

DISS. ETH NO.

Computational analysis of multi-omics data to understand the molecular mechanisms of germline-dependent epigenetic inheritance

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
Deepak Kumar Tanwar
M.Sc., McGill University, Canada

born on 30.10.1991

citizen of India

accepted on the recommendation of

Prof. Isabelle Mansuy, examiner
Prof. Tuncay Baubec, co-examiner
Prof. Mark Robinson, co-examiner

2021

Acknowledgements

I want to thank a few people.

Table of Contents

List of Abbreviations	1
Summary	3
Zusammenfassung	7
Thesis overview	11
Introduction	13
Epigenetics	13
Germline epigenetics	14
Spermatogonial Stem Cells	14
Reprogramming	15
Epigenetic Inheritance	16
Transgenerational Epigenetic Inheritance	17
Reprogramming in epigenetic inheritance	18
Models of Epigenetic Inheritance	18
MSUS mouse model	19
Vectors of epigenetic inheritance	19
DNA methylation	20
Histone modifications	20
RNA	22
Non-coding RNA	22
0.0.1 Transcription factors	23
Extracellular vesicles	23
Challenges in data analysis	24
Multi-omics data integration	24
0.0.1.1 Ad-hoc approaches	25
0.0.1.2 Systematic approaches	25

Short RNA data analysis	26
Goals of thesis	27
Bioinformatics methods	33
Version controlled data analysis using git	33
Data analysis directory organization	34
Pipelines for data analysis	35
Quality check and trimming of sequencing datasets	35
RNA-seq	37
Pseudo-alignment and alignment	37
Duplicated reads analysis	39
Differential expression analysis	39
Exploratory data analysis	40
Differential transcript usage and differential isoform switching	40
Differential exon usage and differential 3' untranslated regions usage .	40
Functional analysis	40
ATAC-seq	41
Alignment, post-alignment quality control, and peak calling	41
Peaks annotation and differential accessibility analysis	42
Functional analysis	43
Differential accessibility analysis at transposable elements	43
ChIP-seq	43
Alignment and peak calling	44
Peaks annotation, differential analysis, and functional analysis	44
Bisulfite sequencing	44
Alignment and methylation calling	45
Differential methylation analysis	45
CNV analysis	45
DML and DMR annotation, and functional analysis	46
Short RNA-seq	46
Multi-omics analysis	46
Chapter 1: Dynamic chromatin accessibility in spermatogonial cells for transcriptional programmings from early postnatal to adult stages	53
1.1 Update on the new data analysis	54
1.1.1 Inclusion of RNA-seq data from the literature	54
1.1.2 New RNA-seq data from lab	55

1.1.2.1	Identification of clusters of genes with distinct expression profiles	57
1.1.2.2	Multi-omics data integration	58
Chapter 2: shortRNA: a flexible framework for the analysis of short RNA sequencing data		63
2.1	Abstract	63
2.2	Introduction and background	64
2.3	Methods	66
2.3.1	Development and testing environment	66
2.3.2	Backbone data structures of shortRNA	66
2.3.2.1	phylo	66
2.3.2.2	FactorList	66
2.3.2.3	DataFrame	67
2.3.2.4	TreeSummarizedExperiment	67
2.3.3	Pipeline	68
2.3.3.1	Quality control and trimming	68
2.3.3.2	Sequence count matrix	69
2.3.3.3	Annotation preparation	69
2.3.3.4	Alignment	71
2.3.3.5	Reads annotation and assignment	71
2.3.3.6	Assignment rules	72
2.3.3.7	Features tree	75
2.3.3.8	Reads assignment ambiguity	75
2.3.3.9	Construction of TreeSummarizedExperiment object	76
2.3.3.10	Differential analysis	76
2.3.3.11	Table of important functions	76
2.4	Results	77
2.4.1	Datasets used for testing the pipeline	77
2.4.2	Comparison of quantification and identification of miRNAs	77
2.4.2.1	shortRNA quantification is positively correlated with Oasis2 and Sports1	77
2.4.2.2	shortRNA accurately identify miRNAs	78
2.4.3	Data simulation to check the alignment and reads assignment by short-RNA	78
2.4.4	Differential expression analysis	80

2.4.5	Validation using human data and quantitative real-time PCR	80
2.4.6	Qualitative comparison with other tools	81
2.5	Discussion & Outlook	81
2.6	Limitations	83
2.7	Conclusion	84
2.8	Data and code availability	84
Conclusion and outlook	101
2.9	Overall conclusion	101
	MSUS-associated differences	101
	RNA sequencing data analysis pipeline	102
	ATAC sequencing data analysis pipeline	102
	Updating manuscript from Chapter 1 in light of latest RNA-seq data	103
	Low correlations of quantification between shortRNA and qPCR	104
Appendix A	105
	Datasets analyzed from lab's MSUS model	105
	Human dataset from the Schratt lab	105
	Mouse datasets analyzed from literature	106
Appendix B	107
	Other publications and manuscripts during doctoral thesis work	107
	Involvement of circulating factors in the transmission of paternal experiences through the germline	107
	Symposium summary: Epigenetic inheritance-impact for biology and society 26–28 August 2019, Zurich, Switzerland	107
	OmniSperm: Multiomic analyses of sperm and offspring production from a single male	108
	Early life stress alters chromatin accessibility landscape and transcript usage in spermatogonial cells during postnatal testis maturation	108
Appendix C	109
	shortRNA analysis steps in R (a short vignette)	109
	Package installation	109
	Quality check and trimming	109
	FastQ files to sequence by counts matrix	109
	Unique sequences for alignment	109

Obtaining databases for analysis	109
Annotation preparation and genome index generation for alignment	110
Alignment	110
Overlapping aligned data with annotations	110
Assignment of overlapping reads to features with assignment rules	111
Converting features to FactorList	111
Assignment of reads to the features tree	111
Creation of TreeSummarizedExperiment object	111
Differential analysis	112
Appendix D	113
Simulated reads for testing shortRNA	113
Libraries	113
miRNAs	113
cluster chr17:17830188-17830879	113
Imputing differences for differential analysis	114
Read counts	114
miRNA precursor	114
Data to be simulated	115
Appendix E	117
2.10 CV	118
References	123

