT, and ARNT..HIF1A manifest in both HMs H3K4me3 and H3K27ac. As expected, most TFs only manifest in a subset of HMs, reflecting their association with certain chromatin states [64, 65].

Unraveling the specificity of TFs with respect to HM ChIP-seq, ATAC-seq, and DNase-seq

The ENCODE project generated a plethora of different assays for cell lines such as K562 and MCF-7, which we used here to determine to what extent different protocols (i.e., ATAC-seq, DNase-seq, and HM-ChIP-seq) are suited to reveal condition-specific TFs.

In total, we discovered 381 unique TFs (339 across 11 HM ChIP-seq experiments, 83 in ATAC-seq, and 96 in DNase-seq) if ATAC-seq and DNase-seq open chromatin peaks were processed with HINT to obtain footprints (Fig. 6, Supplementary Fig. S8A–C, Supplementary Fig. S9A–D). Interestingly, the efficacy of footprinting varies between the protocols significantly. Supplementary Fig. S9 shows differences in the number of footprints detected between both protocols. While the number of open chromatin peaks was nearly the same for both protocols, DNase-seq yields fewer footprints compared to ATAC-seq. In general, TF-Prioritizer reports more TFs when using footprinting compared to using open chromatin peaks. Many of these overlap with ChIP-seq TFs, confirming that footprinting is a meaningful strategy (Supplementary Fig. S8A, B, Fig. 6). We found TFs that can only be detected in a subset of the protocols (Fig. S6A, B, Supplementary Table 6). Using ChIP-seq data, we found the largest number of TFs, likely due to the combination of results from 10 different histone modifications and 1 histone variant, which together cover a wide variety of chromatin states. We found the largest number of detected TFs using the H2AFZ histone variant, possibly due to background peaks because of low antibody sensitivity in this histone variant. Of note, in Supplementary Fig. S10A, B, we investigated how the number of identified TFs differs when excluding H2AFZ. We can see a decrease in the total number of prioritized TFs in ChIP-seq from 339 to 301. We further examined how the number of identified TFs changes when only employing frequently studied HM ChIP-seq data from H3K27ac, H3K4me1, and H3K4me3 (Supplementary Fig. S10C, D). We can observe a decrease in identified TFs from 339 to 152, but again, the overlap with ATAC-seq and/or DNase-seq drops. H2AFZ is predominantly found in CREs and is also associated with cancer [66]. Since we have only investigated cancer cell lines, it remains unclear if this histone variant is generally highly informative of TF binding or if this is limited to cancer cells. Surprisingly, DNase-seq and ATAC-seq show a comparably small overlap even though both protocols are aimed at measuring chromatin accessibility. This corroborates earlier findings where it was observed that both protocols reveal assay-specific sites that contribute to predicting gene expression [67].

Indeed, some TFs known to be important for both cancer cell lines were reported through several protocols, while others were reported by only 1 protocol. For instance, we found MYC, a key TF for cell proliferation in K562 and MCF-7 cells [68, 69], was highly ranked in ATAC-seq and HM ChIP-seq (H3K4me2, H3K79me2). Conversely, GATA1, another TF important for cell differentiation in K562 [70, 71], was prioritized only by DNase-seq. GATA1 regulates MYB, a key hematopoietic TF involved in stem cell self-renewal and lineage decisions that is prioritized in HM ChIP-seq (H2AFZ, H3K27ac, H3K4me2) [71, 72]. TF-Prioritizer found many members of the SP (SP1, SP2, SP3, SP4, SP8, and SP9) and KLF (KLF1, KLF2, KLF3, KLF4, KLF6, KLF7, KL8, KLF9, KLF10, KLF11, KLF12, KLF14, KLF15, and KLF16) family to be important for K562 cell differentiation in a plethora of HM ChIP-seq, ATAC-seq, and DNase-seq experiments. Notably, TF-Prioritizer uses an individual TF energy pattern during the calculation of TF affinity to potential binding (i.e., TRAP) for each TF of a TF family. The incorporation of TF expression data in our score further boosts this differentiation between TFs of the same family. We identified 6 of 9 TFs from the SP TF family and 14 of 16 TFs from the KLF TF family [73]. Hu et al. [74] found that the SP and KLF TF families are most important in erythroid differentiation in K562 cells and that SP1 and SP3 are involved in activating GATA1 [75].

We further investigated if TF-Prioritizer found biologically relevant TFs for the MCF-7 cell line. We found ELF5, an important TF

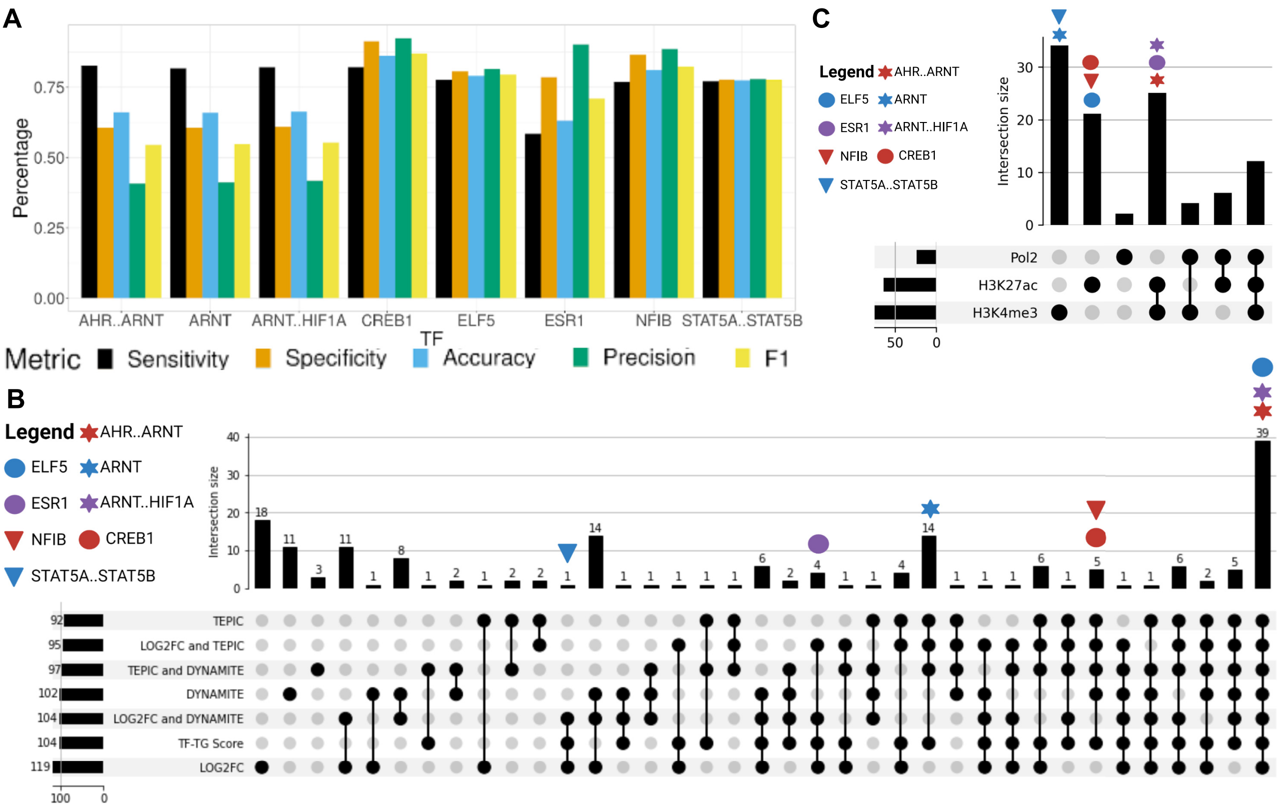


Figure 5: (A) Overview of performance metrics of prioritized TFs discussed in this article. (B) Contributions of individual components of the TF-TG score to the accumulated TF-TG score. We systematically considered different components of the TF-TG score (i.e., the score of TEPIC, LOG2FC, and DYNAMITE) as well as their combinations to determine their importance for the overall results. We find all important TFs exclusively using the TF-TG score. (C) Investigation of which TFs are reported in which assay. We can see that the most important TFs only manifest in a subset of HMs.

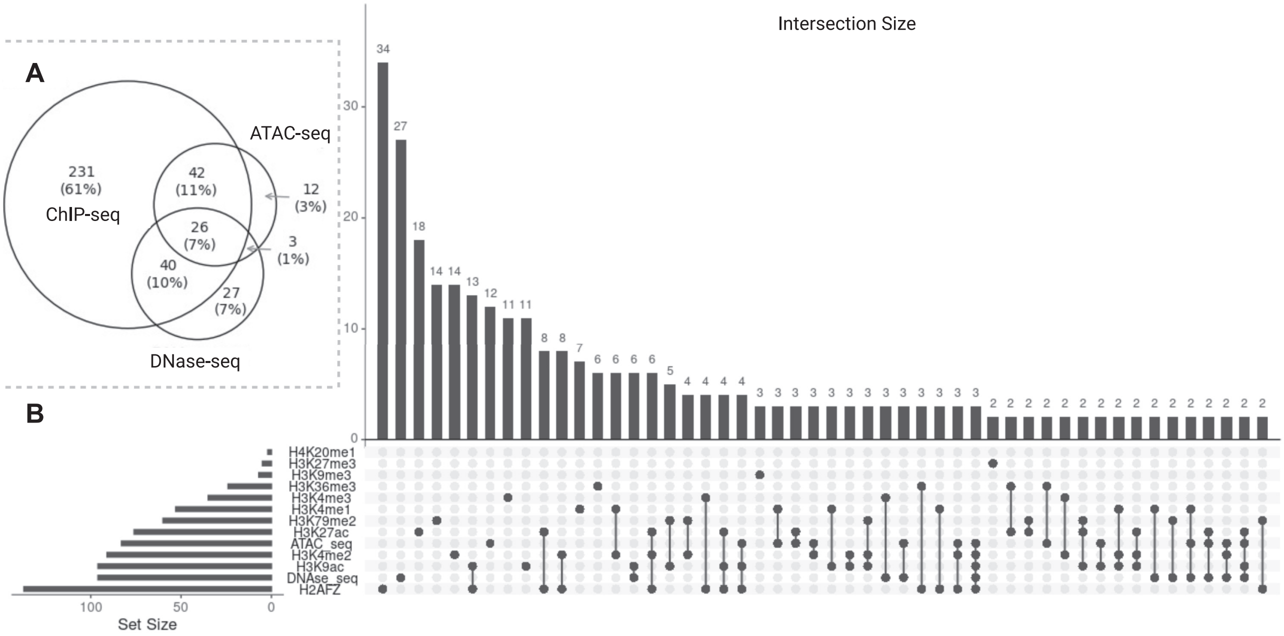


Figure 6: Guide to determine which experiments fit best by the usage of ATAC-seq, DNase-seq, or several histone modifications. (A) We combined all HM ChIP-seq data and investigated the overlap with ATAC-seq and DNase-seq. We found that ATAC-seq and ChIP-seq have a bigger overlap than ATAC-seq and DNase-seq. We found 26 TFs that are prioritized by all 3 protocols. (B) We separated the TFs of the HM ChIP-seq data in which HMs they were discovered. We can see huge differences between the HMs (e.g., while we can discover 137 TFs in H2AFZ, we can only discover 2 in H4K20me1).

in breast cancer, to be prioritized in ATAC-seq, DNase-seq, and HM ChIP-seq (H2AFZ). This is of particular interest, as ELF5 is a strong biomarker in breast cancer, and TF-Prioritizer is capable of prioritizing ELF5 in the ATAC-seq, DNase-seq, and ChIP-seq [76–78]. Piggin et al. [78] also postulated that ELF5 modulates the estrogen receptor. TF-Prioritizer found certain estrogen receptors (e.g., ESR2, ESRRG) to be relevant for cell differentiation in MCF-7. Estrogen receptor proteins are highly relevant in breast cancer [79, 80]. The TF GATA3 was also predicted (ATAC-seq, H3K27ac, H3K9ac) to be important for cell differentiation in MCF-7. GATA3 is a key player when it comes to cell differentiation in the MCF-7 cell line [81, 82] and a regulator of estrogen receptor proteins [83]. FOXA1, predicted by TF-Prioritizer (ATAC-seq), is important in cell differentiation for MCF-7 cell lines, is a critical determinant of estrogen receptor function, and affects the proliferation activity of breast cancer [84, 85].

Conclusion and Outlook

TF-Prioritizer is a pipeline that combines RNA-seq and ChIP-seq data to identify condition-specific TF activity. It builds on several existing state-of-the-art tools for peak calling, TF-affinity analysis, differential gene expression analysis, and machine learning tools. TF-Prioritizer is the first tool to jointly consider multiple types of modalities (e.g., different histone marks and/or time-series data) and provide a summarized list of active TFs. A particular strength of TF-Prioritizer is its ability to integrate all of this in an automated pipeline that produces a feature-rich and user-friendly web report. It allows interpreting results in the light of experimental evidence (TF ChIP-seq data) either retrieved automatically from ChIP-Atlas or user-provided and processed into genome browser screenshots illustrating all relevant information for the target genes. Our approach was heavily inspired by DYNAMITE [25, 86], which follows the same goal but requires manually performing all necessary steps.

We show that TF-Prioritizer is capable of identifying already known and validated TFs (e.g., STAT5, ELF5, NFIB, ESR1) that are involved in the process of mammary gland development or lactation and their experimentally validated target genes (e.g., *Socs2*, *Csn* milk protein family, Rho GTPase associated proteins). Furthermore, we prioritized some not yet recognized TFs (e.g., CREB1, ARNT) that we suggest as potential candidates for further experimental validation. These results led us to hypothesize that the Rho GTPases undergo major changes in their tasks during the stages of pregnancy, mammary gland development, and lactation, which are regulated by TFs.

In conclusion, each protocol and histone modification can unravel unique transcription factor binding sites that provide insight into gene regulatory mechanisms. It is our opinion that employing TF-Prioritizer on as many protocols and HM ChIP-seq experiments as possible could improve our understanding of given conditions.

In the future, we plan to extend TF-Prioritizer to more closely explore the combined effects of enhancers, which are often non-additive, as suggested by our current model [87]. We further plan to test the functionality of TF-Prioritizer on ATAC-seq data and to apply TF-Prioritizer in a single-cell context where histone ChIP-seq is currently hard to retrieve. Furthermore, we plan to include a more detailed ranking of the prioritized TFs. We plan to offer the user the ability to apply raw FASTQ files to TF-Prioritizer, where quality checks of the data will be performed. In summary, TF-Prioritizer is a powerful functional genomics tool that allows biomedical researchers to integrate large-scale ChIP-seq and RNA-seq data, prioritize TFs likely involved in condition-specific gene regulation, and interactively explore the evidence for the generated hypotheses in the light of independent data.

Availability of Source Code and Requirements

Project name: TF-Prioritizer
Project homepage: [89]
Operating system(s): Linux
Programming language: Java
Other requirements: Java version 11.0.14 or higher, Python version 3.8.5 or higher, R version 4.1.2 or higher, C++ version 9.4.0 or higher, CMAKE version 3.16 or higher, Angular CLI version 14.0.1 or higher, Node.js version 16.10.0 or higher, Docker version 20.10.12 or higher, and Docker-Compose version 1.29.2 or higher
Open source license: GNU GPL v. 3.0
RRID:SCR_023222

Additional Files

Supplementary Figure 1: TF-Prioritizer uses nf-core ChIP-seq / ATAC-seq and nf-core RNA-seq preprocessed data as input files (see GitHub repository for detailed formatting instructions). More specifically, broad peaks and gene counts. (1) Once started, the pipeline downloads necessary data (gene lengths, gene symbols, and short descriptions of the genes) from bioMar . (2) The user can then decide to use a transcript per million (TPM) filter or a gene count filter to filter before DESeq2 usage. We also allow for batch correction in DESeq2. TF-Prioritizer uses a TPM filter of 1 as default. DESeq2 normalizes and calculates the log2 fold change (log2fc) from raw gene count data . If the user used ATAC-seq as an input, we use the footprint method HINT to process the peaks, for this process we additionally expect BAM files in the same directory format as the peaks from the user. In parallel, (3) TF-Prioritizer preprocesses the ChIP-seq broad peaks by filtering blacklisted regions . We recommend using the sample combination option to combine similar broad peak samples into one peak file, as the total runtime of the pipeline is reduced significantly without losing the quality of the data. (4) Optionally, the user can decide to use TGene to predict links between target genes and regulatory elements combining distance and histone/expression correlation. If the TGene option is not activated, TEPIC, executed in the next step of the pipeline, uses a window-based approach to link regulatory elements to target genes. (5) TEPIC uses TRAP, an approach that quantifies transcription factor affinity scores based on a biophysical model for regulatory regions . TEPIC “computes TF affinities and uses open-chromatin/HM signal intensity as quantitative measures of TF binding strength”. TEPIC uses “machine learning to find low-affinity binding sites to improve the ability to explain gene expression variability compared to the standard presence/absence classification of binding sites” . In addition, especially for histone modification ChIP-seq data, we extended the TEPIC framework so that it can also search for transcription factor binding sites (TFBS) between two peaks that have close (∼500 bps) genomic positions (default: search between two peaks). (6) The pipeline then executes DYNAMITE an approach that uses a “sparse logistic regression classifier to infer TFs related to gene expression changes between samples” . (7) We added a distribution analysis to the pipeline to further prioritize TFs depending on their distribution compared to the global distribution using (8) a Mann-Whitney U test and the comparison of the means and the medians (for details see Materials and Methods Distribution Analysis Section). (9) We then use a discounted cumulative gain approach to retrieve a global ranking (overall histone modification data provided) of prioritized TFs (see Materials and Methods Discounted Cumulative Gain Section). (10) In the following, TF-Prioritizer generates condition-specific and histone-modification-specific heatmaps for prioritized TFs and their predicted target genes. (11) We then check if we can find publicly available tissue-specific TF ChIP-seq data from ChIP-ATLAS and (12) download the files. (13) Afterward, we take screenshots using the IGV. (14) In the last step, we conclude all analysis and plots in form of an easy-to-use HTML report that could also be used as a webpage.

Supplementary Figure 2: We showcase the discussed TFs and their statistical metrics. We can see the confusion matrix for each TF. We also provide sensitivity, specificity, precision, accuracy, and F1-score. We can see that the metrics differ vastly between the TFs. There is a drop in all metrics when it comes to co-factors. We believe that more research is necessary to obtain better predictions for co-factor TFs.

Supplementary Figure 3: We show the predicted peaks and experimental signals of DDR1. We can see higher Pol2 signals in the area of DDR1 due to higher expression during lactation. We furthermore can also observe a predicted peak of STAT5 in the DDR1 region. Past research showed that DDR1 is necessary to maintain STAT5 signaling during lactation.

Supplementary Figure 4: Validation of selected target genes for Elf5. (a) and (b) show heat maps of predicted target genes. We select Gli1, Lcp1, and Igfals (black arrows) as they are already known to be crucial in either mammary gland development or lactation. We further select the genes Arhgap9, Arhgef2, and Arhgap39 (black arrows) that are known to be essential for Rho GTPases due to their studied role in epithelial morphogenesis during mammary gland development [54,55]. In the heatmaps, we can observe a clear separation of these target genes between the time points p6-L1 and p6-L10. (c) and (d) show IGV screenshots of Arhgap9/ Gli1 and Lcp1 respectively. We included a predicted track in the IGV screenshot that indicates high-affinity binding regions for the TF that are represented by a tick and a black box surrounding it. In (c), we can see predicted Elf5 peaks near Arhgap9 and Gli1. ChIP-Atlas and the experimental TF ChIP-seq data substantiate the prediction near Arhgap9. Experimental data of Elf5 back up the predictions near Gli1. We can also observe upregulated Pol2 activity in L1 in this area. In (d) we can see multiple predictions of Elf5 bindings near Lcp1. ChIP-Atlas and the experimental TF ChIP-seq data corroborate the bindings of Elf5 in this area. We also observe an upregulated Pol2 activity in time points L1 and L10 in this area.

Supplementary Figure 5: We showcase three more examples of predicted peaks to experimental data for the transcription factor ELF5. With our predictions, we found two more genes associated with the Rho GTPase (ARHGAP39, ARHGEF2) and IGFALS that are known to play a role in mammary gland development and lactation (see Suppl. Material 2 for detailed discussion).

Supplementary Figure 6: We show the co-occurrence analysis of prioritized TFs. We can see that CREB1 has a high overlap of peaks with ELF5, NFIB, STAT5A, and STAT5b which are all key players in mammary gland development and lactation.

Supplementary Figure 7: We show the binding sites that co-occur between CREB1 and STAT5A..STAT5B, NFIB, or ELF5. We can see that there is a positive trend between the TFs. IF CREB1 has a higher binding affinity, the other TF that co-occurs on the same binding site also has a higher binding affinity on average.

Supplementary Figure 8: The above plots describe the common TFs across the different methods, ATAC-seq, DNase-seq, and ChIP-seq histone modifications, without correcting for the technical biases between the protocols using HINT. a) shows the overlapping TFs between ChIP-seq, ATAC-seq, and DNase-seq independent of single histone modifications. b) displays individual intersections of TFs between all possible combinations grouped by ATAC- and DNase-seq. c) represents ungrouped intersections between groups of the first X biggest overlaps.

Supplementary Figure 9: a) Analysis of overlaps between open-chromatin peaks in ATAC-seq and DNase-seq. b) Analysis of overlaps between open-chromatin peaks in ATAC-seq, DNase-seq, and ChIP-seq. c) Analysis of protocol bias-corrected footprints between ATAC-seq and DNase-seq. d) Analysis of protocol bias-corrected footprints between ATAC-seq, DNase-seq, and open-chromatin peaks of ChIP-seq.

Supplementary Figure 10: a) and b) Due to the high number of TFs found in H2AFZ, we excluded this histone variant. We can see that the number of identified TFs dropped from a total of 339 to 301. However, it also excluded some TFs that were identified by ATAC-seq and DNase-seq. c) and d) shows how the number of identified TFs behave if one only includes the frequently used HM ChIP-seq data H3K4me3, H3K4me1, and H3K27ac in comparison to DNase-seq and ATAC-seq. We can see a drop in totally identified TFs from 339 to 152 in ChIP-seq. However, also the number of overlaps between ATAC-seq and DNase-seq drops.

Supplementary Material 1: ENCODE file identifiers

Supplementary Material 2: Confusion matrices and the calculation of sensitivity, specificity, precision, accuracy, and F1-score

Supplementary Material 3: Biological findings

Supplementary Table 1: Feature comparison between TEPIC2 + DYNAMITE and TF-Prioritizer

Supplementary Table 2: Feature comparison between TF-Prioritizer and diffTF

Supplementary Table 3: Technical comparison between TF-Prioritizer and diffTF

Supplementary Table 4: Comparison of prioritized transcription factors between TF-Prioritizer and diffTF

Supplementary Table 5: Comparison of prioritized transcription factors before and after the filtering of the background distribution

Supplementary Table 6: Guide which TF was found in which protocol and HM

Abbreviations

Ahr: aryl hydrocarbon receptor; Arhgap9: Rho GTPase activating protein 9; Arhgap12: Rho GTPase activating protein 12; Arhgap39: Rho GTPase activating protein 39; Arhgef1: Rho guanine nucleotide exchange f

actor 1; Arhgef2: Rho/Rac guanine nucleotide exchange factor 2; Arhgef9: Cdc42 guanine nucleotide exchange factor 9; Arhgef18: Rho/Rac guanine nucleotide exchange factor 18; Arhgef40: Rho guanine nucleotide exchange factor 40; Arnt: aryl hydrocarbon receptor nuclear translocator; bp: base pair; ChIP: chromatin immunoprecipitation; CRE: *cis*-regulatory element; Creb1: CAMP responsive element binding protein 1; Csn: Casein proteins; Csn1s2a: Casein alpha S2 like A; Csn1s2b: Casein alpha S2 like B; Csn2: Casein beta; Csnk1e: Casein kinase 1 epsilon; Csnk2a2: Casein kinase 2 alpha 2; Csnk2b: Casein kinase 2 beta; Ddr1: discoidin domain receptor tyrosine kinase 1; Elf5: E74 like ETS transcription factor 5; Esr1: estrogen receptor 1; Ets2: ETS proto-oncogene 2, transcription factor; F1-score: harmonic mean between precision and sensitivity; FN: false negatives; FP: false positives; Gli1: GLI family zinc finger 1; HM: histone modification; Hif1a: hypoxia inducible factor 1 subunit alpha; Igfals: insulin-like growth factor binding protein acid labile subunit; IGV: Integrative Genome Viewer; L1: lactation day 1; L10: lactation day 10; Lcp1:

lymphocyte cytosolic protein 1; mRNA: messenger RNA; MWU: Mann–Whitney *U* test; Nfib: nuclear factor I B; p6: pregnancy day 6; p13: pregnancy day 13; Socs2: suppressor of cytokine signaling 2; Stat5 (composition of Stat5a and Stat5b): signal transducer and activator of transcription 5A + signal transducer and activator of transcription 5B; TF: transcription factor; TFBS: transcription factor binding sites; TG: target gene; TF-Gene score: retrieved by TEPIC; TF-TG score: retrieved by the distribution analysis; TP: true positives; TPM: transcripts per million; Tp53: tumor protein p53; log$_2$fc: log$_2$ fold-change.

Authors' Contributions

M.H., N.T., F.S., J.B., M.S., D.B.B., L.H., and M.L. drafted the concept for this pipeline. M.H., N.T., O.L., and K.Y. implemented the pipeline. M.H., L.S., and N.T. conceptualized and implemented the ATAC-seq and DNase-seq integration. J.J. and H.K.L. created the experimental data in the laboratory. J.J., H.K.L., L.-L.W., K.Y., and N.B. prepared the data and performed quality checks. M.H., N.T., D.B.B., L.H., and M.L. wrote the manuscript. All authors reviewed the manuscript and approved it.

Funding

This work was supported by the Technical University Munich—Institute for Advanced Study, funded by the German Excellence Initiative. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the *e:Med* research and funding concept (*grants 01ZX1908A / 01ZX2208A* and *grants 01ZX1910D / 01ZX2210D*). This work was supported by the Deutsche Forschungsgemeinschaft SFB1531 (S03, project number 456687919). D.B.B. was supported by the German Federal Ministry of Education and Research (BMBF, grant no. 031L0309A). This work was supported in part by the Intramural Research Programs (IRPs) of National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 422216132. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Competing Interests

The authors declare no competing interests.

Data Availability

All supporting data and materials are available in the *GigaScience* GigaDB database [88].

Acknowledgments

We thank Christina Trummer for designing the TF-Prioritizer logo and Andreas Niekler for his help with the statistical analysis of the co-occurrence analysis.

Figures were created with https://www.biorender.com. Parts of the images were designed using resources from Flaticon.com.

References

1. Collins, FS, Green, ED, Guttmacher, AE. US National Human Genome Research Institute: a vision for the future of genomics research. *Nature* 2003;**422**:835–47.
2. Malecová, B, Morris, KV. Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. *Curr Opin Mol Ther* 2010;**12**:214–22.
3. Vaquerizas, JM, Kummerfeld, SK, Teichmann, SA, *et al.* A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;**10**:252–63.
4. Hwa, V. STAT5B deficiency: impacts on human growth and immunity. *Growth Horm IGF Res* 2016;**28**:16–20.
5. Andersson, EI, Tanahashi, T, Sekiguchi, N, *et al.* High incidence of activating STAT5B mutations in CD4-positive T-cell large granular lymphocyte leukemia. *Blood* 2016;**128**:2465–8.
6. Anzalone, AV, Randolph, PB, Davis, JR, *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 2019;**576**:149–57.
7. Scholefield, J, Harrison, PT. Prime editing—an update on the field. *Gene Ther* 2021;**28**:396–401.
8. Ignatieva, EV, Levitsky, VG, Kolchanov, NA. Human genes encoding transcription factors and chromatin-modifying proteins have low levels of promoter polymorphism: a study of 1000 genomes project data. *Int J Genomics Proteomics* 2015;**2015**:260159.
9. Zhou, Q, Liu, M, Xia, X, *et al.* A mouse tissue transcription factor atlas. *Nat Commun* 2017;**8**:1–15.
10. Lee, BH, Rhie, SK. Molecular and computational approaches to map regulatory elements in 3D chromatin structure. *Epigenetics Chromatin* 2021;**14**:14.
11. Keenan, AB, Torre, D, Lachmann, A, *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res* 2019;**47**:W212–24.
12. Roopra, A. MAGIC: a tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. *PLoS Comput Biol* 2020;**16**:e1007800.
13. Holland, CH, Tanevski, J, Perales-Patón, J, *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol* 2020;**21**:36.
14. Ferreira, SS, Hotta, CT, de Carli Poelking, VG, *et al.* Co-expression network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. *Plant Mol Biol* 2016;**1**:15–35.
15. Mason, MJ, Fan, G, Plath, K, *et al.* Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *Bmc Genomics* 2009;**10**:327.
16. Berest, I, Arnold, C, Reyes-Palomares, A, *et al.* Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF. *Cell Rep* 2019;**29**:3147–3159.e12. e12.
17. Oki, S, Ohta, T, Shioi, G, *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* 2018;**19**:e46255.
18. Zhang, Y, Liu, T, Meyer, CA, *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol* 2008;**9**:R137.
19. Hesselberth, JR, Chen, X, Zhang, Z, *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 2009;**6**:283–9.
20. Li, Z, Schulz, MH, Look, T, *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* 2019;**20**:45.
21. Gusmao, EG, Dieterich, C, Zenke, M, *et al.* Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* 2014;**30**:3143–51.

22. Gusmao, EG, Allhoff, M, Zenke, M, *et al.* Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Methods* 2016;**13**:303–9.
23. Roider, HG, Kanhere, A, Manke, T, *et al.* Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 2007;**23**:134–41.
24. Schmidt, F, Gasparoni, N, Gasparoni, G, *et al.* Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res* 2017;**45**:54–66.
25. Schmidt, F, Kern, F, Ebert, P, *et al.* TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics* 2019;**35**:1608–9.
26. Mann, HB, Whitney, DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist* 1947;**18**:50–60.
27. Robinson, JT, Thorvaldsdóttir, H, Winckler, W, *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;**9**:24–26.
28. Thorvaldsdóttir, H, Robinson, JT, Mesirov, JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf* 2013;**14**:178–92.
29. Robinson, JT, Thorvaldsdóttir, H, Wenger, AM, *et al.* Variant review with the Integrative Genomics viewer. *Cancer Res* 2017;**77**:e31–4.
30. Lee, HK, Willi, M, Kuhns, T, *et al.* Redundant and non-redundant cytokine-activated enhancers control Csn1s2b expression in the lactating mouse mammary gland. *Nat Commun* 2021;**12**:2239.
31. Patel, H, Ewels, P, Peltzer, A, *et al.* nf-core/rnaseq: nf-core/rnaseq v3.6—platinum platypus. *Zenodo*. 2022. doi:10.5281/zenodo.6327553
32. Patel, H, Wang, C, Ewels, P *et al.*, nf-core/chipseq: nf-core/chipseq v1.2.2—Rusty Mole. *Zenodo*. 2021. doi:10.5281/zenodo.4711243
33. Ewels, PA, Peltzer, A, Fillinger, S, *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 2020;**38**:276–8.
34. Patro, R, Duggal, G, Love, MI, *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**:417–9.
35. Liu, T. Advanced: call peaks using MACS2 subcommands. *GitHub*. 2016. https://github.com/macs3-project/MACS/wiki/Advanced%3A-Call-peaks-using-MACS2-subcommands
36. Smedley, D, Haider, S, Durinck, S, *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 2015;**43**:W589–98.
37. Love, MI, Huber, W, Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
38. Yan, F, Powell, DR, Curtis, DJ, *et al.* From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 2020;**21**:22.
39. Amemiya, HM, Kundaje, A, Boyle, AP. The ENCODE blacklist: identification of problematic regions of the *Sci Rep* 2019;**9**:9354.
40. Pundhir, S, Bagger, FO, Lauridsen, FB, *et al.* Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. *Nucleic Acids Res* 2016;**44**:4037–51.
41. Description.Pdf at master· SchulzLab/TEPIC. *GitHub*. https://github.com/SchulzLab/TEPIC
42. Ouyang, Z, Zhou, Q, Wong, WH. ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci USA* 2009;**106**:21521–6.
43. Karadimitriou M. *Mann-Whitney U test*. https://maths.shu.ac.uk/mathshelp/Stats%20support%20resources/Resources/Nonparametric/Comparing%20groups/Mann-Whitney/SPSS/stcp-marshall-MannWhitS.pdf
44. Zou, Z, Ohta, T, Miura, F, *et al.* ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and bisulfite-seq data. *Nucleic Acids Res* 2022;**0**:W175–82.
45. Bai, X, Shi, S, Ai, B, *et al.* ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res* 2020;**48**:D51–7.
46. Wiedemann, G, Niekler, A. *Hands-On: A Five Day Text Mining Course for Humanists and Social Scientists in R*. Teach4DH@ GSCL.
47. Rayson, P, Berridge, D, Francis, B. Extending the Cochran rule for the comparison of word frequencies between corpora. *In: 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*. 2004:926–36.
48. Gries, ST, Durrant, P. *et al. Analyzing co-occurrence data*. In: M Paquot, (ed.), A Practical Handbook of Corpus Linguistics. Cham, Switzerland: Springer International Publishing; 2020: 141–59.
49. Tropf H. Multidimensional range search in dynamically balanced trees. *Angew Inform*. http://hermanntropf.de/media/multidimensionalrangequery.pdf (accessed: 02/2023)
50. Cui, Y, Riedlinger, G, Miyoshi, K, *et al.* Inactivation of Stat5 in mouse mammary epithelium during pregnancy reveals distinct functions in cell proliferation, survival, and differentiation. *Mol Cell Biol* 2004;**24**:8037–47.
51. Liu, X, Robinson, GW, Wagner, KU, *et al.* Stat5a is mandatory for adult mammary gland development and lactogenesis. *Genes Dev* 1997;**11**:179–86.
52. Croker, BA, Kiu, H, Nicholson, SE. SOCS regulation of the JAK/STAT signalling pathway. *Semin Cell Dev Biol* 2008;**19**:414–22.
53. Zeng, X, Willi, M, Shin, HY, *et al.* Lineage-specific and non-specific cytokine-sensing genes respond differentially to the master regulator STAT5. *Cell Rep* 2016;**7**:3333–46.
54. Głąb, TK, Boratyński, J. Potential of Casein as a carrier for biologically active agents. *Top Curr Chem (Z)* 2017;**375**:71.
55. Ryskaliyeva, A, Henry, C, Miranda, G, *et al.* Alternative splicing events expand molecular diversity of camel CSN1S2 increasing its ability to generate potentially bioactive peptides. *Sci Rep* 2019;**9**:5243.
56. Groenen, MAM, Dijkhof, RJM, Verstege, AJM, *et al.* The complete sequence of the gene encoding bovine α2-casein. *Gene* 1993;**23**:187–93.
57. Wellberg, E, Metz, RP, Parker, C, *et al.* The bHLH/PAS transcription factor singleminded 2 s promotes mammary gland lactogenic differentiation. *Development* 2010;**137**:945–52.
58. Fiaschi, M, Rozell, B, Bergström, Å, *et al.* Targeted expression of GLI1 in the mammary gland disrupts pregnancy-induced maturation and causes lactation failure. *J Biol Chem* 2007;**282**:36090–101.
59. Ogorevc, J, Dovč, P. Expression of estrogen receptor 1 and progesterone receptor in primary goat mammary epithelial cells. *Anim Sci J* 2016;**87**:1464–71.
60. Van Aelst, L, Symons, M. Role of Rho family gtpases in epithelial morphogenesis. *Genes Dev* 2002;**16**:1032–54.
61. Zuo, Y, Oh, W, Ulu, A, *et al.* Minireview: mouse models of rho gtpase function in mammary gland development, tumorigenesis, and metastasis. *Mol Endocrinol* 2016;**30**:278–89.
62. Joo, E, Olson, MF. Regulation and functions of the RhoA regulatory guanine nucleotide exchange factor GEF-H1. *Small GTPases* 2021;**12**:358–71.
63. Le Provost, F, Riedlinger, G, Hee Yim, S, *et al.* The aryl hydrocarbon receptor (AhR) and its nuclear translocator (Arnt) are dispensable for normal mammary gland development but are required for fertility. *Genesis* 2002;**32**:231–9.

64. Lickwar, CR, Mueller, F, Hanlon, SE, *et al.* Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* 2012;**484**:251–5.
65. Steinfeld, I, Shamir, R, Kupiec, M. A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. *Nat Genet* 2007;**39**:303–9.
66. Giaimo, BD, Ferrante, F, Herchenröther, A, *et al.* The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin* 2019;**12**:37.
67. Nordström, KJV, Schmidt, F, Gasparoni, N, *et al.* Unique and assay specific features of nome-, ATAC- and dnase I-seq data. *Nucleic Acids Res* 2019;**47**:10580–96.
68. Delgado, MD, Lerga, A, Cañelles, M, *et al.* Differential regulation of max and role of c-myc during erythroid and myelomonocytic differentiation of K562 cells. *Oncogene* 1995;**10**:1659–65.
69. Wang, Y-H, Liu, S, Zhang, G, *et al.* Knockdown of c-myc expression by rnai inhibits MCF-7 breast tumor cells growth in vitro and in vivo. *Breast Cancer Res* 2005;**7**:R220–8.
70. Huang, D-Y, Kuo, Y-Y, Chang, Z-F. GATA-1 mediates auto-regulation of gfi-1B transcription in K562 cells. *Nucleic Acids Res* 2005;**33**:5331–42.
71. Halsey, C, Docherty, M, McNeill, M, *et al.* The GATA1s isoform is normally down-regulated during terminal haematopoietic differentiation and over-expression leads to failure to repress MYB, CCND2 and SKI during erythroid differentiation of K562 cells. *J Hematol Oncol* 2012;**5**:45.
72. Sakamoto, H, Dai, G, Tsujino, K, *et al.* Proper levels of c-myb are discretely defined at distinct steps of hematopoietic cell development. *Blood* 2006;**108**:896–903.
73. Suske, G, Bruford, E, Philipsen, S. Mammalian SP/KLF transcription factors: bring in the family. *Genomics* 2005;**85**:551–6.
74. Hu, JH, Navas, P, Cao, H, *et al.* Systematic RNAi studies on the role of Sp/KLF factors in globin gene expression and erythroid differentiation. *J Mol Biol* 2007;**366**:1064–73.
75. Hou, CH, Huang, J, He, QY, *et al.* Involvement of Sp1/Sp3 in the activation of the GATA-1 erythroid promoter in K562 cells. *Cell Res* 2008;**18**:302–10.
76. Qu, X, Li, Q, Tu, S, *et al.* ELF5 inhibits the proliferation and invasion of breast cancer cells by regulating CD24. *Mol Biol Rep* 2021;**48**:5023–32.
77. Li, X, Li, S, Li, B, *et al.* Acetylation of ELF5 suppresses breast cancer progression by promoting its degradation and targeting CCND1. *npj Precis Onc* 2021;**5**:20.
78. Piggin, CL, Roden, DL, Law, AMK, *et al.* ELF5 modulates the estrogen receptor cistrome in breast cancer. *PLoS Genet* 2020;**16**:e1008531.
79. Vantangoli, MM, Madnick, SJ, Huse, SM, *et al.* MCF-7 human breast cancer cells form differentiated microtissues in scaffold-free hydrogels. *PLoS One* 2015;**10**:e0135426.
80. Russo, J, Russo, IH. The role of estrogen in breast cancer. In: *Molecular Basis of Breast Cancer.* 2004:89–135.
81. Chou, J, Provot, S, Werb, Z. GATA3 in development and cancer differentiation: cells GATA have it! *J Cell Physiol* 2010;**222**:42–49.
82. Kouros-Mehr, H, Bechis, SK, Slorach, EM, *et al.* GATA-3 links tumor differentiation and dissemination in a luminal breast cancer model. *Cancer Cell* 2008;**13**:141–52.
83. Eeckhoute, J, Keeton, EK, Lupien, M, *et al.* Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res* 2007;**67**:6477–83.
84. Hurtado, A, Holmes, K, Ross-Innes, C, *et al.* FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* 2011;**43**:27–33.
85. Tachi, K, Shiraishi, A, Bando, H, *et al.* FOXA1 expression affects the proliferation activity of luminal breast cancer stem cell populations. *Cancer Sci* 2016;**107**:281–9.
86. Durek, P, Nordström, K, Gasparoni, G, *et al.* Epigenomic profiling of human CD4+ T cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity* 2016;**45**:1148–61.
87. Shin, H, Willi, M, Yoo, K, *et al.* Hierarchy within the mammary STAT5-driven wap super-enhancer. *Nat Genet* 2016;**48**:904–11.
88. Hoffmann, M, Trummer, N, Schwartz, L, *et al.* Supporting data for "TF-Prioritizer: A Java Pipeline to Prioritize Condition-Specific Transcription Factors." *GigaScience Database.* 2023. http://dx.doi.org/10.5524/102379.
89. Hoffmann, M, Trummer, N. TF-Prioritizaer GitHub. 2023. https://github.com/biomedbigdata/TF-Prioritizer.
90 The ENCODE Consortium. ENCODE portal https://www.encodeproject.org/.