List of Tables

List of Figures

1	History of the term epigenetics	29
2	Graphical summary of the biology of germline stem cell specification and transitions	30
3	Graphical summary of the biology of germline stem cell specification and transitions	31
4	Overview of version control system	34
5	Organization of data analysis projects	35
6	Workflow for quality check and quality control for sequencing data	36
7	Workflow for RNA sequencing data analysis	38
8	Workflow for ATAC sequencing data analysis	48
9	Workflow for ChIP sequencing data analysis	49
10	Workflow for Bisulfite sequencing data analysis	50
11	Workflow for multi-omics data integration	51
1.1	Log ₂ fold-change comparison of RNA-seq data from the literature and lab (old and new)	55
1.2	Volcano plot of differentially expressed genes	56
1.3	K-means clustering of differentially expressed genes	57
1.4	Enriched heatmap of distal DARs	59
1.5	Enriched heatmap of proximal DARs	60
1.6	Enriched heatmap of distal DARs with ChIP-seq and BS-seq data	61
1.7	Enriched heatmap of proximal DARs with ChIP-seq and BS-seq data	62
2.1	Schematic of A: SummarizedExperiment (Morgan et al., 2020), B: TreeSummarizedExperiment (R. Huang et al., 2020)	85
2.2	Overview of shortRNA pipeline	86

2.3	Benchmark of four different Bioconductor tools for reading and storing FASTQ file quality data. Performance of four tools is measured in A : the amount of time they take for reading the file, B : the amount of memory they take to store the data.	87
2.4	Exemplary QC report plots of a sperm sample from (Katharina Gapp et al., 2021). A : Quality control table, B : Quality score lines plot at each nucleotide in the sequencing reads, C : Barplot of reads duplication level with the mean GC ratio represented as lines, D : Histogram of reads distribution, before and after quality control	88
2.5	Schematic of quality check and quality control of FASTQ files, UMIs collapsing (if present), and collapsing FASTQ files to obtain unique sequences for alignment and making a sequence by samples counts matrix	89
2.6	Schematic of obtaining database, preparing annotations and custom genome indexes for alignment	90
2.7	Schematic of alignment, overlapping features and aligned sequences, and assigning reads to features using assignment rules	91
2.8	General framework of feature tree and assignment of sequences to the tree. A : Example of two sequences. B : Alignment of both the sequences to miRNA. C : General outline of the feature tree and assignment of both the sequences to the tree	92
2.9	Hierarchical organization of A : tRNA features and B : mitochondrial tRNA features.	92
2.10	Hierarchical organization of miRNA features	93
2.11	Multiple assignments of a sequence mapping to a tRNA. A : Exemplary sequence. B : Multiple sequence alignment plot of the sequence and the features to which it is mapping. C : Assignment of the reads to the tRNA feature tree	94
2.12	Schematic of assigning reads to the features tree of features, creation of TreeSummarizedExperiment object, and downstream analysis and plotting .	95
2.13	Comparison of Log2 TMM normalized quantification of miRNAs from shortRNA with Oasis2, Sports1, and seqpac	96
2.14	Upset plot of overlaps of detected features between four methods: shortRNA, Oasis2, Sports1, and seqpac	96
2.15	Genomic plots of simulated reads from miRNA precursor.	97
2.16	Hierarchical representation of features and assigned reads to them. Differential expression analysis of features from two miRNA clusters. Simulated read nodes are highlighted in orange colour	97

2.17 Comparison of quantifications of five miRNAs from four different tools with - Delta Ct values from qPCR. A: Overall correlation. B: Correlation at individual miRNA	98
2.18a Genomic plots of validated miRNAs	99
2.18b Genomic plots of validated miRNAs	100

List of Abbreviations

?? ??

?? ??

Summary

The environment has a strong influence on an organism's development, especially in the early stages. Environmental influences and life experiences can alter the phenotypic features of exposed individuals and their offspring in many animals. In mammals, traumatic stress is a type of environmental event that can change their behaviour, cognition, and physiological functions. Some of the impacts of severe stress can be passed down through generations, even if those generations have never been exposed to a traumatic stressor. Environmentally generated effects that involve non-genetic germline alterations can be an important route of trait transmission over generations. Even if the subsequent generations have never been exposed to similar stressful exposures, they may be impaired in similar functions. The host lab's Unpredictable maternal separation combined with unpredictable maternal stress (MSUS) mouse model is perfect for learning more about the mechanisms behind epigenetic inheritance induced by environmental perturbations.

In this thesis, to study the mechanisms of epigenetic inheritance, multi-omics data analysis pipelines were built, Methods, and a number of high-throughput sequencing datasets generated using the MSUS model were analyzed using these pipelines, Appendix 1. Also, the datasets generated from spermatogonial cells were used to study the developmental trajectory across postnatal and adult stages. Further, the effect of early life trauma on epididymal extracellular vesicles transcriptome was investigated using short RNA sequencing data. At last, a tool was built to analyze short RNA sequencing data to address issues in the current tools and pipelines.

Spermatogonial cells are the postnatal initiators of spermatogenesis, and they undertake critical proliferative and differentiation tasks throughout one's life. They are the only type of stem cell capable of transmitting genetic information to an embryo. Open chromatin in spermatogonial cells undergoes significant remodelling throughout testes development. Throughout postnatal testes development, they have significantly different transcriptional pathways. We investigated the dynamics of chromatin accessibility and gene expression during the developmental stages of PND8, PND15, and PNW21 using multi-omics data collected from spermatogonial cells. By comparing chromatin accessibility changes between early postnatal

and steady-state adult cells in the mouse testes, the age-dependent molecular phenotype of spermatogonial cells was identified. The differentially accessible regions were discovered to be relevant to developmental and metabolic processes and enriched for important transcription factors. Also, the accessibility of chromatin at transposable elements in spermatogonia has been examined. Specific transposable elements subtypes become less accessible as the adult testis matures, whereas specific LINE L1 subtypes become more accessible. The genome's olfactory receptor gene regions were discovered to be highly connected to several of the more accessible subtypes. In conclusion, these findings reveal that chromatin accessibility in spermatogonial cells changes from postnatal to adult stage, as well as interesting biological patterns when combined with expression and histone modifications data. Furthermore, we show how bioinformatics approaches may be used to integrate multi-omics datasets. These results are available in Chapter 1.