4.2. Publication 2: circRNA-sponging: a pipeline for extensive analysis of circRNAs and their miRNA sponging effects

Citation

The article titled "circRNA-sponging: a pipeline for extensive analysis of circRNA expression and their role in miRNA sponging" has been published online at Oxford University Press Bioinformatics Advances on 08 July 2023.

Full citation:

Hoffmann, M., Schwartz, L., Ciora, O.-A., Trummer, N., Willruth, L.-L., Jankowski, J., Lee, H. K., Baumbach, J., Furth, P., Hennighausen, L., & List, M. (2023). circRNA-sponging: a pipeline for extensive analysis of circRNA expression and their role in miRNA sponging. *Bioinformatics Advances*, vbad093. PMCID: PMC10359604

Summary

Circular RNAs (circRNAs) are long non-coding RNAs (lncRNAs), and some are already associated with diseases. circRNAs have the potential as biomarkers for diagnosis and treatment. Among their functions, circRNAs can act as microRNA (miRNA) sponges, preventing miRNAs from repressing their targets. However, no existing pipeline systematically assesses the sponging potential of circRNAs. In this study, I developed circRNA-sponging, a nextflow pipeline that identifies and analyzes circRNAs and their sponging potential using RNA-seq and miRNA-seq data. The pipeline performs the following steps: “(1) identification of circRNAs through back-splicing junctions, (2) quantification of circRNA expression in relation to linear counterparts, (3) differential expression analysis, (4) miRNA expression identification and quantification, (5) prediction of miRNA binding sites on circRNAs, (6) investigation of potential circRNA-miRNA sponging events, (7) creation of a competing endogenous RNA network, and (8) identification of potential circRNA biomarkers. I demonstrated the functionality of the circRNA-sponging pipeline using multiple brain tissues and showed that circRNAs were differentially expressed between those” [53].

Availability

The nextflow pipeline circRNA-sponging is maintained and available at https://github.com/biomedbigdata/circRNA-sponging.
The data used in this analysis is freely available at GEO under the IDs GSE100265 and GSE93129.

Contribution

I had a leadership role during the planning and development phase for the circRNA-sponging nextflow pipeline. My main responsibilities were supervising and providing support to Leon Schwartz in the pipeline's implementation as well as parts of the pipeline, and plots were implemented by myself. I was also in charge of drafting the manuscript. After receiving the reviewers' feedback, I worked to address their concerns, enhancing the manuscript to satisfy their requirements. Lastly, I took responsibility for the entire publication process, which involved adhering to submission guidelines, deadlines, and fulfilling other obligations associated with academic publishing.

Rights and permissions

The original article is embedded under the open-access license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/) by Oxford Academic Press. All rights belong to Oxford Academic Press.

Additional supplementary material

Supplementary data are available at Bioinformatics Advances online.

Bioinformatics Advances, 2023, **00**, vbad093
https://doi.org/10.1093/bioadv/vbad093
Advance access publication 8 July 2023
Original Article

Gene regulation

circRNA-sponging: a pipeline for extensive analysis of circRNA expression and their role in miRNA sponging

Markus Hoffmann [1,2,3,*,†], Leon Schwartz [1,†], Octavia-Andreea Ciora [1,†], Nico Trummer [1], Lina-Liv Willruth [1], Jakub Jankowski[3], Hye Kyung Lee [3], Jan Baumbach [4,5], Priscilla A. Furth [3,6], Lothar Hennighausen [2,3] and Markus List [1,*]

[1]Big Data in BioMedicine Group, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising D-85354, Germany
[2]Institute for Advanced Study, Technical University of Munich, Garching D-85748, Germany
[3]Laboratory of Genetics and Physiology, National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
[4]Computational Systems Biology, University of Hamburg, Hamburg, Germany
[5]Computational BioMedicine Lab, University of Southern Denmark, Odense, Denmark
[6]Departments of Oncology & Medicine, Georgetown University, Washington, DC, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Thomas Lengauer

Abstract

Motivation: Circular RNAs (circRNAs) are long noncoding RNAs (lncRNAs) often associated with diseases and considered potential biomarkers for diagnosis and treatment. Among other functions, circRNAs have been shown to act as microRNA (miRNA) sponges, preventing the role of miRNAs that repress their targets. However, there is no pipeline to systematically assess the sponging potential of circRNAs.

Results: We developed circRNA-sponging, a nextflow pipeline that (i) identifies circRNAs via backsplicing junctions detected in RNA-seq data, (ii) quantifies their expression values in relation to their linear counterparts spliced from the same gene, (iii) performs differential expression analysis, (iv) identifies and quantifies miRNA expression from miRNA-sequencing (miRNA-seq) data, (v) predicts miRNA binding sites on circRNAs, (vi) systematically investigates potential circRNA–miRNA sponging events, (vii) creates a network of competing endogenous RNAs and (viii) identifies potential circRNA biomarkers. We showed the functionality of the circRNA-sponging pipeline using RNA sequencing data from brain tissues, where we identified two distinct types of circRNAs characterized by a specific ratio of the number of the binding site to the length of the transcript. The circRNA-sponging pipeline is the first end-to-end pipeline to identify circRNAs and their sponging systematically with raw total RNA-seq and miRNA-seq files, allowing us to better indicate the functional impact of circRNAs as a routine aspect in transcriptomic research.

Availability and implementation: https://github.com/biomedbigdata/circRNA-sponging.

Contact: markus.hoffmann@nih.gov or markus.list@tum.de

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Circular RNAs (circRNAs) are classified as long noncoding RNAs (lncRNAs), even though a few have been reported to encode proteins (Miao *et al.*, 2021). circRNAs are characterized by their loop structure, which makes them less prone to degradation (Jeck and Sharpless, 2014; Yu and Kuo, 2019). The biogenesis of circRNAs is explained by the occurrence of a backsplicing event (see Supplementary Fig. S1) during the alternative splicing process of precursor messenger RNA (pre-mRNA), where the 5′ terminus of an upstream exon and the 3′ terminus of a downstream exon are covalently joined (Yu and Kuo, 2019). The difference between circRNAs and linear RNAs is the lack of a 5′ cap and a 3′ polyadenylation [poly(A)] tail along with its circular shape, which makes circRNAs more stable and, in most cases, resistant to exonuclease activity (Lasda and Parker, 2014; Mester-Tonczar *et al.*, 2020; Suzuki *et al.*, 2006). These circular molecules can be made up of exonic and intronic regions of its spliced pre-mRNA and are thus found in a variety of sizes, ranging from 100 to >4000 nucleotides (Lasda and Parker, 2014; Szabo and Salzman, 2016). circRNAs are conserved across species, and their expression is tissue- and disease-specific (Jeck *et al.*, 2013; Jeck and Sharpless, 2014; Zhang *et al.*, 2018). Hence, they can play an important role as biomarkers and therapeutic targets (Kristensen *et al.*, 2018; Qu *et al.*, 2015; Zhang *et al.*, 2018). Another type of noncoding RNA is microRNA (miRNA) which plays a role in post-transcriptional gene regulation (Hoffmann *et al.*, 2021; List *et al.*, 2019) and is involved in many biological processes and diseases (Kartha and Subramanian, 2014). miRNAs bind

Received: May 12, 2023. **Revised:** June 23, 2023. **Editorial Decision:** July 4, 2023. **Accepted:** July 7, 2023

their target genes via the RNA-induced silencing complex causing their degradation or preventing their translation (Weinstein *et al.*, 2013).

The possible interplay between circRNAs, miRNAs, messenger RNAs (mRNAs) that code for proteins, and other types of RNA that share miRNA binding sites gives rise to a large regulatory network. Salmena *et al.* (2011) proposed that any RNA that carries miRNA binding sites [e.g. mRNAs, circRNAs, pseudogenes, transcripts of 3′-untranslated regions (UTRs) and lncRNAs] can act as a competing endogenous RNA (ceRNA) that competes for the limited pool of available miRNAs in a cell. As a result of this competition, an overexpressed RNA can sponge away miRNAs required for the regulation of other RNAs, which can explain why noncoding RNAs, such as circRNAs, can be implicated in a phenotype.

The enhanced stability of circRNAs might allow them to work as buffers for miRNAs by binding them until sufficient miRNAs are present to outnumber the circRNA binding sites (Zhang *et al.*, 2018). The regulatory function of circRNAs and their alleged association with diseases are the main reasons why identifying sponging activity between circRNAs and miRNAs is of particular interest. The presence of an interaction between miRNAs and circRNAs has been repeatedly proven, and several circRNAs [e.g. CDR1as/CiRS-7, SRY (Hansen *et al.*, 2013) and circNCX1 (Li *et al.*, 2018)] have been recognized as miRNA sponges. Even though individual studies confirmed the existence of circRNA sponges, further studies are needed to elucidate the role of circRNAs in miRNA-mediated gene regulation.

From a computational point of view, the detection of circRNAs is difficult due to their circular shape and the lack of poly(A) tail, which makes it unlikely to observe them in poly(A)-enriched RNA sequencing (RNA-seq) libraries (Szabo and Salzman, 2016). Hence, circRNAs can only be robustly detected in libraries without poly(A) enrichment, such as ribosomal RNA (rRNA) depleted RNA-seq and total RNA-sequencing (total RNA-seq), which do not deplete circRNAs (Szabo and Salzman, 2016). Identification of circRNAs relies on the detection of backsplicing junctions among the unmapped reads, which allows for the estimation of circRNA abundance. By focusing on the backsplicing junction alone, the expression of circRNAs in relation to their linear counterparts is typically underestimated (Yu *et al.*, 2021).

Several approaches for circRNA analysis have been proposed. Chen *et al.* (2021) reviewed 100 existing circRNA-related tools for circRNA detection, annotation, downstream analysis, as well as network analysis. They list a total of 44 circRNA identification tools including, but not limited to, CIRCexplorer (Zhang *et al.*, 2016, 2014), find_circ (Memczak *et al.*, 2013), CIRI (Meng *et al.*, 2017), KNIFE (Szabo *et al.*, 2015) and circRNA_finder (Westholm *et al.*, 2014). They also present a total of 14 circRNA annotation databases collecting circRNA information from the literature, such as circBase (Glažar *et al.*, 2014) and CIRCpedia (Dong *et al.*, 2018; Zhang *et al.*, 2016). Other circRNA-related tools include databases for feature collection and storing circRNA information related to disease and biomarkers. In addition, circRNA network identification tools model the interactions between circRNAs and miRNAs, lncRNAs or RNA-binding proteins. Other tools for downstream analysis of circRNAs cover alternative splicing detection, circRNA assembly and structure prediction and visualization (Chen *et al.*, 2021). To the best of our knowledge, none of the tools provides a comprehensive and automated circRNA-sponging analysis integrating identification and quantification of both circRNAs and miRNAs, a systematic investigation of potential circRNA–miRNA sponging events and a ceRNA network analysis. We developed ‘circRNA-sponging’, a nextflow pipeline integrating state-of-the-art methods to (i) detect circRNAs via identifying backsplicing junctions from total RNA-seq data, (ii) quantify their expression values relative to linear transcripts, (iii) perform differential expression analysis, (iv) identify and quantify miRNA expression from miRNA-sequencing (miRNA-seq) data, (v) predict miRNA binding sites on circRNAs, (vi) systematically investigate potential circRNA–miRNA sponging events, (vii) create a ceRNA network and (viii) identify potential circRNA biomarkers using the ceRNA network (Fig. 1).

We demonstrate the potential of the circRNA-sponging pipeline on matched rRNA-depleted RNA-seq and miRNA-seq data from mouse brain tissues.

2 Methods

2.1 Data

Using circRNA-sponging, we processed a total of 23 samples of matched single-end rRNA-depleted RNA-seq and miRNA-seq data for four brain regions (cerebellum, cortex, hippocampus, olfactory bulb). Samples include three replicates for wild-type (WT) and 2–3 CDR1 knock-out (KO) mouse replicates (GEO accessions: GSE100265, GSE93129) (Piwecka *et al.*, 2017) (see Supplementary Table S1). We use the mm10 genome version for mapping.

2.2 Pipeline architecture

The circRNA-sponging pipeline is implemented in R (v. 4.2.0) and Python (v. 3.8.12) and wrapped with nextflow version 22.04.0.5697. The pipeline is hosted on dockerhub and will pull the required docker image when executed. The relevant image was built under docker version 22.06. It follows the nf-core guidelines (Ewels *et al.*, 2020) and encompasses several state-of-the-art techniques organized into three modules: (1) the circRNA module, (2) the miRNA module and (3) the sponging module, the latter of which can only be performed if both other modules have been executed (Fig. 2). In the following, we provide a deeper insight into each module and highlight important components of the pipeline.

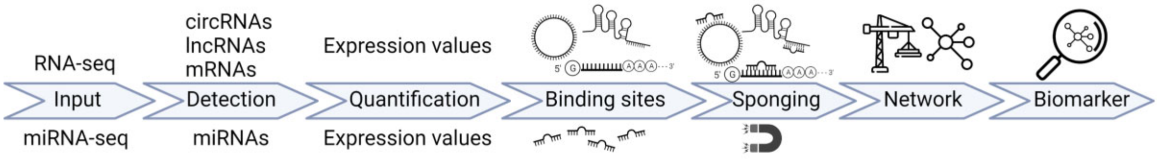


Figure 1. Overview of the individual steps of the pipeline. Total RNA-seq data processing is shown on top, and miRNA-seq processing on the bottom. In the miRNA sponging step, these results are integrated for network analysis and biomarker detection

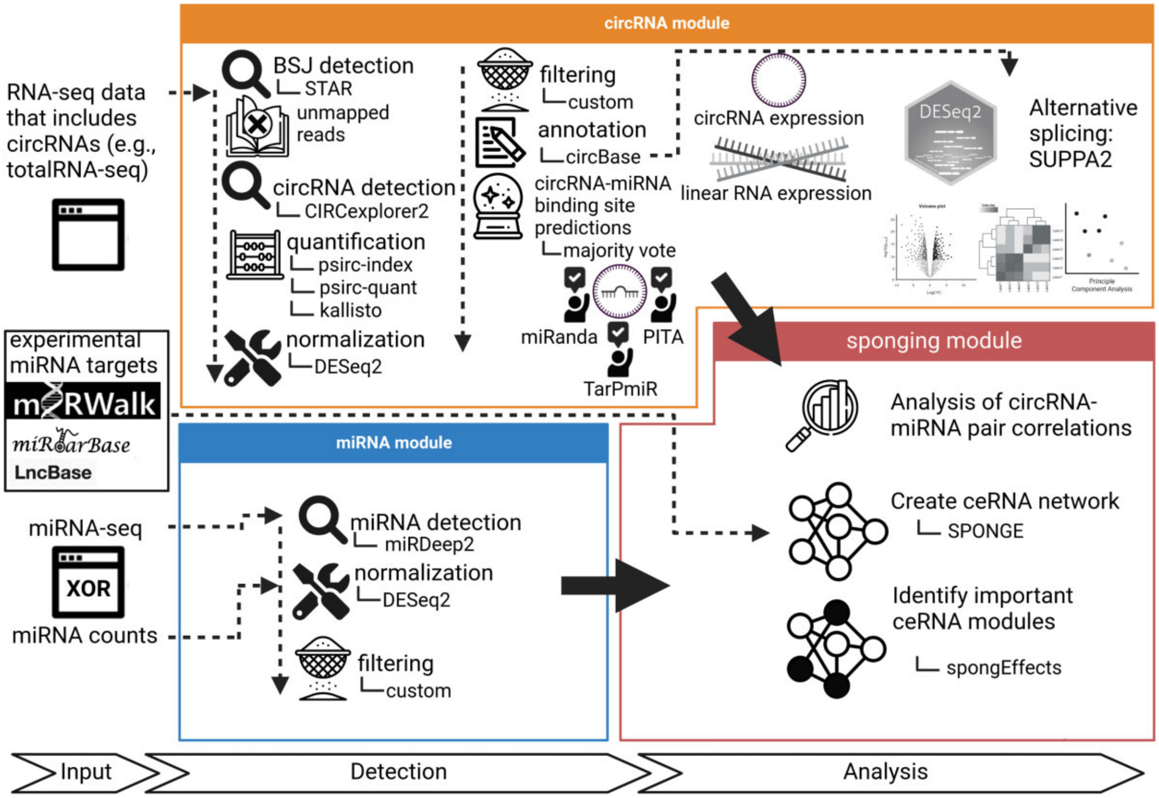


Figure 2. Workflow of the circRNA-sponging pipeline. The pipeline consists of three modules: (1) the circRNA module, (2) the miRNA module and (3) the sponging module. In (1), we detect circRNAs via identifying backsplicing junctions from total RNA-seq data, quantify their expression values, perform a differential expression analysis, and predict miRNA binding sites on circRNAs using a majority vote between three state-of-the-art methods. In (2), we either detect and quantify miRNAs in raw miRNA-seq or directly process miRNA expression data. In (3), we systematically investigate circRNA–miRNA sponging events, create a ceRNA network and use it to identify potential circRNA biomarkers

1) The circRNA module addresses the identification, quantification and miRNA binding site prediction of circRNAs. For read mapping, we employ the STAR (Dobin *et al.*, 2013) aligner, which provides support for the detection of splice-junction and fusion reads. The resulting unmapped split-reads are used by CIRCexplorer2, which uses a combination of methods (i.e. a de novo assembly approach to identify novel circRNA and a reference-based approach, which uses known exon-exon junctions to map backsplicing events to known genes) to increase the accuracy of its predictions (Zhang *et al.*, 2016) to identify backsplicing events. We could confirm the excellent performance of CIRCexplorer2 using data simulated with polyester (Frazee *et al.*, 2015) from the linear mouse reference genome GRCm38 at varying sequencing depth, where we rarely detect false positive backsplicing junctions and zero false positive circRNAs (Supplementary Fig. S2a). Next, raw read counts are normalized with DESeq2 (Love *et al.*, 2014), and circRNAs with low expression levels are excluded to reduce false positives. By default, only circRNAs with a normalized read count >5 in at least 20% of samples are retained. Database annotation is performed using circBase (Glažar *et al.*, 2014), which covers curated circRNAs with experimental evidence of several model organisms.

We use psirc (Yu *et al.*, 2021) to quantify circRNA expression levels, as the detection of backsplicing junctions alone does not reflect circRNA expression levels in relation to the gene's expression. To mitigate this, psirc employs kallisto (Bray *et al.*, 2016) and considers both linear and circular transcripts in the expectation-maximization step to produce comparable expression values. psirc corrects for various sequencing biases that can affect circRNA detection, such as coverage bias, mapping bias, read length bias and alternative splicing bias. Yu *et al.* (2021) showed that psirc provides a more accurate identification of circRNA expression levels by validating their method with experimental data. If the data have been sampled from different conditions (e.g. case and control), the quantified linear transcripts and circRNAs can be used to perform a differential expression analysis using DESeq2 (Love *et al.*, 2014). The pipeline generates heatmaps, volcano plots and principal component analysis (PCA) of the circRNAs and linear transcripts between conditions. We analyze alternative splicing between circular and linear transcripts on a gene level using SUPPA2 (Trincado *et al.*, 2018). In order to integrate circular transcripts, we construct a merged gene annotation file consisting of both linear and circular transcripts. Based on this input, we generate percent spliced-in (PSI) values for linear and circular isoforms with the SUPPA2 step psiPerIsoform (Trincado *et al.*, 2018). We, additionally, normalized the linear and circular PSI values for a gene by their sample-wise mean to account for differences in overall linear and circular splicing frequencies. Both nonnormalized and normalized PSI values are automatically visualized. To boost reliability, we predict circRNA–miRNA binding sites using a majority voting between miRanda (Enright *et al.*, 2003), PITA (Kertesz *et al.*, 2007) and TarPmiR (Ding *et al.*, 2016) since each method has a distinct approach for predicting miRNA binding sites. Testing these tools with random miRNA sequences shows that up to 25% of the reported binding sites may be false positives (Supplementary Fig. S2b and d), which aligns with previous findings (Min and Yoon, 2010). Thus, we consider a circRNA–miRNA binding site as relevant if it is predicted by at least two out of the three methods. miRanda considers seed matching, conservation and free energy, and we consider predictions with a score above the 25% quantile. PITA additionally considers site accessibility and target-site abundance. TarPmiR further integrates machine learning to improve results for supported organisms (Ding *et al.*, 2016). We further incorporate experimentally validated target sites from DIANA-LncBase v3 (Karagkouni *et al.*, 2020), miRTarBase (Hsu *et al.*, 2011; Huang *et al.*, 2020) and miRWalk3.0 (Sticht *et al.*, 2018).

2) The miRNA module covers the quantification and processing of miRNA expression. miRDeep2 (Friedländer *et al.*, 2012) is used to obtain miRNA counts. Alternatively, already mapped miRNA expression data can be provided. Raw counts are normalized with DESeq2 (Love *et al.*, 2014) followed by a filtering step, where by default, miRNAs with a normalized read count >5 in at least 20% of samples are retained.

3) The sponging module is used for the identification of crosstalk between circRNAs, miRNAs as well as ceRNA interactions of circRNAs with other transcripts. To identify potential sponging activity, we perform a correlation analysis of circRNA–miRNA pairs, where a negative correlation coefficient indicates a sponging relationship. For all circRNA–miRNA pairs (i.e. a circRNA that harbors at least one binding site for the miRNA), we compute a Pearson correlation coefficient along with the normalized residual sum of squares and the adjusted *P*-value after the Benjamini-Hochberg correction. Pairs are filtered (e.g. *P*-adjusted < 0.05, RMSE < 1.5, and optionally by the number of binding sites) and are considered potential sponging candidates. We further construct a ceRNA network using SPONGE (List *et al.*, 2019) on matched gene and miRNA expression data. Finally, we apply spongEffects (Boniolo *et al.*, 2023) to extract ceRNA modules consisting of circRNAs with a high node centrality score in the ceRNA network and their direct neighbors. For each module, spongEffects computes a sample-specific enrichment score (i.e. the spongEffects enrichment scores are calculated using one of three gene set enrichment approaches: Gene Set Enrichment Analysis (ssGSEA) and Gene Set Variation analysis (GSVA) algorithms as implemented in the GSVA package (version 1.34.0) (Hänzelmann *et al.*, 2013), or Overall Expression (OE) (Jerby-Arnon *et al.*, 2018)). These approaches can calculate spongEffects scores even if some genes in the ceRNAs modules are missing. The resulting module-by-sample score matrices can be used for further analysis. No major differences were observed between the three methods, and the choice of the optimal tool depends on the specific task and dataset (Boniolo *et al.*, 2023), such as differential analysis between groups or supervised machine learning. The spongEffects scores are then used for training and testing a random model classifier to distinguish between groups of samples (e.g. healthy and control) (Kuhn, 2008). The prediction power is then measured by a 5-fold cross-validation and a comparison to random modules (i.e. sampling modules of the same size as the ones predicted while preserving the size distribution of the real modules). In our example dataset, we used 16 samples for training (four samples for cerebellum, cortex, hippocampus and olfactory bulb) and eight samples for testing (two samples for cerebellum, cortex, hippocampus and olfactory bulb).

3 Results

circRNAs are highly abundant and conserved in the mammalian brain (Hanan *et al.*, 2017; Rybak-Wolf *et al.*, 2015). To demonstrate the capabilities of the circRNA-sponging pipeline, we analyzed a public RNA-seq dataset from the mouse brain. We focused on the sponging capacity of circRNAs and their potential role as ceRNAs.

3.1 Comparing circRNA and host gene expression reveals changes in circRNA splicing

In total, we detected 46 380 and 27 390 circRNAs before and after filtering, respectively. This number aligns with the known high abundance of circRNAs in brain tissue (Hanan *et al.*, 2017; Rybak-Wolf *et al.*, 2015). We could annotate only 1027 (∼4%) of the circRNAs that passed the filter (Supplementary Fig. S3a), as comparably few circRNAs have thus far been annotated in mice using circBase. psirc-estimated expression levels, which take reads mapping to parts other than the backsplicing junction of the circRNA into account, are up to 6-fold higher compared to counts derived from backsplicing junctions only (Fig. 3c, per tissue type: Supplementary Fig. S3d–g). We observed a generally higher expression of circRNAs in the cerebellum compared to other brain regions, which could indicate a higher importance of the circRNAs in this brain region (Supplementary Fig. S3b).

Concerning the miRNA binding sites, we observed that with a higher number of binding sites, the Pearson correlation increases (Supplementary Fig. S4). Despite the high number of shared circRNAs across brain regions (Fig. 3a), we observed a brain region-specific abundance of overall circRNAs levels (Fig. 3b, Supplementary Fig. S3b). However, expression levels of the circRNAs differ considerably, such that samples cluster by brain region (Fig. 3b, Supplementary Fig. S3c). Rybak-Wolf *et al.* (2015) reported circRNAs of 12 host genes (TULP4, RIMS2, ELF2, PHF21A, MYST4, CDR1, STAU2, SV2B, CPSF6, DYM, RMST and RTN4). They speculated on the importance of circRNAs originating from these genes for brain cell identity, but we posit that a change in circRNA expression alone does not necessarily imply a functional role as circRNA expression is coupled to the expression of the host gene, as we expect the number of reads mapping to the backsplicing junctions to correlate. We detected nine of the 12 circRNAs (all but TULP4, SV2B and RMST) in our analysis (Supplementary Table S2, Supplementary Fig. S5a and b). The difference between KO and WT samples is negligible, with the exception of the CDR1 region (mmu_hsa_circ_0001878 in circBase annotation) that was targeted successfully (Supplementary Fig. S5c and d). When excluding the circRNA of the CDR1, we observed a clear separation between the cerebellum and other brain regions, while the cortex and hippocampus are more similar (Galea *et al.*, 2011). Our analysis revealed a total of 33 circRNAs that show significantly different expression between brain regions (*P*-adjusted < 0.01, absolute log2 fold change > 5, Supplementary Table S3). By comparing the expression level of the circRNAs to the linear transcripts, as facilitated by psirc-quant, we can identify cases where the expression of circRNAs increases beyond the level suggested by the overall gene expression. Such cases could offer evidence for the functional importance of a circRNA. For example, mmu_circ_0000595, a circRNA of host gene RIMS2, shows a higher expression in the cortex, whereas the expression of the host gene RIMS2 remains stable in this tissue (Fig. 3d,

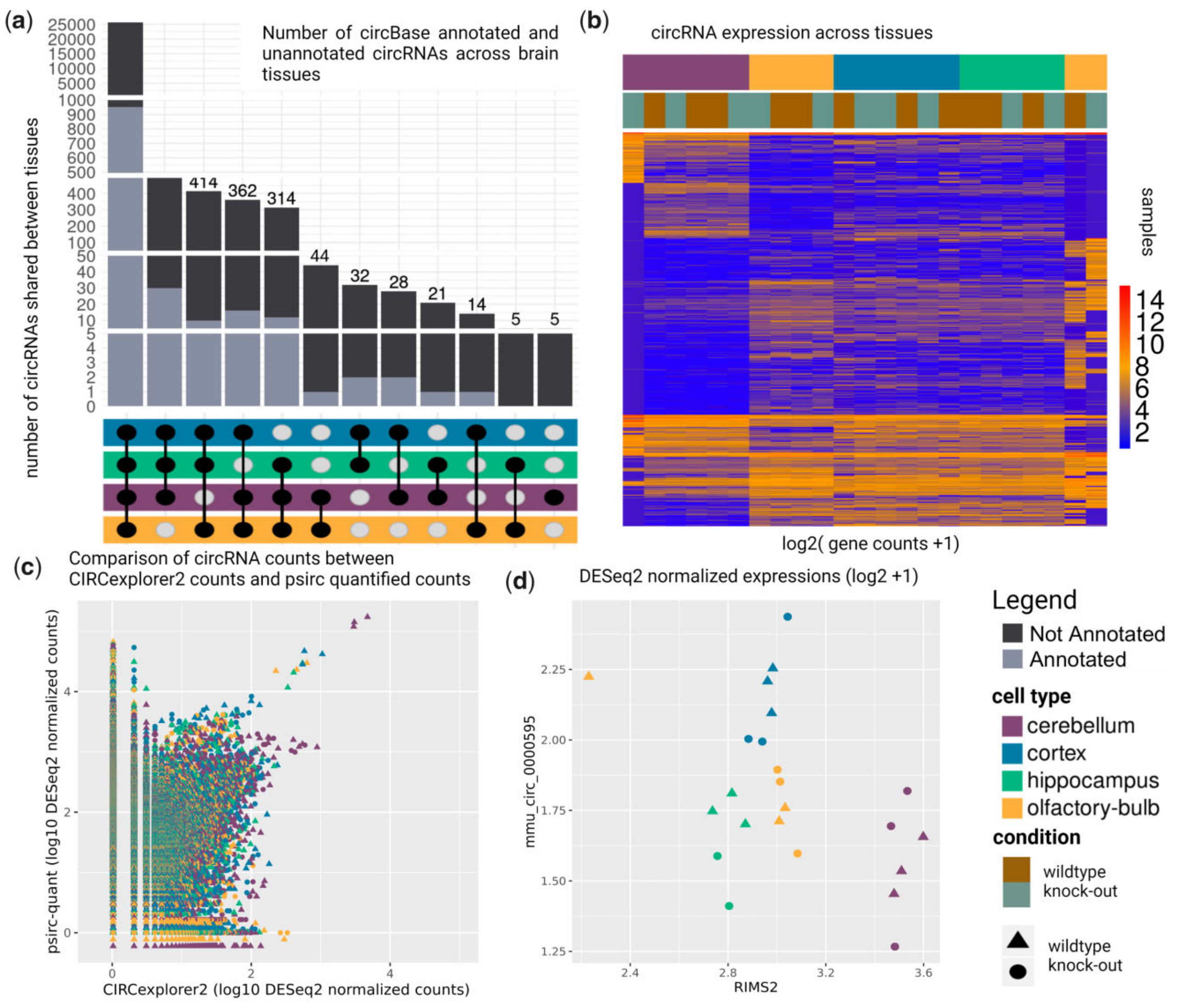


Figure 3. circRNA results of the mouse brain regions dataset. (**a**) circRNAs shared between brain regions. (**b**) Expression of circRNAs across tissues and experimental conditions. (**c**) Comparison of psirc-quant quantified circRNA counts to CIRCexplorer2 counts. (**d**) Comparison between a circRNA originating from RIMS2 and expression of the linear RIMS2 gene

see also Supplementary Fig. S6a–p). We investigate the output of SUPPA2 to explore the relationship between the quantity of circular and linear transcripts per shared gene more thoroughly. As expected, the PSI values of the linear transcripts are mostly close to 100%, whereas circular RNAs are rare and show overall very low PSI values. However, there are instances where circular transcripts show high PSI values (Supplementary Fig. S7a). Differential splicing analysis of circular transcripts revealed that only very few are significantly differentially spliced between cell types, i.e. pass both filters of a change in PSI $\geq 25\%$ and a *P*-value of ≤ 0.01 (Supplementary Fig. S7b). We can observe that two circRNAs (chr15:34600014–34625031_– with its host genes HYDIN between cortex-hippocampus and chr8:110298074–110334816_+ with its host gene NIPAL2 between hippocampus-olfactory-bulb) are considered differentially spliced in (Supplementary Fig. S7). These results are similar for normalized PSI values (Supplementary Fig. S8), where we accounted for sample-specific differences. Our results thus suggest that the splicing ratio of linear to circular RNA expression does not change between different brain regions.

3.2 A ceRNA network reveals circRNAs acting as miRNA sponges

If matched total RNA and miRNA sequencing data are provided, circRNA-sponging infers a ceRNA network using the R package SPONGE (List *et al.*, 2019) and visualizes the result (Fig. 4). Important players in this regulatory network are characterized by a large node degree, i.e. they indirectly regulate many of the connected RNAs via sequestering miRNA copies. Since the network inferred by SPONGE does not offer insights into individual samples or conditions, we subsequently computed spongEffects (Boniolo *et al.*, 2023) enrichment scores which capture the interaction of individual circRNAs and their target genes. As these scores are sample-specific, they can offer insights into condition-specific circRNA-sponging activity. spongeEffects scores can also be used as features for machine learning tasks such as classification (Boniolo *et al.*, 2023). Since the number of available samples for training is rather small here, the random forest reported subset accuracy drops considerably on the holdout set in 10-fold cross-validation. While the cortex and hippocampus are difficult to differentiate, the cerebellum can be

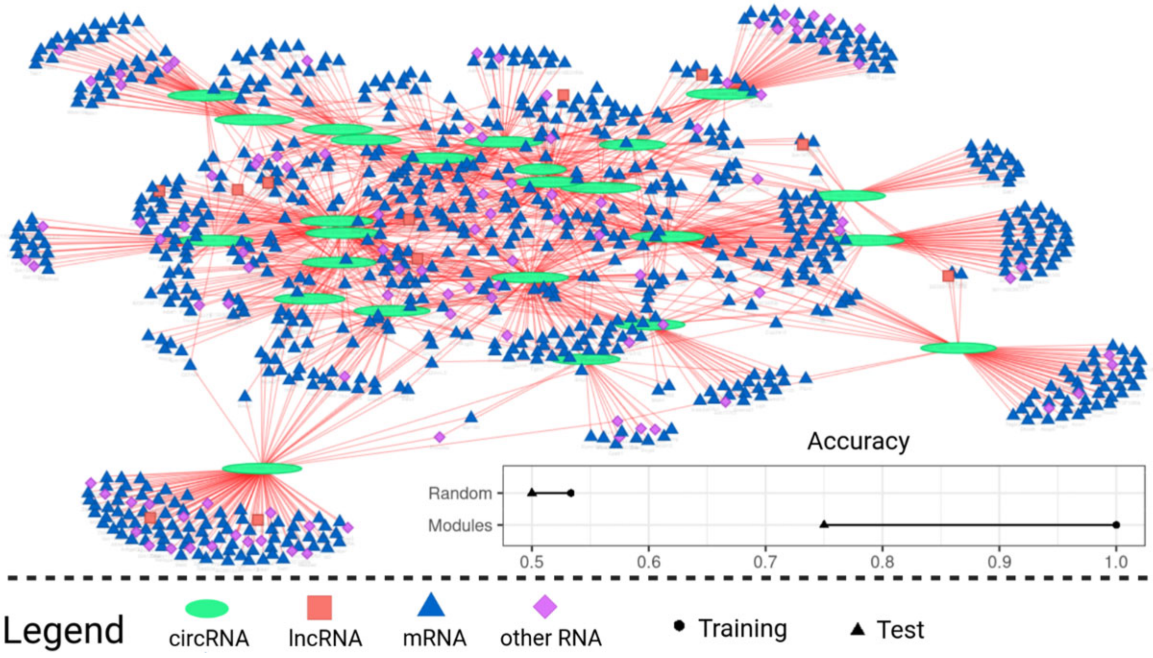


Figure 4. circRNA–ceRNA-subnetwork with the top 25 ceRNA modules ranked by the number of significant interactions (node degree in the network). For each ceRNA module consisting of the circRNA and its target genes, we computed spongEffects enrichment scores and used them as input for a random forest model. The bottom-right corner shows the subset accuracy of this model in distinguishing different brain regions on the training and test set. The results of a model trained on random modules of the same size show random performance

robustly distinguished from other brain regions (Supplementary Fig. S9). In particular, two circRNAs, chr10:9770449–9800068_– and chr10:79860969–79862010_+ stand out as distinctive features of the cerebellum (Supplementary Figs S10 and S11). While the inferred ceRNA network shows that circRNAs in the mouse brain are regulatory active through miRNA sponging, a larger number of samples is likely needed to fully resolve brain region-specific sponging activity.

3.3 Comparing the number of circRNA binding sites with respect to their length reveals two distinct clusters

miRNA sponging has long been considered a potential function of circRNAs (Hansen *et al.*, 2013). To fulfill this function, it would be beneficial for circRNAs to carry a large number of miRNA binding sites, and indeed, some known circRNAs, such as CDR1as harboring over 70 miRNA binding sites for miR-7 alone (Jiang *et al.*, 2020), fit the hypothesis well. To investigate if this is a general property of circRNAs, we systematically compared the length of a transcript to the number of binding sites, expecting to observe a larger ratio for circRNAs compared to the 3′ untranslated regions of linear transcripts, where miRNA binding sites are predominantly located (Fig. 5). While linear transcripts show a very diverse picture, circRNA length correlates well with the number of binding sites.

Compared to the prediction in 3′-UTRs (also based on TarPmiR), circRNAs show a comparably high ratio between the number of binding sites and the length. We observed two distinct circRNA clusters despite employing the same three miRNA binding site prediction methods (miRanda, TarPmiR, PITA) for all of them. We only accept a miRNA binding site for a circRNA if it was predicted by at least two prediction methods. It appears that all three miRNA binding site prediction tools were able to identify miRNA binding sites in the circRNAs of the blue cluster, while only miRanda and PITA, but not TarPmiR, could predict miRNA binding sites of the circRNAs in the red cluster. This observation was not expected as TarPmiR, in general, predicts overall more binding sites than miRanda or PITA. An interesting question is hence if TarPmiR is able to differentiate between circRNAs that are active miRNA sponges and those that have other functions. Previous research has defined three types of circular RNAs based on structural features—exonic circular (ecirc) RNAs, circular intronic RNA (ciRNA) and exon-intron circRNA (EIciRNA) (Xiao *et al.*, 2022; Yang *et al.*, 2018; Zhang *et al.*, 2022). It has been suggested that ecircRNAs function predominantly through a miRNA sponging effect in the cytoplasm, whereas other circular RNA forms (e.g. ciRNA and ElciRNA) function in the nucleus to regulate gene transcription (Li *et al.*, 2015, 2017; Zhang *et al.*, 2013). Hence, circRNAs that are functional in the nucleus could have fewer miRNA binding sites. To test alternative explanations, we checked if clusters differed by (i) biotype of the circRNA host gene (i.e. coding or noncoding gene, Supplementary Fig. S12a), (ii) genesis, i.e. the splicing method of the circRNA [ElciRNA, ciRNA, ecircRNA, Supplementary Fig. S12b (Trincado *et al.*, 2018)] or (iii) circRNA expression level (Supplementary Fig. S12c). The observed clusters did not differ in any of these categories, and further work is needed to elucidate if these results are related to other structural features. It should also be noted that TarPmiR was not trained specifically on circRNAs and that a prediction method tailored toward circRNAs should be developed when suitable experimental data become available. In addition, we investigated the relationship between the number of miRNA binding sites and the SPONGE (List *et al.*, 2019) correlation scores associated with each circRNA and found that these scores seem to have no apparent correlation to the number of miRNA binding sites, although a very large number of

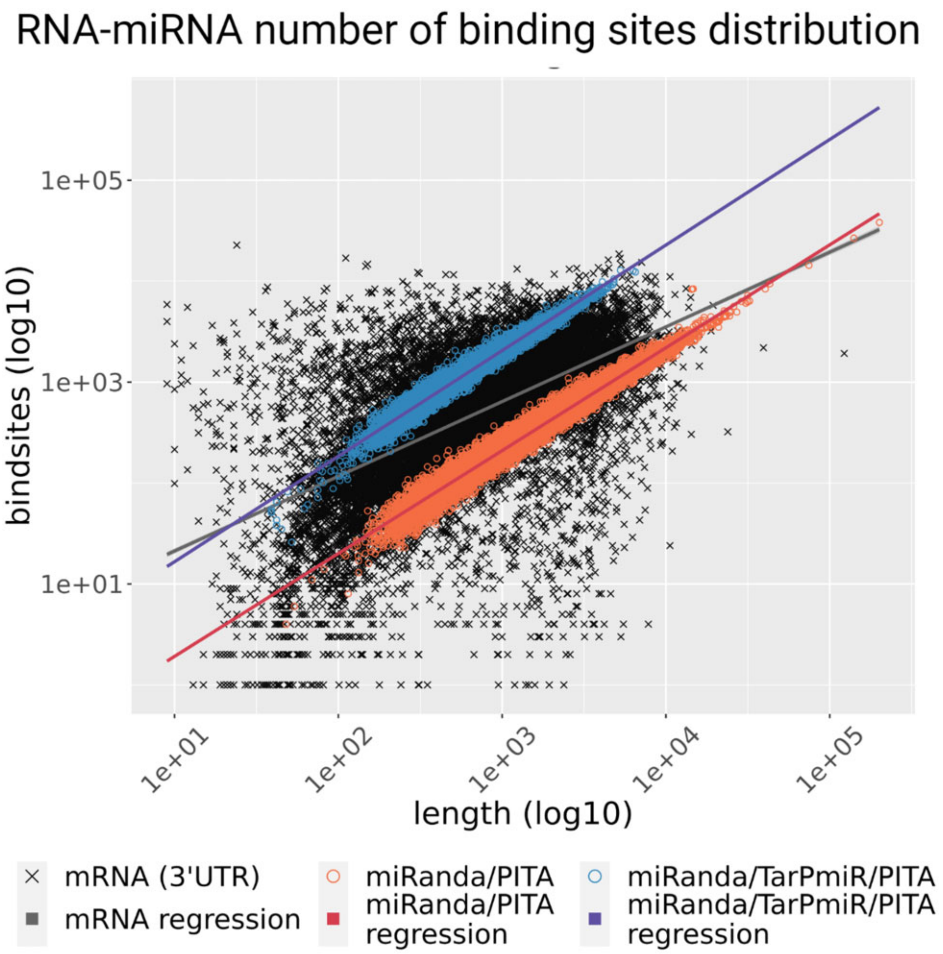


Figure 5. Number of miRNA binding sites versus transcript length for linear and circular RNA. For the 3′-UTRs of mRNAs, the number of binding sites was inferred from miRWalk 3.0. circRNA–miRNA binding sites were counted if they were reported by two of the three prediction methods employed, i.e. miRanda, TarPmiR and PITA. circRNAs form two clusters that can be explained by the different target site prediction methods used. Linear regression models were fit to each of the groups to show the trend of the association

binding sites seems to be advantageous for generating more elevated scores (i.e. over 0.5) in comparison to circRNAs with lower miRNA binding potential, that are only rarely able to reach comparably high correlation values (Supplementary Fig. S13a and b).