If environmentally-triggered marks are not inherited from SSCs, another possible way in which sperm may carry them is through the uptake of extracellular vesicles, in particular those produced by epididymal cells. By using the MSUS model, we studied the impact of early postnatal stress on the transcriptome of epididymal extracellular vesicles. The short RNA profile of epididymal EVs, particularly miRNAs, were found to be altered in adult males exposed to postnatal stress. In certain cases, these miRNA changes have been related to differences in the expression of their target genes in sperm and zygotes produced from those sperm. To summarize, chronic stress in early postnatal life changes miRNAs in mature extracellular vesicles of the male reproductive tract, with implications for mature sperm and zygotes. These results are summarized in Chapter 2. While working on Chapters 1 and 2, we discovered a lot of issues with tools available for analyzing short RNA sequencing data. Short RNAs are significant molecules that play an important role in the control of the genome. Short RNAs have been classified into several categories, including miRNA, tRNA and tRNA fragments, and piRNAs, all of which have a complex biosynthesis. As a result, an analysis framework is necessary that is both sufficiently specialized to reflect the specificities of different classes of short RNA and their biogenesis, as well as sufficiently generic and thorough to conduct global analyses. We created a user-friendly, highly customizable, and comprehensive R package to do a full end-to-end analysis of short RNA sequencing data. Our package has been specifically designed to account for changes in short RNA types and their biogenesis, and it may readily be modified to include further annotation. We also use a customised genome annotation with pseudo chromosomes to account for post-transcriptional modifications, as well as a flexible rule-based approach to allocate reads along a tree of hierarchically ordered attributes. This enables systematic querying, exploration, and differential expression analysis of short RNAs at several levels of granularity, ranging from single sequences to entire RNA

classes. The shortRNA R package, which is platform-agnostic, was designed to carry out all of the analysis steps from within R. These results are available in Chapter 3.

In summary, the work presented in this thesis investigated various mechanisms of epigenetic inheritance and developed several pipelines for data analysis from the MSUS model, investigated changes in spermatogonial cells across development and the impact of early life stress on the transcriptome of epididymal extracellular vesicles, and developed an R package for short RNA sequencing data analysis.

Zusammenfassung

Die Umwelt hat einen starken Einfluss auf die Entwicklung eines Organismus, insbesondere in den frühen Stadien. Umwelteinflüsse und Lebenserfahrungen können bei vielen Tieren die phänotypischen Merkmale exponierter Individuen und ihrer Nachkommen verändern. Bei Säugetieren ist traumatischer Stress eine Art von Umweltereignis, das ihr Verhalten, ihre Wahrnehmung und ihre physiologischen Funktionen verändern kann. Einige der Auswirkungen von starkem Stress können über Generationen weitergegeben werden, auch wenn diese Generationen noch nie einem traumatischen Stressor ausgesetzt waren. Umweltbedingte Effekte, die nicht-genetische Veränderungen der Keimbahn beinhalten, können ein wichtiger Übertragungsweg von Merkmalen über Generationen hinweg sein. Auch wenn die nachfolgenden Generationen noch nie ähnlichen Belastungen ausgesetzt waren, können sie in ähnlichen Funktionen beeinträchtigt sein. Das Mausmodell Unvorhersehbare mütterliche Trennung in Kombination mit unvorhersehbarem mütterlichem Stress (MSUS) des Wirtslabors ist perfekt, um mehr über die Mechanismen hinter der epigenetischen Vererbung durch Umweltstörungen zu erfahren.

Um die Mechanismen der epigenetischen Vererbung zu untersuchen, wurden in dieser Dissertation Multi-Omics-Datenanalysepipelines erstellt, Methods, und eine Reihe von Hochdurchsatz-Sequenzierungsdatensätzen, die mit dem MSUS-Modell generiert wurden, wurden unter Verwendung dieser Pipelines analysiert, Anhang 1. Außerdem wurden die aus Spermatogonien generierten Datensätze verwendet, um den Entwicklungsverlauf über postnatale und adulte Stadien hinweg zu untersuchen. Darüber hinaus wurde die Wirkung eines frühen Lebenstraumas auf das Transkriptom der extrazellulären Vesikel der Nebenhoden unter Verwendung von kurzen RNA-Sequenzierungsdaten untersucht. Schließlich wurde ein Tool entwickelt, um kurze RNA-Sequenzierungsdaten zu analysieren, um Probleme in den aktuellen Tools und Pipelines zu beheben.

Spermatogoniale Zellen sind die postnatalen Initiatoren der Spermatogenese und übernehmen lebenslang kritische Proliferations- und Differenzierungsaufgaben. Sie sind die einzige Art von Stammzellen, die genetische Informationen an einen Embryo übertragen können. Offenes Chromatin in Spermatogonien wird während der Hodenentwicklung

signifikant umgebaut. Während der postnatalen Hodenentwicklung weisen sie signifikant unterschiedliche Transkriptionswege auf. Wir untersuchten die Dynamik der Zugänglichkeit von Chromatin und der Genexpression während der Entwicklungsstadien von PND8, PND15 und PNW21 unter Verwendung von Multi-Omics-Daten, die von Spermatogonien gesammelt wurden. Durch den Vergleich der Veränderungen der Chromatin-Zugänglichkeit zwischen frühen postnatalen und stationären adulten Zellen in den Hoden von Mäusen wurde der altersabhängige molekulare Phänotyp von Spermatogonien identifiziert. Die unterschiedlich zugänglichen Regionen wurden als relevant für Entwicklungs- und Stoffwechselprozesse entdeckt und um wichtige Transkriptionsfaktoren angereichert. Auch die Zugänglichkeit von Chromatin an transponierbaren Elementen in Spermatogonien wurde untersucht. Spezifische Subtypen der transponierbaren Elemente werden mit zunehmender Reife des adulten Hodens weniger zugänglich, während spezifische LINE L1-Subtypen leichter zugänglich werden. Es wurde entdeckt, dass die olfaktorischen Rezeptor-Genregionen des Genoms stark mit mehreren leichter zugänglichen Subtypen verbunden sind. Zusammenfassend zeigen diese Ergebnisse, dass sich die Zugänglichkeit von Chromatin in Spermatogonien vom postnatalen bis zum adulten Stadium ändert, sowie interessante biologische Muster, wenn sie mit Expressions- und Histonmodifikationsdaten kombiniert werden. Darüber hinaus zeigen wir, wie bioinformatische Ansätze verwendet werden können, um Multi-Omics-Datensätze zu integrieren. Diese Ergebnisse sind inverfügbar Kapitel 1.