4 Conclusion and outlook

We developed a new circRNA processing and analysis pipeline consisting of three modules harboring multiple current state-of-the-art methods: (i) circRNA detection, (ii) miRNA detection and (iii) detection of sponging events between circRNAs, ceRNAs and miRNAs. To the best of our knowledge, it is the first comprehensive circRNA pipeline to detect, quantify and annotate circRNAs as well as to determine their sponging activity. The latter allows users to bring circRNAs into a functional context with other RNAs, such as mRNAs and lncRNAs, through a joint ceRNA network which is mediated by miRNA sponging. Wen *et al.* (2021) recently highlighted the need for a further extensive investigation into circRNAs due to their enormous potential to explain human diseases like cardiovascular, autoimmune and cancers. circRNAs are also known to be involved in brain development, brain cell differentiation and neuronal signaling (Hanan *et al.*, 2017; Piwecka *et al.*, 2017). To demonstrate the capabilities of the circRNA-sponging pipeline, we hence re-analyzed a public dataset where we investigated circRNAs across different mouse brain regions. Using our pipeline, we could offer novel insights into circRNA biology across tissues of the brain. We showed that differences in circRNA splicing could be revealed when considering the expression of circRNAs relative to the expression of a host gene, similar to how alternative splicing events are detected by considering exon or intron inclusion. Our pipeline is the first to routinely incorporate differential splicing analysis between linear and circular transcripts of the same genes, allowing to better differentiate between changes in expression and changes in splicing. We further placed our findings into the context of miRNA sponging, demonstrating that circRNA exerts regulatory control over a vast number of transcripts. Finally, we showed that the number of binding sites in circRNAs correlated well with their length and observed that TarPmiR's machine-learning strategy identifies a subset of circRNAs that could indicate promising candidates for miRNA sponging. Further work is needed to investigate if these two classes represent structurally different circRNAs, such as ecircRNAs, ciRNAs or ElciRNAs, or if this observation can be explained by differences in the miRNA prediction methods with no biological implication at all.

In the future, we plan to extend the circRNA-sponging pipeline with additional features. As various functions other than miRNA sponging have been attributed to circRNAs (Nielsen *et al.*, 2022), we see room for expanding the features toward, e.g. investigating the protein-coding potential of circRNAs (Miao *et al.*, 2021). We further seek to integrate circRNA-sponging into ongoing community efforts such as nf-core (Ewels *et al.*, 2020) to build up long-term support for maintaining and expanding this pipeline. In summary, the circRNA-sponging pipeline is a powerful tool to detect, investigate and analyze circRNAs and their sponging effects and thus, it helps researchers consider circRNAs as a routine aspect in RNA-seq and miRNA-seq data analysis.

Acknowledgements

The authors thank Bahar Tercan, Mauro A. A. Castro, A. Gordon Robertson, Katherine A. Hoadley and Victoria K. Cortessis for the fruitful discussions. They thank Christina Trummer for the logo design. Figures were created with https://www.biorender.com. Parts of the images were designed using resources from Flaticon.com.

Author Contributions

Markus Hoffmann (Conceptualization [lead], Investigation [lead], Methodology [lead], Project administration [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Leon Schwartz (Data curation [lead], Formal analysis [lead], Methodology [lead], Software [lead], Validation [lead], Writing—original draft [supporting], Writing—review & editing [supporting]), Octavia Ciora (Methodology [equal], Software [lead], Visualization [lead], Writing—original draft [supporting], Writing—review & editing [supporting]), Nico Trummer (Data curation [equal], Formal analysis [supporting], Software [supporting]), Lina-Liv Willruth (Data curation [lead], Resources [equal]), Jakub Jankowski (Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal]), Hye-Kyung Lee (Methodology [supporting], Writing—original draft [equal], Writing—review & editing [equal]), Jan Baumbach (Funding acquisition [lead], Writing—original draft [equal], Writing—review & editing [equal]), Priscilla Furth (Validation [equal], Writing—original draft [lead], Writing—review & editing [lead]), Lothar Hennighausen (Funding acquisition [lead], Writing—original draft [lead], Writing—review & editing [lead]) and Markus List (Conceptualization [lead], Funding acquisition [lead], Methodology [lead], Project administration [lead], Writing—original draft [lead], Writing—review & editing [lead]).

Funding

This work was supported by the Technical University Munich—Institute for Advanced Study, funded by the German Excellence Initiative. This work was supported in part by the Intramural Research Programs (IRPs) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). J.B. was partially funded by his VILLUM Young Investigator [grant number 13154]. This project was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [422216132]. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the *e: Med* research and funding concept [*grant numbers 01ZX1908A/01ZX2208A* and *grant numbers 01ZX1910D/01ZX2210D*]. This project has received funding from the European Union's Horizon 2020 research and innovation program [grant agreement number 777111]. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Conflict of interest

None declared.

Data availability

Data are available in GEO under GSE100265 (rRNA-depleted RNA-seq) and GSE93129 (miRNA). circRNA-sponging is available under the GPL v.3.0 license at: https://github.com/biomedbigdata/circRNA-sponging. Dockerhub image is available under: https://hub.docker.com/r/bigdatainbiomedicine/circrna-sponging. The results presented in this manuscript are available as RData objects at: https://doi.org/10.6084/m9.figshare.21864948.

References

Boniolo,F. *et al.* (2023) spongEffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape. *Bioinformatics*, **39**, btad276.

Bray,N.L. *et al.* (2016) Erratum: near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 888.

Chen,L. *et al.* (2021) The bioinformatics toolbox for circRNA discovery and analysis. *Brief. Bioinform.*, **22**, 1706–1728.

Ding,J. *et al.* (2016) TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics*, **32**, 2768–2775.

Dobin,C.A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Dong,R. *et al.* (2018) CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison. *Genomics Proteomics Bioinf.*, **16**, 226–233.

Enright,A.J. *et al.* (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.

Ewels,P.A. *et al.* (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.*, **38**, 276–278.

Frazee,A.C. *et al.* (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.

Friedländer,M.R. *et al.* (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.

Galea,J.M. *et al.* (2011) Dissociating the roles of the cerebellum and motor cortex during adaptive learning: the motor cortex retains what the cerebellum learns. *Cereb. Cortex*, **21**, 1761–1770.

Glažar,P. *et al.* (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.

Hanan,M. *et al.* (2017) CircRNAs in the brain. *RNA Biol.*, **14**, 1028–1034.

Hansen,T.B. *et al.* (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, **495**, 384–388.

Hänzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

Hoffmann,M. *et al.* (2021) SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions. *NAR Cancer*, **3**, zcaa042.

Hsu,S.-D. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res.*, **39**, D163–D169.

Huang,H.-Y. *et al.* (2020) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.

Jeck,W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.

Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–461.

Jerby-Arnon,L. *et al.* (2018) A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, **175**, 984–997.e24.

Jiang,C. *et al.* (2020) The emerging picture of the roles of CircRNA-CDR1as in cancer. *Front. Cell Dev. Biol.*, **8**, 590478.

Karagkouni,D. *et al.* (2020) DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic Acids Res.*, **48**, D101–D110.

Kartha,R.V. and Subramanian,S. (2014) Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. *Front. Genet.*, **5**, 8.

Kertesz,M. *et al.* (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Kristensen,L.S. *et al.* (2018) Circular RNAs in cancer: opportunities and challenges in the field. *Oncogene*, **37**, 555–565.

Kuhn,M. (2008) Building predictive models in R using the caret package. *J. Stat. Soft.*, **28**, 1–26.

Lasda,E. and Parker,R. (2014) Circular RNAs: diversity of form and function. *RNA*, **20**, 1829–1842.

Li,Z. *et al.* (2015) Exon-intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.*, **22**, 256–264.

Li,Z. *et al.* (2017) Corrigendum: exon–intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.*, **24**, 194.

Li,M. *et al.* (2018) A circular transcript of ncx1 gene mediates ischemic myocardial injury by targeting miR-133a-3p. *Theranostics*, **8**, 5855–5869.

List,M. *et al.* (2019) Large-scale inference of competing endogenous RNA networks with sparse partial correlation. *Bioinformatics*, **35**, i596–i604.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Memczak,S. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.

Meng,X. *et al.* (2017) CircPro: an integrated tool for the identification of circRNAs with protein-coding potential. *Bioinformatics*, **33**, 3314–3316.

Mester-Tonczar,J. *et al.* (2020) Circular RNAs in cardiac regeneration: cardiac cell proliferation, differentiation, survival, and reprogramming. *Front. Physiol.*, **11**, 580465.

Miao,Q. *et al.* (2021) Coding potential of circRNAs: new discoveries and challenges. *PeerJ*, **9**, e10718.

Min,H. and Yoon,S. (2010) Got target? Computational methods for microRNA target prediction and their extension. *Exp. Mol. Med.*, **42**, 233–244.

Nielsen,A.F. *et al.* (2022) Best practice standards for circular RNA research. *Nat. Methods*, **19**, 1208–1220.

Piwecka,M. *et al.* (2017) Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science*, **357**, eaam8526.

Qu,S. *et al.* (2015) Circular RNA: a new star of noncoding RNAs. *Cancer Lett.*, **365**, 141–148.

Rybak-Wolf,A. *et al.* (2015) Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell*, **58**, 870–885.

Salmena,L. *et al.* (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–358.

Sticht,C. *et al.* (2018) miRWalk: an online resource for prediction of microRNA binding sites. *PLoS One.*, **13**, e0206239.

Suzuki,H. *et al.* (2006) Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. *Nucleic Acids Res.*, **34**, e63.

Szabo,L. and Salzman,J. (2016) Detecting circular RNAs: bioinformatic and experimental challenges. *Nat. Rev. Genet.*, **17**, 679–692.

Szabo,L. *et al.* (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.*, **16**, 126.

Trincado,J.L. *et al.* (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40.

Weinstein,J.N. *et al.*; Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Wen,G. *et al.* (2021) The potential of using blood circular RNA as liquid biopsy biomarker for human diseases. *Protein Cell.*, **12**, 911–946.

Westholm,J.O. *et al.* (2014) Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.*, **9**, 1966–1980.

Xiao,J. *et al.* (2022) Circular RNAs acting as miRNAs' sponges and their roles in stem cells. *J. Clin. Med. Res*, **11**, 2909.

Yang,L. *et al.* (2018) Circular RNAs and their emerging roles in immune regulation. *Front. Immunol.*, **9**, 2977.

Yu,C.-Y. and Kuo,H.-C. (2019) The emerging roles and functions of circular RNAs and their generation. *J. Biomed. Sci.*, **26**, 29.

Yu,K.H.-O. *et al.* (2021) Quantifying full-length circular RNAs in cancer. *Genome Res.*, **31**, 2340–2353.

Zhang,X.-O. *et al.* (2014) Complementary sequence-mediated exon circularization. *Cell*, **159**, 134–147.

Zhang,X.-O. *et al.* (2016) Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.*, **26**, 1277–1287.

Zhang,Y. *et al.* (2013) Circular intronic long noncoding RNAs. *Mol. Cell.*, **51**, 792–806.

Zhang,Y. *et al.* (2022) Emerging functions of circular RNA in the regulation of adipocyte metabolism and obesity. *Cell Death Discov.*, **8**, 268.

Zhang,Z. *et al.* (2018) Circular RNAs: promising biomarkers for human diseases. *EBioMedicine*, **34**, 267–274.

5. GENERAL DISCUSSION AND OUTLOOK

The increasing availability of biological data could help to improve our comprehension of biological systems. Biological data provides a snapshot of an organism or cell at a specific time under particular circumstances. This data also brings complications, such as bias, noise, and natural variability, making it complicated to analyze and interpret. Additionally, due to financial and timely complications, biological data is produced for mostly a small number of samples and a certain OMICS or a restricted number of OMICS layers in the central dogma of molecular biology. Consequently, important information about the natural variability of the samples and the global implications of the findings could be missed in this scenario. From a bioinformatics perspective, this comes with further restrictions since most tools only integrate one OMICS layer, and Multi-OMICS approaches are still scarce for particular experimental settings. Future directions in wet-lab, clinical, and bioinformatics research should include more comprehensive cooperation. This collaboration would result in methods that integrate more OMICS into one study to grasp a more global view of the implications of the findings.

Numerous computational methods have been developed to address various research questions in wet-lab and bioinformatics. However, one faces challenges in their adoption and usability for laboratory personnel with limited computational experience as they struggle with adoption and usability [268]. Many developed computational methods are not properly documented [269], implemented [270], and maintained after publication [271,272], costing time and effort to apply these methods to new data. Some methods cannot be utilized anymore due to missing scripts, intermediate formatting steps, or poor implementation and documentation, rendering them ultimately useless for analyses of new data [273,274]. Converting one data format into another is time-consuming [275], especially for non-computational experts, and includes the risk of introducing errors during faulty conversions [276,277]. Furthermore, many tools are tailored to work with data in (i) specific formats (i.e., very commonly, each method introduces its own input format [278]), (ii) a specific operating system (e.g., Windows, a specific Linux distribution,...), or (iii) specific technical dependencies (e.g., version of a programming language, version of various software packages) restricting their usability for inexperienced users. Additionally, some methods provide very limited, if any, formatting instructions in their documentation. Only very few methods provide converting tools from common formats into their specific format. Many tools also lack documentation of their resulting output format. The above-mentioned reasons ultimately cost time and resources, including the budget to hire experienced bioinformaticians. To tackle these issues, future efforts in bioinformatics should focus on (i) enhancing the interoperability of these tools by providing detailed data formatting instructions or using already established data formats, (ii) creating a fully functional pipeline instead of a staple of scripts, (iii) creating robust documentation, (iv) ensuring compatibility across various systems and software packages, and (v) creating a user-friendly deployment version of the software for non-computational experts [273,274].

In this thesis, one focus is to provide ready-to-use pipelines for frequently used tools in epigenomics and transcriptomics, allowing laboratory personnel with limited computational knowledge to execute. The main purpose of this thesis is to develop analysis pipelines that enable researchers to gain further insights into two regulatory layers of the central dogma of molecular biology: the epigenomics and transcriptomics layers. On the epigenomics layer, I developed TF-Prioritizer, a pipeline that captures the differential activity of TFs between conditions. On the transcriptomics layer, I developed circRNA-sponging, a pipeline that detects circRNAs and investigates their sponging capability. In the remainder of this thesis, I discuss the limitations, considerations, and future plans of the two pipelines introduced in this thesis.

5.1. TF-Prioritizer: a java pipeline to prioritize condition-specific transcription factors

In this thesis, I introduced TF-Prioritizer, a pipeline that utilizes ChIP-seq, ATAC-seq, or DNase-seq in combination with RNA-seq to prioritize potential condition-specific TFA. TF-Prioritizer is the first pipeline that automatically confirms its predictions through the visualization of automatically found and downloaded context-specific experimental data in a feature-rich web application. TF-Prioritizer is the first tool for this kind of analysis and experimental setting that is completely dockerized and documented to ensure easier usability.

Limitations and considerations for computational methods investigating transcription factor activity

With TF-Prioritizer, I aim to detect significant differential TFA in conditions (e.g., healthy versus disease). However, TFA methods have limitations that one needs to consider. In general, many methods used for TFA estimation depend on information related to the binding of TFs. This data is only available for TFs that can be measured or predicted. Measuring techniques, such as ChIP-seq, only work well when a high-affinity antibody is available [279]. Moreover, these techniques are limited by the need for large numbers of similar samples and can only measure one TF at a time. This limitation makes it challenging to get a full picture of TF binding [141,142].

Furthermore, the ways TFs bind also vary, which has led to their classification into (i) pioneers, (ii) settlers, and (iii) migrants. (i) Pioneers have the unique ability to bind to closed chromatin, which is compacted DNA that is typically inaccessible. By binding to this condensed form of DNA, pioneer TFs can initiate the process of 'opening up' the chromatin, making it more accessible to other proteins and TFs. This chromatin remodeling ability of pioneer TFs sets the chromatin environment for further gene regulation activities [280]. (ii) Once the chromatin is accessible, settlers come into play. They primarily bind to their motifs in this open chromatin. Unlike pioneers, settlers don't possess the capability to bind to closed chromatin or facilitate its opening. Instead, they further stabilize the open chromatin state and play roles in fine-tuning the gene expression process [281–283]. (iii) Migrants are more selective in their binding. They do not necessarily bind to all available motifs in the accessible chromatin. Instead, they bind to a subset of their accessible motifs. Their selective binding behavior can be influenced by various factors, including the presence of other TFs or specific conditions within the cell [284]. However, most TFA estimation tools overlook these distinctions [281,283]. The presence of similar binding motifs across different

TFs also complicates accurate predictions [285]. Also, TF-Prioritizer falls short in regard to this expectation due to the sheer complexity of implementing such features.

Several models wrongly assume that higher TF expression implies greater regulatory significance. Studies have shown that this assumption isn’t always valid [286,287]. This oversight might be due to the differences between RNA and protein levels and the influence of modifications after transcription [288–290]. Also, models often oversimplify by assuming a direct relationship between TFA and gene regulation. They usually don’t account for the number of different TFs or their isoforms [291]. Another aspect often overlooked is the role of chromatin compartmentalization. Chromatin compartmentalization refers, within the nucleus, to its organization in distinct compartments and domains. Among these are the A/B Compartments, which are expansive chromatin regions [292]. The 'A' compartments teem with genes and are bustling with transcription activity, whereas the 'B' compartments have fewer genes and display reduced activity [292]. Another integral structure is the Topologically Associating Domains (TADs) [293]. These represent genomic regions that have a propensity for internal interactions, meaning that DNA sequences within a TAD have a heightened likelihood of interacting with each other rather than with sequences situated outside the domain. These structures can influence gene regulation by creating specialized environments, and while TFs help form these compartments, they’re also influenced by them [294]. TF-Prioritizer is capable of utilizing Hi-C data (i.e., provides a comprehensive snapshot of the three-dimensional interactions between different regions of the genome within the cell nucleus) to accommodate chromatin compartmentalization [295,296].

Efforts to validate TFA tools have shown mixed results, with many performing no better than random chance [297–299]. Yet, despite these challenges, computational tools for TFA estimation have produced results that align with existing literature [300,301], highlighting the urgent need for better cooperation between computational biology and biological wet-bench science.

Future plans for TF-Prioritizer

While TF-Prioritizer has streamlined the process of detecting differential TF activity, its utility currently depends heavily on the nfcore pipelines for the preprocessing of raw sequencing FASTQ files. This adds complexity and affects the overall user experience. To resolve this, I aim to develop the first end-to-end pipeline that integrates TF-Prioritizer, including the nfcore pipeline's preprocessing capabilities. This proposed solution could enhance usability by offering a seamless workflow from raw sequencing data analysis to a visualized differential TF activity interpretation that can be handled with a few clicks.

As part of our future endeavors, I intend to extend the functionality of the end-to-end pipeline to include the prediction of condition-specific associations among active (super-)enhancers, promoters, transcription factors, and target genes. The inclusion of such predictive features in our pipeline would facilitate a more comprehensive view of the gene regulatory landscape and could be useful for researchers in the planning of further wet-bench experiments. I plan to establish a comprehensive database incorporating open chromatin state data and RNA-seq for primary cells, tissues, and cell lines from Homo sapiens and Mus musculus from the ENCODE database [199]. This database will serve as a resource for active associations between enhancers, promoters, transcription factors, and target genes across different conditions and cell types.

TF-Prioritizer currently utilizes data from the epigenomics and transcriptomics layers, making it already a Multi-OMICS tool. However, TF-Prioritizer could miss important insights into the genomics layer. In the future, I intend to include the genomics layers to identify genetic variations (i.e., SNPs) and interactions between genetic variations that could be the underlying causes of deregulated associations between CREs, TFs, and target genes. SNPs inside CREs, TFs, or target genes could cause serious harm to a biological system.

SNPs inside CREs could render a TFBS motif dysfunctional by blocking the binding of a TF. This directly influences a target gene's expression (e.g., depending on the importance of the target gene, this could be fatal) [302]. SNPs can also affect the methylation of DNA and could, therefore, affect the accessibility of CREs and target genes and ultimately affect the expression levels of the target gene [303]. Additionally, SNPs within regulatory elements or the mRNA sequence of a TF could impact expression levels, the protein structure of the TF, and TF binding to TFBS, resulting in altered expression levels of target genes [303,304]. However, it could be the case that one SNP individually has little or no effect on the expression level of a target gene, but a combination of such SNPs (e.g., one SNP in the CRE and one SNP in the TF combination) could cause a serious change in binding affinity of the TF to the TFBS and target gene expression levels. This effect is called epistatic interaction of SNPs [305–308]. Since more than 84.7 million SNPs are known [309], detecting higher-order interaction (i.e., SNP combination of two SNPs or higher that have an effect on a disease) leads to a combinatorial explosion and is, with currently available hardware and algorithms, not feasible. I plan to develop algorithms that could make such investigation feasible by employing network-medicine approaches. In the future, I plan to transform the predicted associations between CREs, TFs, and target genes of TF-Prioritizer into an SNP-SNP interaction network by mapping SNPs from dbSNP [310] to the predicted associations and thereby limiting the search space with network-medicine algorithms. With this approach, I intend to potentially understand the underlying genetic cause of deregulated associations (epigenomics and transcriptomics layer) on the grounds of SNPs (genomics layer) of heritable diseases, creating a Multi-OMICS integration of three OMICS layers.

5.2. circRNA-sponging: a pipeline for extensive analysis of circRNAs and their miRNA sponging effects

In this thesis, I introduced the circRNA-sponging pipeline, an end-to-end pipeline to detect, investigate, and analyze circRNAs and their sponging effects from totalRNA-seq and miRNA-seq FASTQ files eliminating additional preprocessing steps. This user-friendly and dockerized pipeline helps researchers consider circRNAs as a routine aspect.

Limitations and considerations for computational detection and functional investigation of circRNAs

With circRNA-sponging, I aim to detect and investigate the functions of circRNAs that have become increasingly relevant in medical research due to their potential roles in various biological processes and diseases [223]. The identification process includes a risk of detecting false positive circRNAs. Some methods could mistakenly classify linear RNAs as circRNAs due to overlaps in sequences or other confounding factors [311]. Zeng et al. performed a benchmarking study of several circRNA detection tools (such as CIRCexplorer2) and showed that there was room for improvement [312]. Jakobi et al., on the other hand, concluded in their review that circRNA detection algorithms have developed to a stage where they can make high-quality predictions of circRNAs in datasets [238,239,313]. In any case, determining the authenticity of computationally predicted circRNAs experimentally is crucially needed (see Sec. 5.2 “Opportunities for experimental investigation of circRNAs”). Additionally, a universally accepted naming system for circRNAs is important since, currently, many different ways to name circRNAs are established and could, hence, lead to confusion [111,245,247,249,314]. With respect to the functional analysis of circRNAs, to date, only a few circRNAs have been associated with a defined mechanism and function (e.g., CDR1as [315]).

A review of several studies has shown that circRNAs could be important for clinical use [316]. A comprehensive pan-cancer study analyzed circRNA expression in over 2000 patient samples, revealing distinct circRNA expression profiles across different cancer types [317]. This, combined with their inherent stability, underscores their potential as cancer biomarkers [318]. Furthermore, circRNAs are understood to play roles in cancer independent of their linear RNA counterparts. For instance, a study on prostate cancer revealed that several circRNAs were vital for cell proliferation, even when their corresponding linear transcripts were not [319]. This highlights the oncogenic or antitumor roles of circRNAs, highlighting their therapeutic potential. Another popular example is circRNA circAGFG1, which has been observed to be upregulated in triple-negative breast cancer tissues. circAGFG1 has demonstrated oncogenic properties by sponging miR-195-5p [320]. Techniques such as short hairpin RNA (shRNA) targeting circAGFG1 have effectively lowered cell proliferation, migration, and invasion while also augmenting apoptosis in vitro. These interventions also demonstrated a decrease in tumor growth, angiogenesis, and metastasis in vivo [320]. Although silencing with shRNA seems to be promising, circRNA-based therapeutic methods with shRNAs have currently only advanced to preclinical studies due to the prime concern of off-target gene silencing, which could be fatal [316,321–323].

Opportunities for experimental investigation of circRNAs

Computational detection and computational functional investigation of circRNAs is the first step to determining their possible applications in clinical therapies. However, experimental validation of hypotheses generated by computational pipelines is strictly necessary. In this chapter, I present a selection of experiments that could be performed to confirm computationally generated hypotheses with circRNAs.

Quantitative PCR with reverse transcription (RT-qPCR) can be used to quantify circRNAs experimentally by designing primers that span the BSJ to amplify circRNA-specific sequences. Additionally, RT-qPCR allows for the simultaneous analysis of linear RNAs, supporting the assessment of circRNA-to-linear RNA ratios [324]. However, Northern blots (i.e., separating RNA molecules by size through electrophoresis and detecting specific sequences by hybridization with a labeled complementary probe) can validate circRNAs without an intervening reverse transcription step which could introduce errors [111,325–327].

After validation that the computationally detected circRNA is real and is expressed in the studied cells or tissues, the functional roles of circRNAs can be studied using methods such as RNA interference or genetic manipulation. For example, RNA interference techniques could selectively target the circular form without affecting the linear transcripts if siRNAs are designed to target the unique BSJ. Genetic deletion could elucidate the physiological role of circRNAs in cellular processes, as demonstrated by the study of CDR1as [315,328]. Additionally, RNA immunoprecipitation followed by sequencing can help identify the RNA-binding proteins associated with specific circRNAs [48,223]. The manipulation of circRNA levels, either through depletion or overexpression, could help to elucidate their functional roles. RNA-guided RNA-targeting systems can be employed for sequence-specific degradation of circRNAs, particularly targeting the BSJ [329,330]. By contrast, genome editing can be used to modify or delete genomic loci or specific elements that facilitate circRNA formation, thereby reducing or preventing back-splicing events [331,332]. Both methods must be carefully optimized to achieve specificity while minimizing off-target effects. It is also crucial to validate the efficiency of these approaches using quantitative methods, such as RT–qPCR, to confirm a reduction in circRNA expression without affecting the linear counterparts [111].

When it comes to experimental validation of miRNA binding (e.g., circRNA-miRNA binding), one must consider the stoichiometry (i.e., properties of the miRNA such as concentration levels, binding affinities, etc.) of the miRNA [333]. The ratio between miRNA and its target RNAs can either enhance or dampen the regulatory impact [334–336]. Binding affinity (i.e., by sequence complementarity and specific motifs) further modulates this interaction [337,338]. Multiple miRNA binding sites on a single RNA can lead to synergistic effects, while competition among miRNAs for the same site introduces another layer of complexity [339,340]. The use of antisense oligonucleotides (ASOs), including anti-miRs, has shown clinical promise in treating certain diseases, yet the extension to miRNA targeting remains challenging [341–343]. The issue of stoichiometry in miRNA-anti-miR interactions is crucial for gene regulation, but there's a prevalent presumption in the literature that anti-miRs will unambiguously and specifically inhibit their target miRNAs, an assumption that risks oversimplification of such a complicated interaction [333,344–349]. Despite the interest in miRNA research, the slow pace of clinical development for miRNA-targeting drugs highlights the need for basic experiments beyond computational analysis, particularly in validating the stoichiometry and specificity of miRNA-anti-miR interactions in in-vivo and in-vitro experiments [333,347–351].

As an example, Zhu et al. [352] investigated if the circRNA circPan3 mediates the promotion of intestinal stem cell self-renewal by utilizing several of the aforementioned methods to experimentally investigate circPan3's functions. Zhu et al. selected the circRNA circPan3 due to its high expression in intestinal stem cells by generating reporter mice (i.e., generation of mice that have a fluorescent marker at the circRNA) [352]. circPan3 was further confirmed to be a circRNA through the utilization of PCR, RNA sequencing, and RNase R treatment. The authors also silenced circPan3 through shRNA targeting and observed no measurable effects. However, they generated circPan3-depleted mice by genetically deleting the circRNA while confirming no measurable off-target effects. They concluded that with this method, they could confirm a measurable physiological change [352]. In conclusion, one can see that not one but multiple experimental methods that are labor intensive need to be employed to understand the functions of circRNAs.

Future plans for circRNA-sponging

In the future, I plan to incorporate the circRNA-sponging pipeline into the nfcore's circRNA pipeline in order to reach a broader user base. This integration would enhance usability by allowing researchers to seamlessly move from the analysis of circRNA sequences to the understanding of their sponging capabilities in a streamlined workflow.

To further enhance the interpretability of the circRNA-sponging pipeline, I aim to develop an automatically generated interactive web application as a part of our pipeline. This application will enable users to visualize complex gene regulatory networks and interactions effectively. Moreover, the application will feature links to multiple external resources, encouraging in-depth downstream investigation of the results.

In a future project, I aim to investigate the roles of circRNA in estrogen receptor (ER, i.e., plays a substantial part in female breasts) alpha signaling, particularly in the context of breast cancer pathogenesis. Furthermore, I intend to explore how changes in ER expression and intra-mammary estrogen levels impact the expression of circular RNAs, utilizing comprehensive databases for the role of circular RNAs in estrogen signaling and breast cancer risk. While the study of estrogen links to breast cancer risk has primarily focused on cisgender women, I acknowledge that breast cancer can develop across a spectrum of gender expressions and hormonal exposure histories, including transgender women and men. Therefore, my future research will not only have significant implications for cisgender women but will also contribute to a broader understanding of breast cancer across diverse gender expressions. Moreover, I plan to extend our study across five different species to conduct trans-species explorations and search for circRNAs in ER signaling that are conserved across species.

I furthermore plan to extend the circRNA-sponging pipeline with an automatic calculation of false-positively detected circRNAs. For this, I will accept polyA-enriched RNA-seq (i.e., no circRNA should be found) and totalRNA-seq (i.e., circRNAs should be found). I will run the detection pipeline on both datasets and compare their results.

5.3. General conclusion

In general, Bioinformatics and the development of robust computational tools for biological data analysis have been proven a crucial addition to purely wet-lab-driven research to advance our understanding of biological systems. With a growing amount of data in the future to analyze, it is our strong opinion that computational work in the biomedical field will become indispensable. In our perspective, I advocate for a model where every wet-bench scientist is partnered with an analytical bioinformatician capable of utilizing the latest analysis tools. This direct collaboration could ensure that responsibilities are divided equally between wet-bench experiments and computational analysis, resulting in more shared first-author manuscripts, which, in my opinion, should be honored equally as a single first-author publication to all involved individuals. Building upon our previous point, I strongly believe that it is vital to encourage the establishment of dedicated laboratories focused on professional computational method development and software architecture. These labs would specialize in creating new tools for analyzing biological data, which is becoming increasingly complex and voluminous. This three-pronged approach - wet-bench research, initial data analysis, and methodological innovation - would ensure that I not only generate data but are also equipped to extract complex insights from it.

This thesis has focused on the development and enhancement of user-friendly, inclusive, and comprehensive pipelines that tackle the complexities of biological data. The work carried out has produced tools that enable the integration of diverse OMICS layers, facilitating a deeper understanding of biological systems. Future endeavors related to the work presented in this thesis include the integration of genomics layers into TF-Prioritizer and enhancing the circRNA-sponging pipeline's reach and interpretability.

By enhancing the predictions of TF-Prioritizer with data from the genomics layer, I aim to understand the cause of deregulated associations of CREs, TFs, and the expression of target genes on the grounds of SNPs of heritable diseases, creating a Multi-OMICS integration of three OMICS layers. With respect to the circRNAs, I aim to explore the roles of circRNA in estrogen receptor alpha signaling in the context of breast cancer pathogenesis across diverse gender expressions, contributing to a broader understanding of breast cancer.

In conclusion, the symbiotic growth of computational resources and the expanding availability of biological data offer a promising landscape for advancing our collective scientific understanding of biological systems.

PUBLICATION AND PREPRINT LIST

Peer-reviewed publications

1. **Hoffmann, M.***, Pachl, E.*, Hartung, M.*, Stiegler, V.*, Baumbach, J., Schulz, M. H., & List, M. (2021). SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions. *NAR Cancer*, *3*(1), zcaa042. PMCID: PMC8210024
2. Blumenthal, D. B., Baumbach, J., **Hoffmann, M.**, Kacprowski, T., & List, M. (2020). A framework for modeling epistatic interaction. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btaa990. PMID: 33252645
3. Hernández-Lorenzo, L., **Hoffmann, M.**, Scheibling, E., List, M., Matías-Guiu, J. A., & Ayala, J. L. (2022). On the limits of graph neural networks for the early diagnosis of Alzheimer's disease. *Scientific Reports*, *12*(1), 17632. PMCID: PMC9587223
4. Boniolo, F.*, **Hoffmann, M.***, Roggendorf, N., Tercan, B., Baumbach, J., Castro, M. A. A., Robertson, A. G., Saur, D., & List, M. (2023). spongEffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape. *Bioinformatics* , *39*(5), btad276. PMCID: PMC10220456
5. **Hoffmann, M.***, Trummer, N.*, Schwartz, L., Jankowski, J., Lee, H. K., Willruth, L.-L., Lazareva, O., Yuan, K., Baumgarten, N., Schmidt, F., Baumbach, J., Schulz, M. H., Blumenthal, D. B., Hennighausen, L., & List, M. (2023). TF-Prioritizer: a Java pipeline to prioritize condition-specific transcription factors. *GigaScience*, *12*, giad026. PMCID: PMC10155229
6. **Hoffmann, M.***, Schwartz, L.*, Ciora, O.-A.*, Trummer, N., Willruth, L.-L., Jankowski, J., Lee, H. K., Baumbach, J., Furth, P., Hennighausen, L., & List, M. (2023). circRNA-sponging: a pipeline for extensive analysis of circRNA expression and their role in miRNA sponging. *Bioinformatics Advances*, vbad093. PMCID: PMC10359604

Preprints (currently under revision)

7. Maier, A., Hartung, M., Abovsky, M., Adamowicz, K., Bader, G. D., Baier, S., Blumenthal, D. B., Chen, J., Elkjaer, M. L., Garcia-Hernandez, C., **Hoffmann, M.**, Jurisica, I., Kotlyar, M., Lazareva, O., Levi, H., List, M., Lobentanzer, S., Loscalzo, J., Malod-Dognin, N., … Baumbach, J. (2023). Drugst.One -- A plug-and-play solution for online systems medicine and network-based drug repurposing. *ArXiv*. https://doi.org/10.1101/2020.03.22.002386

* shared first author

ACCOMPLISHED COURSES

1. Leadership and personality (Technical University of Munich, Graduate School)
2. Basic certificate for teaching and mentoring on a university level (Technical University of Munich, PROLehre)
3. Advanced certificate for teaching and mentoring on a university level (Technical University of Munich, PROLehre)
4. Expert certificate for teaching and mentoring on a university level (Technical University of Munich, PROLehre)
5. Personal leadership development (National Institutes of Health, Foundation for Advanced Education in the Sciences)
6. Leading with emotional intelligence (National Institutes of Health, Foundation for Advanced Education in the Sciences)

REFERENCES

1. Santos AX da S dos, dos Santos AX da S, Liberali P. From single cells to tissue self‐organization. The FEBS Journal. 2019. pp. 1495–1513. doi:10.1111/febs.14694

2. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Ann Hum Biol. 2013;40: 463–471.

3. Gartner LP, Hiatt JL. Color Textbook of Histology E-Book. Elsevier Health Sciences; 2006.

4. Khan YS, Farhana A. Histology, Cell. StatPearls Publishing; 2023.

5. Zhang W, Walker E, Tamplin OJ, Rossant J, Stanford WL, Hughes TR. Zfp206 regulates ES cell gene expression and differentiation. Nucleic Acids Res. 2006;34: 4780–4790.

6. Alberts B. Molecular Biology of the Cell. Garland; 2004.

7. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. An Overview of Gene Control. Garland Science; 2002.

8. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227: 561–563.

9. Malecová B, Morris KV. Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. Curr Opin Mol Ther. 2010;12: 214–222.

10. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10: 252–263.

11. Hwa V. STAT5B deficiency: Impacts on human growth and immunity. Growth Horm IGF Res. 2016;28: 16–20.

12. Andersson EI, Tanahashi T, Sekiguchi N, Gasparini VR, Bortoluzzi S, Kawakami T, et al. High incidence of activating STAT5B mutations in CD4-positive T-cell large granular lymphocyte leukemia. Blood. 2016;128: 2465–2468.

13. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. 2019;576: 149–157.

14. Scholefield J, Harrison PT. Prime editing - an update on the field. Gene Ther. 2021;28: 396–401.

15. Bardet AF, Steinmann J, Bafna S, Knoblich JA, Zeitlinger J, Stark A. Identification of transcription factor binding sites from ChIP-seq data at high resolution. Bioinformatics. 2013;29: 2705–2713.

16. Ignatieva EV, Levitsky VG, Kolchanov NA. Human Genes Encoding Transcription Factors and Chromatin-Modifying Proteins Have Low Levels of Promoter Polymorphism: A Study of 1000 Genomes Project Data. Int J Genomics Proteomics. 2015;2015: 260159.

17. Zhou Q, Liu M, Xia X, Gong T, Feng J, Liu W, et al. A mouse tissue transcription factor atlas. Nat Commun. 2017;8: 1–15.

18. Lee BH, Rhie SK. Molecular and computational approaches to map regulatory elements in 3D chromatin structure. Epigenetics Chromatin. 2021;14: 14.

19. Berest I, Arnold C, Reyes-Palomares A, Palla G, Rasmussen KD, Giles H, et al. Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. Cell Rep. 2019;29: 3147–3159.e12.

20. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019;47: W212–W224.

21. Roopra A. MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. PLoS Comput Biol. 2020;16: e1007800.

22. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. 2020;21: 36.

23. Beelman CA, Parker R. Degradation of mRNA in eukaryotes. Cell. 1995;81: 179–183.

24. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45: 1113–1120.

25. Li Z, Rana TM. Molecular mechanisms of RNA-triggered gene silencing machineries. Acc Chem Res. 2012;45: 1122–1131.

26. Ranganathan K, Sivasankar V. MicroRNAs - Biology and clinical applications. J Oral Maxillofac Pathol. 2014;18: 229–234.

27. List M, Dehghani Amirabad A, Kostka D, Schulz MH. Large-scale inference of competing endogenous RNA networks with sparse partial correlation. Bioinformatics. 2019;35: i596–i604.

28. Hoffmann M, Pachl E, Hartung M, Stiegler V. SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions. Narodonaselenie. 2021. Available: https://academic.oup.com/narcancer/article-abstract/3/1/zcaa042/6066568

29. Kartha RV, Subramanian S. Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. Front Genet. 2014;5: 8.

30. O’Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol . 2018;9: 402.

31. Dykes IM, Emanueli C. Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA. Genomics Proteomics Bioinformatics. 2017;15: 177–186.

32. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011;146: 353–358.

33. Zhang Y, Zhang Y, Feng Y, Zhang N, Chen S, Gu C, et al. Construction of circRNA-based ceRNA network and its prognosis-associated subnet of clear cell renal cell carcinoma. Cancer Med. 2021;10: 8210–8221.

34. Yao S, Jia X, Wang F, Sheng L, Song P, Cao Y, et al. CircRNA ARFGEF1 functions as a ceRNA to promote oncogenic KSHV-encoded viral interferon regulatory factor induction of cell invasion and angiogenesis by upregulating glutaredoxin 3. PLoS Pathog. 2021;17: e1009294.

35. Zhuang X, Lin Z, Xie F, Luo J, Chen T, Xi Q, et al. Identification of circRNA-associated ceRNA networks using longissimus thoracis of pigs of different breeds and growth stages. BMC Genomics. 2022;23: 294.