Wenn umweltbedingte Markierungen nicht von SSCs vererbt werden, können Spermien sie auch durch die Aufnahme extrazellulärer Vesikel, insbesondere solcher, die von Nebenhodenzellen produziert werden, tragen. Unter Verwendung des MSUS-Modells untersuchten wir den Einfluss von frühem postnatalem Stress auf das Transkriptom extrazellulärer Nebenhodenvesikel. Es wurde festgestellt, dass das kurze RNA-Profil von Nebenhoden-EVs, insbesondere miRNAs, bei erwachsenen Männern, die postnatalem Stress ausgesetzt waren, verändert war. In bestimmten Fällen wurden diese miRNA-Veränderungen mit Unterschieden in der Expression ihrer Zielgene in Spermien und Zygoten, die aus diesen Spermien produziert wurden, in Verbindung gebracht. Zusammenfassend lässt sich sagen, dass chronischer Stress im frühen postnatalen Leben miRNAs in reifen extrazellulären Vesikeln des männlichen Fortpflanzungstrakts verändert, mit Auswirkungen auf reife Spermien und Zygoten. Diese Ergebnisse sind in Kapitel 2 zusammengefasst.

Während der Arbeit an den Kapiteln 1 und 2 entdeckten wir viele Probleme mit verfügbaren Werkzeugen für die Analyse von kurzen RNA-Sequenzierungsdaten. Kurze RNAs sind bedeutende Moleküle, die eine wichtige Rolle bei der Kontrolle des Genoms spielen. Kurze RNAs wurden in verschiedene Kategorien eingeteilt, darunter miRNA, tRNA und tRNA-Fragmente sowie piRNAs, die alle eine komplexe Biosynthese aufweisen. Daher ist

ein Analyserahmen erforderlich, der sowohl ausreichend spezialisiert ist, um die Spezifitäten verschiedener Klassen kurzer RNA und deren Biogenese widerzuspiegeln, als auch ausreichend generisch und gründlich ist, um globale Analysen durchzuführen. Wir haben ein benutzerfreundliches, hochgradig anpassbares und umfassendes R-Paket erstellt, um eine vollständige End-to-End-Analyse von kurzen RNA-Sequenzierungsdaten durchzuführen. Unser Paket wurde speziell entwickelt, um Änderungen in kurzen RNA-Typen und deren Biogenese zu berücksichtigen, und es kann leicht modifiziert werden, um weitere Anmerkungen aufzunehmen. Wir verwenden auch eine angepasste Genom-Annotation mit Pseudochromosomen, um posttranskriptionelle Modifikationen zu berücksichtigen, sowie einen flexiblen regelbasierten Ansatz, um Reads entlang eines Baums hierarchisch geordneter Attribute zuzuordnen. Dies ermöglicht eine systematische Abfrage, Exploration und differentielle Expressionsanalyse von kurzen RNAs auf mehreren Granularitätsebenen, die von einzelnen Sequenzen bis hin zu ganzen RNA-Klassen reichen. Das plattformunabhängige Paket short-RNA R wurde entwickelt, um alle Analyseschritte innerhalb von R durchzuführen. Diese Ergebnisse sind inverfügbar Kapitel 3.

Zusammenfassend untersuchten die in dieser Dissertation präsentierten Arbeiten verschiedene Mechanismen der epigenetischen Vererbung und entwickelten mehrere Pipelines für die Datenanalyse aus dem MSUS-Modell, untersuchten Veränderungen in Spermatogonien während der Entwicklung und den Einfluss von Stress im frühen Leben auf das Transkriptom extrazellulärer Nebenhodenvesikel, und ein R-Paket für die Analyse von kurzen RNA-Sequenzierungsdaten entwickelt.

Thesis overview

This thesis is divided into three parts: introduction and methods, scientific contributions, and concluding remarks. First, I introduce the biological context for my thesis, followed by the method section with the flowcharts and description of various data analysis pipelines. Next, in scientific contributions, there are three chapters: Chapter 1 is the latest version of the manuscript submitted to BioRxiv (Lazar-Contes et al. 2020), Chapter 2 is published (Alshanbayeva et al. 2021), and chapter 3 is included as a descriptive chapter on shortRNA and a manuscript is in preparation. My contribution to each paper is summarized at the beginning of each chapter. In the end, the thesis concludes with a summary and outlook.

Introduction

Epigenetics

Conrad Hal Waddington, a British developmental biologist, coined the term “epigenetics” in 1942. He was also a pioneer in the fields of epigenetics, systems biology, and evolutionary developmental biology. What came to be known as “Waddington’s landscape” describes how genes interact with the environment to produce, from the same genotype, the variety of cellular phenotypes of a multicellular organism (Waddington, 1942). While epigenetics has been given various meanings throughout history (Figure 1, taken from (Pisco, Fouquier d’Herouel, & Huang, 2016), it can be broadly understood as the study of how our genes, which we inherit from our forefathers and mothers, are controlled and interact with the environment: how our genes form who we are. Epigenetics is the study of how additional factors communicate with genes to guide the processes that make our cells and bodies function. “Epi” means “on top of” or “in addition to.”

Epigenetic mechanisms regulate the transcriptional programs in response to developmental and environmental cues and encompass a variety of pre-transcriptional to post-translational events, including DNA methylation, post-translational histone modifications (e.g. acetylation and methylation), histone variations, non-coding RNA (ncRNA) and their modifications, and nucleosome remodelling (Suziarjo, 2016). Epigenetic mechanisms in the somatic cell regulate vital cellular function, including differentiation or adaptation of cellular function in response to environmental indications. However, epigenetic control in germ cells can trigger changes in phenotype over many successive generations, known as transgenerational epigenetic inheritance, section Transgenerational epigenetic inheritance, and contribute to changes in susceptibility to health and disease.

Germline epigenetics

Spermatogonial Stem Cells

The germ stem cells of the seminiferous epithelium in the testis are known as spermatogonial stem cells (SSCs). They are the building blocks of spermatogenesis, the mechanism of producing mature male germ cells. The SSCs are responsible for the continuous creation of the massive number of sperm cells required for a male's reproductive lifespan. They are present in the testis from birth and for the rest of the male's life. SSCs have a dual biological function: they replicate themselves to maintain the stem cell pool while also generating progenitors that will differentiate into spermatozoa. They begin life in the testicular seminiferous tubules and continue to serve this dual purpose throughout adulthood. SSCs are the male adult stem cells of the seminiferous epithelium in the testis that are capable of transferring a species' genome from generation to generation while also possessing the ability to differentiate into pluripotent stem cells.

In mice, the primordial germ cells (PGCs), the precursors of prospermatozoa, appear at 5.5 days post-coitum (dpc). Prospermatozoa establishes SSCs shortly after birth (gonocytes). PGCs undergo global DNA demethylation (which involves the full erasure of parental imprints) through both active and passive mechanisms as they migrate toward the gonads. DNA methylation (DNAm) is gradually restored after that (between embryonic day 13.5 [E13.5] and E16.5), and both maternal and paternal-specific imprints are believed to be completely formed by birth in male prospermatozoa (Hammoud et al., 2015).

The prospermatozoa are mitotically arrested at birth, but they begin cycling on postnatal days 1–2 (PND1 – PND2). A population of spermatogonial cells continues to proliferate throughout the first week after birth, giving rise to the undifferentiated A_{single} , A_{paired} , and A_{aligned} spermatogonial pools (Kolasa, Misiakiewicz, Marchlewicz, & Wiszniewska, 2012). Cycling spermatogonia proliferate and populate the seminiferous tubule over the next week (PND3–PND10); here, a subset gives rise to self-renewing SSCs, while the rest differentiates without self-renewal, resulting in the first wave of spermatogenesis or commits apoptosis. As a result, this stage of germ cell growth is critical for initiating and maintaining male fertility during adulthood. Following puberty, after 35–37 days, SSCs go through many phases of proliferation, eventually forming sister stem cells that will maintain a stable pool or more differentiated intermediates that will initiate meiosis I and II and form mature sperm cells. Please see Figure 2 for an overview. These transitions are timed to correlate with global epigenome shifts, which represent the spermatogonial cell's developmental path, current transcriptional program, and future commitment to differentiation. Since the transcriptome and epigenetic landscape of spermatogonial stem cells is so complex during proliferation and

differentiation, drastic changes in the environment may not go unnoticed.

It is difficult to distinguish SSCs from progenitor spermatogonia that don't have stem cell activity and are committed to differentiation. Specific cell surface markers like Thy1 and TSPAN8 can be used to enrich SSCs. Further, cells extracted from Id4-eGFP transgenic mice that express a high amount of GFP signal, and combinations of reporter gene and cell surface marker signal can be used to further enrich the population (F. Chan et al., 2014; Mutoji et al., 2016).

Reprogramming

Reprogramming refers to the erasure and remodelling of epigenetic markers such as DNAmc throughout mammalian development (Reik, Dean, & Walter, 2001). Diverse cells and tissues acquire different gene expression patterns during the evolution of multicellular organisms. Some cells undergo extensive epigenetic reprogramming in normal developmental or pathological conditions, which involves the removal of epigenetic marks in the nucleus and the installation of a new set of marks. Genome-wide epigenetic reprogramming occurs at stages when the developmental potency of cells changes. There is passive DNA demethylation and further reorganization of histone modifications. The epigenetic reprogramming is likely to be needed for totipotency, correct initiation of embryonic gene expression, and early lineage development in the embryo. Extensive epigenetic reprogramming of DNA and histone marks occur in mammalian development in totipotent early embryos and pluripotent germ cells (Canovas & Ross, 2016).

The germ cell lineage is unique in carrying genetic and epigenetic information from one generation to the next. In the germline, epigenetic reprogramming resets genomic potential and erases epigenetic memory. Epigenetic reprogramming involves genome-wide demethylation of 5mC, which is important for genome imprinting, X inactivation, transposon silencing, centromeric/telomeric structural stability, and gene expression (Surani, Hayashi, & Hajkova, 2007). After a sperm fertilizes an ovum, nearly all epigenetic marks are reprogrammed in two short intervals early in development. By eliminating random changes in epigenetic marks (i.e., epimutations) that have occurred in the germ cells, restores the ability of the fertilized egg cell (i.e., zygote) to develop into all the different cell types and tissues (Reik et al., 2001). During embryo development, there are two waves of epigenetic reprogramming, Figure 3. After fertilization in the preimplantation embryo (i.e., the blastocyst), the first phase begins; preimplantation refers to the time between the embryo implanting and forming a placenta. The embryo in mice undergoes genome-wide demethylation after fertilization, which is finished by embryonic day 5. (E5) (Reik et al., 2001; Smallwood & Kelsey, 2012).

The paternal genome undergoes rapid, active demethylation, whereas the maternal genome experiences passive demethylation. The embryonic DNA begins to be re-methylated on day E5 and is finished before birth. The second wave of epigenetic reprogramming takes place in the germ cells of the developing embryo, which will eventually give rise to gametes with sex-specific epigenetic markers (gametogenesis). The methylation signatures of the parental genomes are found in the primordial germ cells (PGCs), which are progenitors to both male and female germ cells in the developing embryo. PGCs are epiblast cells that first appear in the posterior primitive streak at E7.5, after which they begin to migrate from E8.5 to the genital ridge, arriving by E11.5. PGCs undergo rapid demethylation beginning around E7–8, which is completed by E15–16. In PGCs, parental imprints are erased and totipotency is restored, followed by the restoration of sex-specific methylation. Reprogramming is complete in the male germline at birth, but not in the female germline until adolescence.

Embryonic development involves epigenetic reprogramming to replenish totipotency from a germline state in each generation. To begin with, because there is no hereditary germplasm and germ cell precursors are formed from a post-implantation epiblast that has acquired epigenetic markers throughout early development, a germline reprogramming phase may be required to ensure that this acquired information is erased. Then, in the post-implantation epiblast, the parent-of-origin markers of mature gametes fuse in the zygote and are passed down to nascent PGCs from their predecessors. Another goal could be to reverse the effects of accumulating epigenetic changes that are influenced by both genetic and environmental factors. However, epigenetic information can be inherited through the germline (or after fertilization) on occasion, which is most likely due to insufficient germline erasure. Finally, demethylation in the germline may have the added benefit of reducing the frequency of mutations caused by 5-methylcytosine deamination.