36. Miao Q, Ni B, Tang J. Coding potential of circRNAs: new discoveries and challenges. PeerJ. 2021;9: e10718.

37. Yu C-Y, Kuo H-C. The emerging roles and functions of circular RNAs and their generation. J Biomed Sci. 2019;26: 29.

38. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. Nat Biotechnol. 2014;32: 453–461.

39. Lasda E, Parker R. Circular RNAs: diversity of form and function. RNA. 2014;20: 1829–1842.

40. Enuka Y, Lauriola M, Feldman ME, Sas-Chen A, Ulitsky I, Yarden Y. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. Nucleic Acids Res. 2016;44: 1370–1383.

41. Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A. Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. Nucleic Acids Res. 2006;34: e63.

42. Mester-Tonczar J, Hašimbegović E, Spannbauer A, Traxler D, Kastner N, Zlabinger K, et al. Circular RNAs in Cardiac Regeneration: Cardiac Cell Proliferation, Differentiation, Survival, and Reprogramming. Front Physiol. 2020;11: 580465.

43. Szabo L, Salzman J. Detecting circular RNAs: bioinformatic and experimental challenges. Nat Rev Genet. 2016;17: 679–692.

44. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, et al. Circular

RNAs are abundant, conserved, and associated with ALU repeats. RNA. 2013;19: 141–157.

45. Zhang Z, Yang T, Xiao J. Circular RNAs: Promising Biomarkers for Human Diseases. EBioMedicine. 2018;34: 267–274.

46. Kristensen LS, Hansen TB, Venø MT, Kjems J. Circular RNAs in cancer: opportunities and challenges in the field. Oncogene. 2018;37: 555–565.

47. Qu S, Yang X, Li X, Wang J, Gao Y, Shang R, et al. Circular RNA: A new star of noncoding RNAs. Cancer Lett. 2015;365: 141–148.

48. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, et al. Natural RNA circles function as efficient microRNA sponges. Nature. 2013;495: 384–388.

49. Li M, Ding W, Tariq MA, Chang W, Zhang X, Xu W, et al. A circular transcript of ncx1 gene mediates ischemic myocardial injury by targeting miR-133a-3p. Theranostics. 2018;8: 5855–5869.

50. Yu KH-O, Shi CH, Wang B, Chow SH-C, Chung GT-Y, Lung RW-M, et al. Quantifying full-length circular RNAs in cancer. Genome Res. 2021. doi:10.1101/gr.275348.121

51. Hoffmann M, Trummer N, Schwartz L, Jankowski J, Lee HK, Willruth L-L, et al. TF-Prioritizer: a Java pipeline to prioritize condition-specific transcription factors. Gigascience. 2023;12: giad026.

52. Chen L, Wang C, Sun H, Wang J, Liang Y, Wang Y, et al. The bioinformatics toolbox for circRNA discovery and analysis. Brief Bioinform. 2021;22: 1706–1728.

53. Hoffmann M, Schwartz L, Ciora O-A, Trummer N, Willruth L-L, Jankowski J, et al. circRNA-sponging: a pipeline for extensive analysis of circRNA expression and their role in miRNA sponging. Bioinformatics Advances. 2023; vbad093.

54. Boniolo F, Hoffmann M, Roggendorf N, Tercan B, Baumbach J, Castro MAA, et al. spongEffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape. Bioinformatics. 2023. doi:10.1093/bioinformatics/btad276

55. The Editors of Encyclopedia Britannica. macromolecule. Encyclopedia Britannica. 2019. Available: https://www.britannica.com/science/macromolecule

56. Staudinger H, Fritschi J. Über Isopren und Kautschuk. 5. Mitteilung. Über die Hydrierung des Kautschuks und über seine Konstitution. Helv Chim Acta. 1922;5: 785–806.

57. Knippers R. Molekulare Genetik. Georg Thieme Verlag; 2006.

58. Brown TA. The Human Genome. Wiley-Liss; 2002.

59. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for

deoxyribose nucleic acid. Nature. 1953;171: 737–738.

60. Berg JM, Tymoczko JL. pdf Biochemistry. [cited 22 Dec 2022]. Available: https://www.stembook.org/sites/default/files/scf_members_attachment/pdf-biochemistry-jeremy-m-berg-john-l-tymoczko-lubert-stryer-pdf-download-free-book-3050c7d.pdf

61. transcription / DNA transcription. [cited 22 Dec 2022]. Available: https://www.nature.com/scitable/definition/transcription-dna-transcription-87/

62. Parry GS. The crystal structure of uracil. Acta Crystallogr. 1954;7: 313–320.

63. Pearl LH, Savva R. The problem with pyrimidines. Nat Struct Biol. 1996;3: 485–487.

64. Vértessy BG, Tóth J. Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. Acc Chem Res. 2009;42: 97–106.

65. Brenner S, Jacob F, Meselson M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature. 1961;190: 576–581.

66. Gros F, Hiatt H, Gilbert W, Kurland CG, Risebrough RW, Watson JD. Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. Nature. 1961;190: 581–585.

67. Sanger HL, Klotz G, Riesner D, Gross HJ, Kleinschmidt AK. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. Proc Natl Acad Sci U S A. 1976;73: 3852–3856.

68. Brown SD. XIST and the mapping of the X chromosome inactivation centre. Bioessays. 1991;13: 607–612.

69. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75: 843–854.

70. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391: 806–811.

71. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Principles of Membrane Transport. Garland Science; 2002.

72. Protein Structure. [cited 22 Dec 2022]. Available: https://www.nature.com/scitable/topicpage/protein-structure-14122136/

73. Sun PD, Foster CE, Boyington JC. Overview of protein structural and functional folds. Curr Protoc Protein Sci. 2004;Chapter 17: Unit 17.1.

74. Dehner C. Why Do Proteins Have Quaternary Structure: Non-allosteric Proteins. Molecular Life Sciences. 2014. pp. 1–7. doi:10.1007/978-1-4614-6436-5_21-1

75. Agarwal PK. Enzymes: An integrated view of structure, dynamics and function. Microb Cell Fact. 2006;5: 2.

76. Matson SW, Bean DW, George JW. DNA helicases: enzymes with essential roles in all aspects of DNA metabolism. Bioessays. 1994;16: 13–22.

77. Cooper GM. Transport of Small Molecules. Sinauer Associates; 2000.

78. Numata K. How to define and study structural proteins as biopolymer materials. Polym J. 2020;52: 1043–1056.

79. Morris R, Black KA, Stollar EJ. Uncovering protein function: from classification to complexes. Essays Biochem. 2022;66: 255–285.

80. Truman DES. The Cartoon Guide to Genetics. By Larry Gonick and Mark Wheelis. London: Harper and Row. £3.50 (paperback). ISBN 0 L6 460416 0. Genetical Research. 1984. pp. 104–104. doi:10.1017/s0016672300025799

81. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell. 1982;31: 147–157.

82. Stryer L. The Citric Acid Cycle: The Citric Acid Cycle Must Be Capable of Being Rapidly Replenished. U: Stryer L., Berg JM, Tymozcko JL (ur). Biochemistry.

83. Berg JM, Tymoczko JL, Gatto GJ, Stryer L. Stryer Biochemie. 2018. doi:10.1007/978-3-662-54620-8

84. Rodgers K, McVey M. Error-Prone Repair of DNA Double-Strand Breaks. J Cell Physiol. 2016;231: 15–24.

85. Vesely P. Molecular biology of the cell. By Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walter. ISBN 0-8153-3218-1; hardback; 1,616 pages; $110.00 Garland Science Inc., New York, 2002. Scanning. 2006. pp. 153–153. doi:10.1002/sca.4950260309

86. Uzman A. Molecular Cell Biology (4th edition) Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore and James Darnell; Freeman & Co., New York, NY, 2000, 1084 pp., list price $102.25, ISBN 0-7167-3136-3. Biochemistry and Molecular Biology Education. 2001. pp. 126–128. doi:10.1016/s1470-8175(01)00023-6

87. Uzman A. Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Biochemistry and Molecular Biology Education. 2003. pp. 212–214. doi:10.1002/bmb.2003.494031049999

88. Al Aboud NM, Tupper C, Jialal I. Genetics, Epigenetic Mechanism. StatPearls Publishing; 2022.

89. Cutter AR, Hayes JJ. A brief review of nucleosome structure. FEBS Lett. 2015;589: 2914–2922.

90. Lee H-T, Oh S, Ro DH, Yoo H, Kwon Y-W. The Key Role of DNA Methylation and Histone Acetylation in Epigenetics of Atherosclerosis. J Lipid Atheroscler. 2020;9: 419–434.

91. Miller JL, Grant PA. The role of DNA methylation and histone modifications in transcriptional regulation in humans. Subcell Biochem. 2013;61: 289–317.

92. Tissue-specific regulation by transcription factors. Transcription Factors. Garland Science; 2003. pp. 247–272.

93. Phillips T. Regulation of transcription and gene expression in eukaryotes. Nature Education.

94. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013;155: 934–947.

95. Saint-André V, Alexander J. Federation, Lin CY, Abraham BJ, Reddy J, Lee TI, et al. Models of human core transcriptional regulatory circuitries. Genome Research. 2016. pp. 385–396. doi:10.1101/gr.197590.115

96. Pott S, Lieb JD. What are super-enhancers? Nat Genet. 2015;47: 8–12.

97. Shin HY, Willi M, HyunYoo K, Zeng X, Wang C, Metser G, et al. Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat Genet. 2016;48: 904–911.

98. Lee HK, Willi M, Kuhns T, Liu C, Hennighausen L. Redundant and non-redundant cytokine-activated enhancers control Csn1s2b expression in the lactating mouse mammary gland. Nat Commun. 2021;12: 2239.

99. Liu H, Lei C, He Q, Pan Z, Xiao D, Tao Y. Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. Mol Cancer. 2018;17: 64.

100. Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. 2005;33: 2374–2383.

101. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. RNA. 2003;9: 277–279.

102. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004;23: 4051–4060.

103. Speiser JJ, Erşahin C, Osipo C. The functional role of Notch signaling in triple-negative breast cancer. Vitam Horm. 2013;93: 277–306.

104. Lam JKW, Chow MYT, Zhang Y, Leung SWS. siRNA Versus miRNA as Therapeutics for Gene Silencing. Mol Ther Nucleic Acids. 2015;4: e252.

105. Rozov A, Demeshkina N, Khusainov I, Westhof E, Yusupov M, Yusupova G.

Novel base-pairing interactions at the tRNA wobble position crucial for accurate reading of the genetic code. Nat Commun. 2016;7: 10457.

106. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014;15: 7–21.

107. Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol. 2005;6: 376–385.

108. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116: 281–297.

109. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP. Prediction of plant microRNA targets. Cell. 2002;110: 513–520.

110. Xiao J. Circular RNAs: Biogenesis and Functions. Springer; 2018.

111. Nielsen AF, Bindereif A, Bozzoni I, Hanan M, Hansen TB, Irimia M, et al. Best practice standards for circular RNA research. Nat Methods. 2022;19: 1208–1220.

112. Ma Y, Zheng L, Gao Y, Zhang W, Zhang Q, Xu Y. A Comprehensive Overview of circRNAs: Emerging Biomarkers and Potential Therapeutics in Gynecological Cancers. Front Cell Dev Biol. 2021;9: 709512.

113. Bose R, Ain R. Regulation of Transcription by Circular RNAs. Adv Exp Med Biol. 2018;1087: 81–94.

114. Zhou W-Y, Cai Z-R, Liu J, Wang D-S, Ju H-Q, Xu R-H. Circular RNA: metabolism, functions and interactions with proteins. Mol Cancer. 2020;19: 172.

115. Panda AC. Circular RNAs Act as miRNA Sponges. Adv Exp Med Biol. 2018;1087: 67–79.

116. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, et al. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014;56: 55–66.

117. Das A, Sinha T, Shyamal S, Panda AC. Emerging Role of Circular RNA–Protein Interactions. Non-Coding RNA. 2021;7: 48.

118. Patop IL, Wüst S, Kadener S. Past, present, and future of circRNAs. EMBO J. 2019;38: e100836.

119. Okholm TLH, Nielsen MM, Hamilton MP, Christensen L-L, Vang S, Hedegaard J, et al. Circular RNA expression is abundant and correlated to aggressiveness in early-stage bladder cancer. NPJ Genom Med. 2017;2: 36.

120. Ebert MS, Sharp PA. MicroRNA sponges: progress and possibilities. RNA. 2010;16: 2043–2050.

121. Subedi P, Moertl S, Azimzadeh O. Omics in Radiation Biology: Surprised but

Not Disappointed. Radiation. 2022;2: 124–129.

122. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

123. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13: 278–289.

124. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

125. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;2010: db.prot5384.

126. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4: 651–657.

127. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320: 1344–1349.

128. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA, et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science. 2002;296: 916–919.

129. McLafferty FW. A century of progress in molecular mass spectrometry. Annu Rev Anal Chem . 2011;4: 1–22.

130. National Academies of Sciences Engineering, Medicine, Others. Current Methods for Studying the Human Microbiome. Environmental Chemicals, the Human Microbiome, and Health Risk: A Research Strategy. National Academies Press (US); 2017.

131. Ning K. Methodologies of Multi-Omics Data Integration and Data Mining: Techniques and Applications. Springer Nature; 2023.

132. Krassowski M, Das V, Sahu SK, Misra BB. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front Genet. 2020;11: 610798.

133. Pammi M, Aghaeepour N, Neu J. Multiomics, artificial intelligence, and precision medicine in perinatology. Pediatr Res. 2023;93: 308–315.

134. Danchin A. In vivo, in vitro and in silico: an open space for the development of microbe-based applications of synthetic biology. Microb Biotechnol. 2022;15: 42–64.

135. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018;122: e59.

136. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr. 2020;9: 163–173.

137. Li N, Jin K, Bai Y, Fu H, Liu L, Liu B. Tn5 Transposase Applied in Genomics Research. Int J Mol Sci. 2020;21. doi:10.3390/ijms21218329

138. Gunter HM, Youlten SE, Madala BS, Reis ALM, Stevanovski I, Wong T, et al. Library adaptors with integrated reference controls improve the accuracy and reliability of nanopore sequencing. Nat Commun. 2022;13: 6437.

139. Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, Lind K, et al. The real-time polymerase chain reaction. Mol Aspects Med. 2006;27: 95–125.

140. Oda Y, Sadakane K, Yoshikawa Y, Imanaka T, Takiguchi K, Hayashi M, et al. Highly concentrated ethanol solutions: Good solvents for DNA as revealed by single-molecule observation. Chemphyschem. 2016;17: 471–473.

141. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10: 669–680.

142. Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nat Rev Genet. 2012;13: 840–852.

143. Nakato R, Sakata T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods. 2021;187: 44–53.

144. Sambrook J, Russell DW. Fragmentation of DNA by sonication. CSH Protoc. 2006;2006. doi:10.1101/pdb.prot4538

145. Hoffman EA, Frey BL, Smith LM, Auble DT. Formaldehyde crosslinking: a tool for the study of chromatin complexes. J Biol Chem. 2015;290: 26404–26411.

146. Chiesa R, Moscatelli M, Giordano C, Siccardi F, Cigada A. Influence of Heat Treatment on Structural, Mechanical and Wear Properties of Cross-Linked UHMWPE. J Appl Biomater Biomech. 2004;2: 20–28.

147. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38: 276–278.

148. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35: 316–319.

149. Babraham bioinformatics - FastQC A quality control tool for high throughput sequence data. [cited 11 Jan 2023]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

150. Andrews S, Others. FastQC: a quality control tool for high throughput

sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

151. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048.

152. Babraham Bioinformatics - Trim Galore! [cited 11 Jan 2023]. Available: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

153. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

154. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.

155. Rajawat J. Transcriptomics. Omics Approaches, Technologies And Applications. 2018. pp. 39–56. doi:10.1007/978-981-13-2925-8_3

156. Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature. 1970;226: 1211–1213.

157. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013. pp. 15–21. doi:10.1093/bioinformatics/bts635

158. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

159. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8: 118–127.

160. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2: 160.

161. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2022; 1–28.

162. Goecks J, Jalili V, Heiser LM, Gray JW. How Machine Learning Will Transform Biomedicine. Cell. 2020;181: 92–101.

163. Machine Learning: Algorithms, Models and Applications. BoD – Books on Demand; 2021.

164. Ansari S, Nassif AB. A Comprehensive Study of Regression Analysis and the Existing Techniques. 2022 Advances in Science and Engineering Technology International Conferences (ASET). 2022. pp. 1–10.

165. Lewis TG. Network Science: Theory and Applications. John Wiley & Sons; 2011.

166. Wiredja D, Bebek G. Identifying Gene Interaction Networks. Methods Mol Biol. 2017;1666: 539–556.

167. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1: 2.

168. Comte B, Baumbach J, Benis A, Basílio J, Debeljak N, Flobak Å, et al. Network and Systems Medicine: Position Paper of the European Collaboration on Science and Technology Action on Open Multiscale Systems Medicine. Netw Syst Med. 2020;3: 67–90.

169. Vijayabaskar MS, Goode DK, Obier N, Lichtinger M, Emmett AML, Abidin FNZ, et al. Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets. PLoS Comput Biol. 2019;15: e1007337.

170. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell. 2022;34: 718–741.

171. Staal FJ, Weerkamp F, Langerak AW, Hendriks RW, Clevers HC. Transcriptional control of t lymphocyte differentiation. Stem Cells. 2001;19: 165–179.

172. Huilgol D, Venkataramani P, Nandi S, Bhattacharjee S. Transcription Factors That Govern Development and Disease: An Achilles Heel in Cancer. Genes . 2019;10. doi:10.3390/genes10100794

173. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447: 799–816.

174. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21: 447–455.

175. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016;48: 838–847.

176. Dujon B. The yeast genome project: what did we learn? Trends Genet. 1996;12: 263–270.

177. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet. 2001;27: 167–171.

178. Keleş S, van der Laan M, Eisen MB. Identification of regulatory elements using a feature selection method. Bioinformatics. 2002;18: 1167–1175.

179. Wang W, Cherry JM, Botstein D, Li H. A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae. Proc Natl Acad Sci U S A. 2002;99: 16893–16898.

180. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci U S A. 2003;100: 3339–3344.

181. Liao JC, Boscolo R, Yang Y-L, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci U S A. 2003;100: 15522–15527.

182. Chang C, Ding Z, Hung YS, Fung PCW. Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics. 2008;24: 1349–1358.

183. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. Metab Eng. 2005;7: 128–141.

184. Noor A, Ahmad A, Serpedin E, Nounou M, Nounou H. ROBNCA: robust network component analysis for recovering transcription factor activities. Bioinformatics. 2013;29: 2410–2418.

185. Boscolo R, Sabatti C, Liao JC, Roychowdhury VP. A generalized framework for network component analysis. IEEE/ACM Trans Comput Biol Bioinform. 2005;2: 289–301.

186. Noor A, Ahmad A, Serpedin E. SparseNCA: Sparse Network Component Analysis for Recovering Transcription Factor Activities with Incomplete Prior Information. IEEE/ACM Trans Comput Biol Bioinform. 2018;15: 387–395.

187. Shi Q, Zhang C, Guo W, Zeng T, Lu L, Jiang Z, et al. Local network component analysis for quantifying transcription factor activities. Methods. 2017;124: 25–35.

188. Ma CZ, Brent MR. Inferring TF activities and activity regulators from gene expression data with constraints from TF perturbation data. Bioinformatics. 2021;37: 1234–1245.

189. Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS Comput Biol. 2014;10: e1003908.

190. Jiang P, Freedman ML, Liu JS, Liu XS. Inference of transcriptional regulation in cancers. Proc Natl Acad Sci U S A. 2015;112: 7731–7736.

191. Karabacak Calviello A, Hirsekorn A, Wurmus R, Yusuf D, Ohler U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. Genome Biol. 2019;20: 42.

192. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun. 2020;11: 4267.

193. Baek S, Goldstein I, Hager GL. Bivariate Genomic Footprinting Detects

Changes in Transcription Factor Activity. Cell Rep. 2017;19: 1710–1722.

194. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019;20: 45.

195. Gusmao EG, Dieterich C, Zenke M, Costa IG. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics. 2014;30: 3143–3151.

196. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods. 2016;13: 303–309.

197. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep. 2018;19: e46255.

198. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

199. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019;9: 9354.

200. Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Res. 2017;45: 54–66.

201. Schmidt F, Kern F, Ebert P, Baumgarten N, Schulz MH. TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. Bioinformatics. 2018;35: 1608–1609.

202. Pundhir S, Bagger FO, Lauridsen FB, Rapin N, Porse BT. Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. Nucleic Acids Res. 2016;44: 4037–4051.

203. Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics. 2007;23: 134–141.

204. Karadimitriou, Marshall. Mann-Whitney U test. Sheffield: Sheffield Hallam. Available: https://maths.shu.ac.uk/mathshelp/Stats%20support%20resources/Resources/Nonparametric/Comparing%20groups/Mann-Whitney/SPSS/stcp-marshall-MannWhitS.pdf

205. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics. 1947. pp. 50–60. doi:10.1214/aoms/1177730491

206. Järvelin K, Kekäläinen J. Discounted Cumulated Gain. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer US; 2009. pp. 849–853.

207. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011. pp. 24–26. doi:10.1038/nbt.1754

208. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14: 178–192.

209. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. Cancer Res. 2017;77: e31–e34.

210. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43: W589–98.

211. Maza E, Frasse P, Senin P, Bouzayen M, Zouine M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. Commun Integr Biol. 2013;6: e25849.

212. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;371: 215–220.

213. Achrem M, Szućko I, Kalinka A. The epigenetic regulation of centromeres and telomeres in plants and animals. Comp Cytogenet. 2020;14: 265–311.

214. Description.pdf at master · SchulzLab/TEPIC. Github; Available: https://github.com/SchulzLab/TEPIC

215. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci U S A. 2009;106: 21521–21526.

216. Mann PS. Introductory Statistics. John Wiley & Sons, Incorporated; 1995.

217. Ross SM. Introductory Statistics. Academic Press; 2017.

218. Hsu MT, Coca-Prados M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature. 1979;280: 339–340.

219. Kos A, Dijkema R, Arnberg AC, van der Meide PH, Schellekens H. The hepatitis delta (delta) virus possesses a circular RNA. Nature. 1986;323: 558–560.

220. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, et al. Scrambled exons. Cell. 1991;64: 607–613.

221. Chen CY, Sarnow P. Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. Science. 1995;268: 415–417.

222. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell

types. PLoS One. 2012;7: e30733.

223. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495: 333–338.

224. Li Y, Zheng Q, Bao C, Li S, Guo W, Zhao J, et al. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. Cell Res. 2015;25: 981–984.

225. Li X, Yang L, Chen L-L. The Biogenesis, Functions, and Challenges of Circular RNAs. Mol Cell. 2018;71: 428–442.

226. Zhang J, Hossain MT, Liu W, Peng Y, Pan Y, Wei Y. Evaluation of CircRNA Sequence Assembly Methods Using Long Reads. Front Genet. 2022;13: 816825.

227. Pandey PR, Rout PK, Das A, Gorospe M, Panda AC. RPAD (RNase R treatment, polyadenylation, and poly(A)+ RNA depletion) method to isolate highly pure circular RNA. Methods. 2019;155: 41–48.

228. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol. 2015;16: 4.

229. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Brief Bioinform. 2018;19: 803–810.

230. Zheng Y, Ji P, Chen S, Hou L, Zhao F. Reconstruction of full-length circular RNAs enables isoform-level quantification. Genome Med. 2019;11: 2.

231. Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L. Complementary sequence-mediated exon circularization. Cell. 2014;159: 134–147.

232. Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. 2016;26: 1277–1287.

233. Pan X, Xiong K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. Mol Biosyst. 2015;11: 2219–2226.

234. Pan X, Xiong K, Anthon C, Hyttel P, Freude KK, Jensen LJ, et al. WebCircRNA: Classifying the Circular RNA Potential of Coding and Noncoding RNA. Genes . 2018;9. doi:10.3390/genes9110536

235. Liu Z, Han J, Lv H, Liu J, Liu R. Computational identification of circular RNAs based on conformational and thermodynamic properties in the flanking introns. Comput Biol Chem. 2016;61: 221–225.

236. Wang J, Wang L. Deep learning of the back-splicing code for circular RNA formation. Bioinformatics. 2019;35: 5235–5242.

237. Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, Kjems J. The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet. 2019;20: 675–691.

238. Hansen TB. Improved circRNA Identification by Combining Prediction Algorithms. Front Cell Dev Biol. 2018;6: 20.

239. Hansen TB, Venø MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. Nucleic Acids Res. 2016;44: e58.

240. Gaffo E, Bonizzato A, Kronnie GT, Bortoluzzi S. CirComPara: A Multi-Method Comparative Bioinformatics Pipeline to Detect and Study circRNAs from RNA-seq Data. Noncoding RNA. 2017;3. doi:10.3390/ncrna3010008

241. Li L, Zheng Y-C, Kayani MUR, Xu W, Wang G-Q, Sun P, et al. Comprehensive analysis of circRNA expression profiles in humans by RAISE. Int J Oncol. 2017;51: 1625–1638.

242. Chen L, Yu Y, Zhang X, Liu C, Ye C, Fan L. PcircRNA_finder: a software for circRNA prediction in plants. Bioinformatics. 2016;32: 3528–3529.

243. Li L, Bu D, Zhao Y. CircRNAwrap - a flexible pipeline for circRNA identification, transcript prediction, and abundance estimation. FEBS Lett. 2019;593: 1179–1189.

244. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, et al. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Mol Cell. 2015;58: 870–885.

245. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20: 1666–1670.

246. Meng X, Hu D, Zhang P, Chen Q, Chen M. CircFunBase: a database for functional circular RNAs. Database . 2019;2019. doi:10.1093/database/baz003

247. Dong R, Ma X-K, Li G-W, Yang L. CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison. Genomics Proteomics Bioinformatics. 2018;16: 226–233.

248. Chen X, Han P, Zhou T, Guo X, Song X, Li Y. circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. Sci Rep. 2016;6: 34985.

249. Liu M, Wang Q, Shen J, Yang BB, Ding X. Circbank: a comprehensive database for circRNA with standard nomenclature. RNA Biol. 2019;16: 899–905.

250. Liang G, Yang Y, Niu G, Tang Z, Li K. Genome-wide profiling of Sus scrofa circular RNAs across nine organs and three developmental stages. DNA Res. 2017;24: 523–535.

251. Ye J, Wang L, Li S, Zhang Q, Zhang Q, Tang W, et al. AtCircDB: a tissue-specific database for Arabidopsis circular RNAs. Brief Bioinform. 2019;20:

58–65.

252. Chu Q, Zhang X, Zhu X, Liu C, Mao L, Ye C, et al. PlantcircBase: A Database for Plant Circular RNAs. Mol Plant. 2017;10: 1126–1128.

253. Zhang P, Meng X, Chen H, Liu Y, Xue J, Zhou Y, et al. PlantCircNet: a database for plant circRNA-miRNA-mRNA regulatory networks. Database . 2017;2017. doi:10.1093/database/bax089

254. Wang K, Wang C, Guo B, Song K, Shi C, Jiang X, et al. CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress. Database . 2019;2019. doi:10.1093/database/baz053

255. Bray NL, Pimentel H, Melsted P, Pachter L. Erratum: Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34: 888.

256. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5: R1.

257. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39: 1278–1284.

258. Ding J, Li X, Hu H. TarPmiR: a new approach for microRNA target site prediction. Bioinformatics. 2016;32: 2768–2775.

259. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19: 40.

260. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40: 37–52.

261. Lai EC. Predicting and validating microRNA targets. Genome Biol. 2004;5: 115.

262. Karagkouni D, Paraskevopoulou MD, Tastsoglou S, Skoufos G, Karavangeli A, Pierros V, et al. DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. Nucleic Acids Res. 2020;48: D101–D110.

263. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, et al. miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic Acids Research. 2011. pp. D163–D169. doi:10.1093/nar/gkq1107

264. Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res. 2020;48: D148–D154.

265. Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: An online resource for prediction of microRNA binding sites. PLoS One. 2018;13: e0206239.

266. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). Curr Protoc Hum Genet. 2015;87: 11.16.1–11.16.14.

267. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25.

268. Laub V, Devraj K, Elias L, Schulte D. Bioinformatics for wet-lab scientists: practical application in sequencing analysis. BMC Genomics. 2023;24: 382.

269. Karimzadeh M, Hoffman MM. Top considerations for creating bioinformatics software documentation. Brief Bioinform. 2018;19: 693–699.

270. Mangul S, Martin LS, Eskin E, Blekhman R. Improving the usability and archival stability of bioinformatics software. Genome Biol. 2019;20: 47.

271. Gardner PP, Paterson JM, McGimpsey S, Ashari-Ghomi F, Umu SU, Pawlik A, et al. Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software. Genome Biol. 2022;23: 56.

272. Mangul S, Mosqueiro T, Abdill RJ, Duong D, Mitchell K, Sarwal V, et al. Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLoS Biol. 2019;17: e3000333.

273. Brandies PA, Hogg CJ. Ten simple rules for getting started with command-line bioinformatics. PLoS Comput Biol. 2021;17: e1008645.

274. Seemann T. Ten recommendations for creating usable bioinformatics command line software. Gigascience. 2013;2: 15.

275. Reiter T, Brooks PT, Irber L, Joslin SEK, Reid CM, Scott C, et al. Streamlining data-intensive biology with workflow systems. Gigascience. 2021;10. doi:10.1093/gigascience/giaa140

276. Abeysooriya M, Soria M, Kasu MS, Ziemann M. Gene name errors: Lessons not learned. PLoS Comput Biol. 2021;17: e1008984.

277. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics. 2004;5: 80.

278. Hoffmann N, Mayer G, Has C, Kopczynski D, Al Machot F, Schwudke D, et al. A Current Encyclopedia of Bioinformatics Tools, Data Formats and Resources for Mass Spectrometry Lipidomics. Metabolites. 2022;12. doi:10.3390/metabo12070584

279. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol. 2011;12: 918–922.

280. Mayran A, Drouin J. Pioneer transcription factors shape the epigenetic landscape. J Biol Chem. 2018;293: 13795–13804.

281. Ehsani R, Bahrami S, Drabløs F. Feature-based classification of human transcription factors into hypothetical sub-classes related to regulatory function. BMC Bioinformatics. 2016;17: 459.

282. Srivastava D, Mahony S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. Biochim Biophys Acta Gene Regul Mech. 2020;1863: 194443.

283. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014;32: 171–178.

284. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci. 2014;39: 381–399.

285. Inukai S, Kock KH, Bulyk ML. Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev. 2017;43: 110–119.

286. Nakatake Y, Ko SBH, Sharov AA, Wakabayashi S, Murakami M, Sakota M, et al. Generation and Profiling of 2,135 Human ESC Lines for the Systematic Analyses of Cell States Perturbed by Inducing Single Transcription Factors. Cell Rep. 2020;31: 107655.

287. Kreimer A, Ashuach T, Inoue F, Khodaverdian A, Deng C, Yosef N, et al. Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. Nat Commun. 2022;13: 1504.

288. Larsen SJ, Röttger R, Schmidt HHHW, Baumbach J. E. coli gene regulatory networks are inconsistent with gene expression data. Nucleic Acids Res. 2019;47: 85–92.

289. Gillespie MA, Palii CG, Sanchez-Taltavull D, Shannon P, Longabaugh WJR, Downes DJ, et al. Absolute Quantification of Transcription Factors Reveals Principles of Gene Regulation in Erythropoiesis. Mol Cell. 2020;78: 960–974.e11.

290. Weidemüller P, Kholmatov M, Petsalaki E, Zaugg JB. Transcription factors: Bridge between cell signaling and gene regulation. Proteomics. 2021;21: e2000034.

291. Joung J, Ma S, Tay T, Geiger-Schuller KR, Kirchgatterer PC, Verdine VK, et al. A transcription factor atlas of directed differentiation. Cell. 2023;186: 209–229.e26.

292. Harris HL, Gu H, Olshansky M, Wang A, Farabella I, Eliaz Y, et al. Chromatin alternates between A and B compartments at kilobase scale for subgenic organization. Nat Commun. 2023;14: 3303.

293. Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. Nat Genet. 2020;52: 8–16.

294. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. Annu Rev Cell Dev Biol. 2019;35: 357–379.

295. Yardımcı GG, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol. 2019;20: 57.

296. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72: 65–75.

297. Trescher S, Münchmeyer J, Leser U. Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. BMC Syst Biol. 2017;11: 41.

298. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics. 2018;19: 232.

299. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17: 147–154.

300. Fröhlich H. biRte: Bayesian inference of context-specific regulator activities and transcriptional networks. Bioinformatics. 2015;31: 3290–3298.

301. Balwierz PJ, Pachkov M, Arnold P, Gruber AJ, Zavolan M, van Nimwegen E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res. 2014;24: 869–884.

302. Degtyareva AO, Antontseva EV, Merkulova TI. Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. Int J Mol Sci. 2021;22. doi:10.3390/ijms22126454

303. Vohra M, Sharma AR, Prabhu B N, Rai PS. SNPs in Sites for DNA Methylation, Transcription Factor Binding, and miRNA Targets Leading to Allele-Specific Gene Expression and Contributing to Complex Disease Risk: A Systematic Review. Public Health Genomics. 2020;23: 155–170.

304. Hernández-Lorenzo L, Hoffmann M, Scheibling E, List M, Matías-Guiu JA, Ayala JL. On the limits of graph neural networks for the early diagnosis of Alzheimer's disease. Sci Rep. 2022;12: 17632.

305. Heap GA, Trynka G, Jansen RC, Bruinenberg M, Swertz MA, Dinesen LC, et al. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med Genomics. 2009;2: 1.

306. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8: e1002822.

307. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS

Catalog). Nucleic Acids Res. 2017;45: D896–D901.

308. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

309. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74.

310. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29: 308–311.

311. Gao Y, Zhao F. Computational Strategies for Exploring Circular RNAs. Trends Genet. 2018;34: 389–400.

312. Zeng X, Lin W, Guo M, Zou Q. A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput Biol. 2017;13: e1005420.

313. Jakobi T, Dieterich C. Computational approaches for circular RNA analysis. Wiley Interdiscip Rev RNA. 2019;10: e1528.

314. Seal RL, Chen L-L, Griffiths-Jones S, Lowe TM, Mathews MB, O’Reilly D, et al. A guide to naming human non-coding RNA genes. EMBO J. 2020;39: e103777.

315. Piwecka M, Glažar P, Hernandez-Miranda LR, Memczak S, Wolf SA, Rybak-Wolf A, et al. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science. 2017;357. doi:10.1126/science.aam8526

316. He AT, Liu J, Li F, Yang BB. Targeting circular RNAs as a therapeutic approach: current strategies and challenges. Signal Transduct Target Ther. 2021;6: 185.

317. Vo JN, Cieslik M, Zhang Y, Shukla S, Xiao L, Zhang Y, et al. The Landscape of Circular RNA in Cancer. Cell. 2019;176: 869–881.e13.

318. Li F, Yang Q, He AT, Yang BB. Circular RNAs in cancer: Limitations in functional studies and diagnostic potential. Semin Cancer Biol. 2021;75: 49–61.

319. Chen S, Huang V, Xu X, Livingstone J, Soares F, Jeon J, et al. Widespread and Functional RNA Circularization in Localized Prostate Cancer. Cell. 2019;176: 831–843.e22.

320. Chen Y, Wang D, Peng H, Chen X, Han X, Yu J, et al. Epigenetically upregulated oncoprotein PLCE1 drives esophageal carcinoma angiogenesis and proliferation via activating the PI-PLCε-NF-κB signaling pathway and VEGF-C/ Bcl-2 expression. Mol Cancer. 2019;18: 1.

321. Schultz N, Marenstein DR, De Angelis DA, Wang W-Q, Nelander S, Jacobsen

A, et al. Off-target effects dominate a large-scale RNAi screen for modulators of the TGF-β pathway and reveal microRNA regulation of TGFBR2. Silence. 2011;2: 3.

322. Jackson AL, Burchard J, Schelter J, Chau BN, Cleary M, Lim L, et al. Widespread siRNA “off-target” transcript silencing mediated by seed region sequence complementarity. RNA. 2006;12: 1179–1187.

323. Anderson EM, Birmingham A, Baskerville S, Reynolds A, Maksimova E, Leake D, et al. Experimental validation of the importance of seed complement frequency to siRNA specificity. RNA. 2008;14: 853–861.

324. Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017;27: 1759–1768.

325. Tang C, Yu T, Xie Y, Wang Z, McSwiggin H, Zhang Y, et al. Template switching causes artificial junction formation and false identification of circular RNAs. bioRxiv. 2018. p. 259556. doi:10.1101/259556

326. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008;26: 317–325.

327. Kristensen LS. Profiling of circRNAs using an enzyme-free digital counting method. Methods. 2021;196: 11–16.

328. Pamudurti NR, Patop IL, Krishnamoorthy A, Ashwal-Fluss R, Bartok O, Kadener S. An in vivo strategy for knockdown of circular RNAs. Cell Discov. 2020;6: 52.

329. Konermann S, Lotfy P, Brideau NJ, Oki J, Shokhirev MN, Hsu PD. Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. Cell. 2018;173: 665–676.e14.

330. Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, et al. RNA targeting with CRISPR-Cas13. Nature. 2017;550: 280–284.

331. Zhang Y, Xue W, Li X, Zhang J, Chen S, Zhang J-L, et al. The Biogenesis of Nascent Circular RNAs. Cell Rep. 2016;15: 611–624.

332. Zheng Q, Bao C, Guo W, Li S, Chen J, Chen B, et al. Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. Nat Commun. 2016;7: 11215.

333. Kilikevicius A, Meister G, Corey DR. Reexamining assumptions about miRNA-guided gene silencing. Nucleic Acids Res. 2022;50: 617–634.

334. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136: 215–233.

335. Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. Target mRNA abundance dilutes microRNA and siRNA activity. Mol Syst Biol. 2010;6: 363.

336. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol. 2011;18: 1139–1146.

337. Zhao X, Chen D, Cai Y, Zhang F, Xu J. RBPvsMIR: A Computational Pipeline to Identify Competing miRNAs and RNA-Binding Protein Pairs Regulating the Shared Transcripts. Genes . 2018;9. doi:10.3390/genes9090426

338. Ustianenko D, Chiu H-S, Treiber T, Weyn-Vanhentenryck SM, Treiber N, Meister G, et al. LIN28 Selectively Modulates a Subclass of Let-7 MicroRNAs. Mol Cell. 2018;71: 271–283.e5.

339. Posner R, Laubenbacher R. Connecting the molecular function of microRNAs to cell differentiation dynamics. J R Soc Interface. 2019;16: 20190437.

340. Boettger T, Braun T. A new level of complexity: the role of microRNAs in cardiovascular development. Circ Res. 2012;110: 1000–1013.

341. Aartsma-Rus A, Corey DR. The 10th Oligonucleotide Therapy Approved: Golodirsen for Duchenne Muscular Dystrophy. Nucleic Acid Ther. 2020;30: 67–70.

342. Mendell JR, Khan N, Sha N, Eliopoulos H, McDonald CM, Goemans N, et al. Comparison of Long-term Ambulatory Function in Patients with Duchenne Muscular Dystrophy Treated with Eteplirsen and Matched Natural History Controls. J Neuromuscul Dis. 2021;8: 469–479.

343. Bennett CF, Kordasiewicz HB, Cleveland DW. Antisense Drugs Make Sense for Neurological Diseases. Annu Rev Pharmacol Toxicol. 2021;61: 831–852.

344. Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, et al. Silencing of microRNAs in vivo with “antagomirs.” Nature. 2005;438: 685–689.

345. Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, et al. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. Cell Metab. 2006;3: 87–98.

346. Hogan DJ, Vincent TM, Fish S, Marcusson EG, Bhat B, Chau BN, et al. Anti-miRs competitively inhibit microRNAs in Argonaute complexes. PLoS One. 2014;9: e100951.

347. Shen X, Corey DR. Chemistry, mechanism and clinical status of antisense oligonucleotides and duplex RNAs. Nucleic Acids Res. 2018;46: 1584–1600.

348. Crooke ST, Baker BF, Crooke RM, Liang X-H. Antisense technology: an overview and prospectus. Nat Rev Drug Discov. 2021;20: 427–453.

349. Aoki Y, Wood MJA. Emerging Oligonucleotide Therapeutics for Rare Neuromuscular Diseases. J Neuromuscul Dis. 2021;8: 869–884.

350. Li Z, Rana TM. Therapeutic targeting of microRNAs: current status and future challenges. Nat Rev Drug Discov. 2014;13: 622–638.

351. Gagnon KT, Corey DR. Guidelines for Experiments Using Antisense Oligonucleotides and Double-Stranded RNAs. Nucleic Acid Ther. 2019;29: 116–122.

352. Zhu P, Zhu X, Wu J, He L, Lu T, Wang Y, et al. IL-13 secreted by ILC2s promotes the self-renewal of intestinal stem cells through circular RNA circPan3. Nat Immunol. 2019;20: 183–194.