Epigenetic Inheritance

Inheritance is the process of passing traits or information from one generation of individuals or cells to the next. Charles Darwin proposed natural selection as the primary mechanism for species evolution in his 1859 book: “On the origin of species”. According to Darwin, natural selection is a key mechanism of evolution, characterized by changes in the heritable traits of a population over generations. Jean-Baptiste Lamark was an early proponent of the idea that biological evolution occurred and proceeded in accordance with natural laws. He is well known for the principle of “inheritance of acquired characteristics”, an idea known as Lamarckism. According to Lamarckian theories of inheritance, physical traits inherited by a generation during its lifetime (e.g. strength and body size) are passed down to offspring. In

1868, Darwin proposed pangenesis theory as a supplement to his 1859 theory of evolution by natural selection in his book “The Variation of Animals and Plants Under Domestication.” The term pangenesis was coined by Darwin, with “pan-” referring to the whole and “genesis” referring to the origin. Using the term pangenesis, Darwin proposed that all parts of the parents could contribute to the evolution and development of the offspring. Darwin’s theory of pangenesis was based on the idea that somatic cells would release ‘gemmules’ or ‘pangenes’ in response to environmental stimulation (use and disuse), which travelled around the body but not necessarily in the bloodstream. These pangenes were microscopic particles that supposedly contained information about the characteristics of their parent cell, and Darwin believed that they eventually accumulated in germ cells, where they could pass on the newly acquired characteristics of the parents to the next generation (Darwin, 1871). As a result, Darwin advocated Lamarckian use and disuse inheritance, as well as Lamarckian soft inheritance. In the 1880s, August Weismann challenged this concept. Weismann proposed the germ plasm theory, which states that germ cells in the gonads contain information that is passed down from generation to generation unaffected by experience and independent of somatic cells (Weismann, 1893). Hereditary information is only transferred from germline cells to somatic cells. This is referred to as the Weismann barrier.

However, recent epigenetic discoveries have revealed that acquired traits can be passed down to future generations. Our recent work showed that serum from adult male mice exposed to postnatal stress, injected chronically to naïve males can induce metabolic symptoms associated with stress in the offspring of the injected males (van Steenwyk et al., 2020). Also, we identified circulating extracellular vesicles as vectors of communication with germ cells that can reproduce symptoms of exposure in the offspring when injected into fathers *in vivo* (Alshanbayeva, Tanwar, Roszkowski, Manuella, & Mansuy, 2021). These results highlight the importance of circulating factors in the mechanisms of epigenetic inheritance, Chapter 2.

Transgenerational Epigenetic Inheritance

The transmission of epigenetic information from one generation to the next is referred to as intergenerational epigenetic inheritance, while that across multiple generations is referred to as transgenerational epigenetic inheritance. Both intergenerational and transgenerational imply the parents’ exposure to the environmental factor in question, as well as the gametes they already have. Transgenerational exposure refers to people who have never been specifically exposed to the environmental factor during their lifetime (either as individuals or as future gametes). This may be the third generation after *in utero* exposure or the second generation after preconception exposure. It is assumed that fetal or *in utero* exposure to an environmental

factor is not heritable because the offspring are specifically exposed. However, the mechanisms of adverse development following *in utero* exposure to drugs, alcohol, diet, and other factors also influence offspring through epigenetic mechanisms.

Reprogramming in epigenetic inheritance

Epigenetic reprogramming refers to the process of removing and resetting relevant epigenetic information in germ cells, which is primarily accomplished through DNA demethylation. Epigenetic reprogramming prevents the transmission of epigenetic information from mature gametes, which is in line with the germ-plasm theory of Weissman (Weismann, 1893). The extent of epigenetic reprogramming, however, is unknown, leaving the question of how much epigenetic information from adult gametes is preserved in the embryo unanswered. In mammals, DNAm and histone marks are effectively reset during reprogramming both in the germline and in the zygote immediately after conception, leaving epigenetic changes with minimal possibility of passing down. Evidence for the survival of histone modifications across reprogramming has been identified in sperm, where histones are largely replaced by protamines. Surprisingly, H2K27me3 enrichment has been discovered in reprogramming resistant sperm nucleosomes (Casas & Vavouri, 2020), suggesting that this mark may be a hereditary signal. Further, findings in (Kremsky & Corces, 2020) imply that transcription factors can operate as epigenetic information carriers during germ cell and pre-implantation development by maintaining the CpG methylation state.

Models of Epigenetic Inheritance

Given limitations such as the inability to conduct controlled experiments in humans and the availability of data for successive generations, rodent models have become critical for correlating human findings (Jawaid, Jehle, & Mansuy, 2021) and to investigate the penetrance of effects across generations to provide mechanistic insight. There are several experimental paradigms for studying epigenetic inheritance. In mammals, dietary changes, stress, medications, and environmental toxins have all been linked to altered phenotypes in offspring (K. Gapp & Bohacek, 2018). Epigenetic inheritance appears to be initiated or modulated by mutations in the epigenetic machinery, such as DNA/RNA methylation enzymes or histone modification complexes, such as overexpression of Kdm1a, an H3K4me3 demethylase in mice (Siklenka et al., 2015).

Alternatively, environmental perturbations may initiate epigenetic inheritance, with various insults leading to offspring phenotypes in a variety of model systems (Jawaid & Mansuy,

2019). Environmental factors like vinclozolin (an antiandrogenic compound), methoxychlor (an estrogenic compound), stress responses, high-fat diet, high-sugar diet, low-protein diet, air pollutant exposure in pregnant mice, In utero Bisphenol-A bis(2- Ethylhexyl) phthalate (DEHP) and dibutyl phthalate (DBP) exposure in rats, In utero exposure to jet fuel (JP-8) in mice, In utero valproic acid exposure in mice, Olfactory fear conditioning in F0 mice, and Female mice fed a high-fat/high-sugar diet have been linked to epigenetic inheritance (Bohacek & Mansuy, 2015; Hanson & Skinner, 2016; Heard & Martienssen, 2014; Mao, Li, Liu, & Chi, 2017; Nagy & Turecki, 2015; Rando, 2016; Schaefer & Nadeau, 2015). These models shed new light on the mechanisms of epigenetic inheritance. However, distinguishing evidence of epigenetic mark transmission from the evidence of exposure transmission is difficult.

MSUS mouse model

A unique transgenerational model to assess the effects of trauma: unpredictable maternal separation combined with unpredictable maternal stress (MSUS) is a firmly established transgenerational model in the Mansuy lab, that does not involve any chemical or mutagenic agent but depends only on psychological and emotional trauma, Figure 4. It is an ideal model to study the possible transgenerational effects of life experiences on the genome and the epigenome. MSUS consists in separating mouse pups from their mother unpredictably (any time during the day) each day for 3 hours from postnatal day 1 (PND1) to PND14 and in exposing the mothers to stressful situations such as restraint stress or forced swim in cold water (18°C) unpredictably during separation. Filial 1 (F1) generation males are mated with naïve control females, and F2 males are mated with control females to obtain F3. MSUS induces several behavioural symptoms including depression, increased risk-taking, antisocial behaviours and memory deficits, and metabolic alterations across several generations (Franklin et al., 2010; van Steenwyk, Roszkowski, Manuella, Franklin, & Mansuy, 2018). Further, metabolic phenotypes like decreased bodyweight, insulin hypersensitivity and glucose intolerance are also observed. For studying the molecular mechanisms of epigenetic inheritance, a number of datasets have been collected and sequenced, Figure 5.

Vectors of epigenetic inheritance

Epigenetic marks, such as DNAme, histone modifications, and non-coding RNAs, have been proposed as information-carrying vectors passed down through generations. These three candidates have been investigated in a variety of EI paradigms. It is hypothesized that stable changes in epigenetic marks caused by an environmental insult (epimutations) may

affect gene expression or chromatin stability, increasing the risk of a phenotype or disease in offspring (Miska and Ferguson-Smith, 2016). DNAme, histone modifications, RNA, and ncRNAs are all interconnected and interdependent within individual cells. They are likely to act in a coordinated and collaborative manner to cause disease phenotypes in the context of epigenetic inheritance.

DNA methylation

The most common type of DNAme in mammals is cytosine residue methylation. Cytosine residues are methylated on the 5' carbon and are referred to as 5-methylcytosines (5mC). Methylation of a cytosine residue is a binary process – either methylation or non-methylation. In mammals, DNA methylation is typically found in the context of CpG dinucleotides (Bird, 2002). Cytosine methylation can be copied from one DNA strand to the next in this manner during replication and meiosis. DNAme patterns, on the other hand, are dynamic. The patterns of DNAme vary over time and space, changing during development and differing between cell types (Luo, Hajkova, & Ecker, 2018). DNA methyltransferase (DNMT) enzymes methylate cytosine residues. DNMT1 is a methyltransferase that reads hemimethylated DNA after cell division and adds methylation to the newly synthesized strand (Z. Chen & Riggs, 2011). The de novo methyltransferases DNMT3a and 3b can add methylation to previously unmethylated residues (Smith & Meissner, 2013). Although DNAme in gene control regions is generally associated with gene repression, its precise role in gene regulation is likely locus dependent (Deaton & Bird, 2011). Methylated DNA is a true epigenetic mark because there are “readers” that preferentially recognize 5mC and can interpret the methylation mark’s meaning (H. Zhu, Wang, & Qian, 2016). Many properties of DNAme make it an appealing mechanistic candidate for epigenetic inheritance, including the following: it can be environmentally modulated; machinery exists to replicate methylation patterns onto newly synthesised DNA, making it mitotically heritable; and methylated loci resistant to epigenetic reprogramming have been identified in the zygote and germline. Disruption of the DNAme machinery may also play a role in the initiation of EI.

Histone modifications

In somatic cells, DNA is wrapped around an octameric core of histone proteins (H2A, H2B, H3, and H4) to form a nucleosome (Marmorstein & Trievel, 2009). Histones’ N-terminal tails and globular domains undergo post-translational modifications (PTMs). Histone modifications are dynamic during development, vary across tissues, are regulated by specific enzymes, play a significant role in gene expression, and interact with other epigenetic

control systems such as DNAm (Barth & Imhof, 2010). On the histone tails, there are over 60 different amino acid residues where modifications have been detected (Iwasaki et al., 2013). Histones can undergo various modifications, including acetylation, methylation, phosphorylation, ubiquitination and sumoylation among others (Kouzarides, 2007). Histone methyltransferase (HMT) enzymes methylate histones on Lysine (K) and Arginine (R) amino acids. Histone lysine residues can be mono-, di-, or tri-methylated, and these modifications can function as either active or repressive marks. Methylation of H3K4, H3K36, and H3K79 is associated with transcriptional activation, whereas methylation of H3K9, H3K27, and H4K20 is associated with transcriptional repression (Sims, Houston, Magazinnik, & Rice, 2006; Vakoc, Sachdeva, Wang, & Blobel, 2006). These modifications can influence chromatin packaging and nucleosome positioning. Histone modifications, like DNAm, are spatially and temporally dynamic. Certain histone modifications, such as trimethylation of lysine 27 on histone 3 (H3K27me3), are generally associated with gene repression, whereas acetylation of the same residue (H3K27ac) is generally associated with gene activation (Lawrence, Daujat, & Schneider, 2016). Histone modifications regulate gene expression by recruiting protein complexes and controlling the accessibility of DNA to the transcription machinery (Lawrence et al., 2016). Histone modifications are a plausible mechanistic candidate in EI because they can be modulated by the environment. During epigenetic reprogramming, some histone modifications in the germline and zygote are retained. Histones are spread onto newly assembled chromatin by sophisticated machinery (Marmorstein & Trievel, 2009). However, it is still unclear how histone codes are maintained during mitosis and meiosis (Erkek et al., 2013). Furthermore, during spermatogenesis, most histones are removed and replaced with protamines (Casas & Vavouri, 2014). Sperm contains protamine modifications that are similar to histone modifications (Brunner, Nanni, & Mansuy, 2014), despite this, their functional significance remains unknown, particularly during epigenetic inheritance. Disruption of the machinery that regulates histone modification could also be a key initiator of EI (Siklenka et al., 2015). H3K9 and H3K27 methylation appear to be the most likely to be epigenetic among the various histone lysine methylation states, not only because they are key regulators of classic epigenetic phenomena (C. Huang, Xu, & Zhu, 2013). According to one model of epigenetic inheritance, marked parental histones inherited via DNA replication recruit histone modifiers to deposit similar marks on new adjacent nucleosomes, reestablishing the previous local landscape of histone modification (Saxton & Rine, 2019). These findings suggest that histone modifications can aid in epigenetic inheritance, though such a mechanism is normally obscured by H3K9 demethylation activity (Saxton & Rine, 2019).

RNA

One haploid copy of the genome exists in the sperm nucleus, which is transcriptionally silent and not replicated due to the packaging of sperm chromatins during spermiogenesis. Also, the sources of sperm RNAs are unknown. Some RNAs are probably expressed endogenously within germ cells, but considering that short RNAs have been demonstrated to impact gene expression outside of their tissue of origin, sperm RNA could come from elsewhere. Because RNAs are easily degraded, they are most likely transported with protein chaperones and, more likely, enclosed in extracellular vesicles. For example, in epididymosomes, which are extracellular vesicles that are released from the epididymal epithelium contain proteins, noncoding RNAs and a distinct set of lipids that are transferred to spermatozoa while they pass through the different epididymal regions (Sullivan, 2016). Aside from epididymosomes, RNAs have been found in exosomes and high-density lipoproteins (HDLs) in the bloodstream, as well as vesicles secreted by prostate cells (protostomes) and Sertoli cells, which come into direct contact with developing sperm. These interactions have not been studied in the context of transgenerational inheritance. Microinjections of RNA mixtures were used to recapitulate mutant phenotypes in functional tests of RNAs contribution to inheritance patterns. These findings imply that RNA plays a significant role in non-genetic inheritance. It is unclear how these transient RNA molecules can mediate epigenetic inheritance across cell divisions and generations on their own. Nonetheless, a growing body of evidence suggests that RNA may aid in the stable propagation of paternal effects from the soma to the germline, where they will be inherited by the next generation (Andersson et al., 2014; K. Gapp et al., 2020; van Steenwyk et al., 2020).

Non-coding RNA

ncRNAs are broadly classified into two types: 1) long-ncRNA, greater than 200 nucleotides in length (lncRNAs) (Z. Li et al., 2012) and 2) small-ncRNA, less than 200 nucleotides (sncRNA) (Bouckenheimer et al., 2018). Small regulatory RNAs (siRNAs, miRNAs, tRNAs, and piRNAs) are non-coding RNAs that are expressed in the majority of eukaryotic cells. These small RNAs are bound by the Argonaute (AGO) or PIWI proteins and function as specificity factors (guide RNAs) to regulate other homologous RNAs at the transcription, stability, and translation levels. Small RNAs are found in the nuclei of most eukaryotes, where they regulate genome stability, heterochromatin formation, and transcription. ncRNAs are uniquely suited to transmit information across generations. A variety of ncRNA species can be found in both sperm and oocytes (Rando, 2016; Veselovska et al., 2015). There is evidence that sperm can acquire ncRNAs from surrounding somatic tissues (Robles, Valcarce,

& Riesco, 2019). Thus, RNAs can circumvent and penetrate the theoretically impenetrable wall between somatic cells and the germline (the so-called Weismann barrier). This has significant implications for our understanding of epigenetic inheritance as a result of adult-life environmental insults. However, it is currently unknown how epigenetic messages transmitted by RNA are maintained across multiple cell divisions, let alone multiple generations.

0.0.1 Transcription factors

Transcription factors (TFs) may also act as epigenetic information carriers. Transcription factors are proteins that play a role in the process of transcribing DNA to RNA. TFs are found on the genomes of mature gametes, and there is evidence that TF binding can influence DNAm at the binding sites (Jones & Takai, 2001; Smith & Meissner, 2013). CpG sites bound by TFs during germ cell re-methylation are protected from methylation throughout development (Kremsky & Corces, 2020). During PGC and preimplantation development, DNAm can only occur at CpGs that are resistant to demethylation. This is due to the fact that TF binding during the re-methylation phase maintains the methylation status between generations (Kremsky & Corces, 2020). This suggests that transcription factors can act as carriers of epigenetic information during germ cell and pre-implantation development by maintaining CpG methylation status and serve as the foundation for a mechanistic description of the transgenerational transmission of such information in mammals (Kremsky & Corces, 2020). Mammalian gametes contain complex patterns of 3D interactions that can also be transmitted to the zygote after fertilization (Jung et al., 2019). TFs may remain bound to chromatin after fertilization and influence gene expression in the developing embryo. According to a recent publication, BPA, either directly or indirectly, induces the binding of TFs at different regions of the genome in F1 generation sperm, altering the 3D organization of chromatin in mature F1 generation sperm, many of which are conserved in subsequent generations (Jung et al., 2020). Thus, TFs are a plausible candidate for transmitting epigenetic information.

Extracellular vesicles

All cells can secrete various types of membrane vesicles known as extracellular vesicles (EVs). EVs vary greatly in size (30 nm to several micrometres in diameter), chemical composition, secretion mechanism and carrying molecules, all of the properties that lead to their different functions. EVs mediate cell-to-cell communication by activating receptors on the surface of acceptor cells (Bonsergent et al., 2021) and play roles in a variety of physiological and pathological processes, including blood coagulation, inflammation, stem cell expansion,

neuronal communication, and tumorigenesis (Lo Cicero, Stahl, & Raposo, 2015). There are two major EV biogenesis pathways. The first biogenesis pathway buds EVs directly from the plasma membrane, resulting in what is known as microvesicles. The second biogenesis pathway involves intraluminal vesicle release caused by multivesicular endosome fusion with the plasma membrane, resulting in the formation of exosomes. Exosomes and microvesicles contain cargoes such as nucleic acids, proteins and lipids, RNAs (extracellular RNAs, exRNAs) and DNA (O'Brien, Breyne, Ughetto, Laurent, & Breakefield, 2020). The majority of known ncRNA biotypes, including small nuclear RNAs, small nucleolar RNAs (snoRNAs), rRNAs, lncRNAs, PIWI-interacting RNAs (piRNAs), transfer RNAs (tRNAs), mitochondrial RNAs, Y RNAs, and vault RNAs (vtRNAs), have been discovered in EVs in more comprehensive studies (Abels & Breakefield, 2016; O'Brien et al., 2020). ExRNA extraction, purification, and sequencing methods are much more susceptible to contamination and artefacts than cellular RNA preparations, due to the low amount of RNA available from EVs. As a result, analyzing exRNA samples can be difficult, and a specific tool for EV ncRNA data analysis has been developed (Rozowsky et al., 2019). A number of studies have shown the role of extracellular vesicles in epigenetic inheritance. Also, we found that extracellular vesicles are vectors of communication with the germline that play a role in epigenetic inheritance (Alshanbayeva et al., 2021), Chapter 2.

Challenges in data analysis

Multi-omics data integration

A single omics analysis may not provide enough information to provide a deep understanding of a biological system, but combining multiple 'omics experiments and data analyses can provide a more holistic view of a system. Performing different experiments on the same or different samples from the same biological conditions, such as RNA-seq, ATAC-seq, ChIP-seq, and BS-seq, results in multi-dimensional omics datasets, which enable the study of relationships between different biological processes, such as gene expression, chromatin accessibility, histone modifications, transcription factor binding, and DNAm, as well as the leveraging of multiple data types to draw inferences about these processes. Data integration is the process of combining disparate sources of information in order to gain a better understanding of complex systems. The researchers execute two forms of data integration: custom or ad-hoc integration and more systematic integration.

0.0.1.1 Ad-hoc approaches

Integrating BS-seq data with RNA-seq is one example of an ad-hoc technique. The expressed genes are known to be demethylated. As a result, the methylation values of the promoters can be linked to the genes that are expressed. Integrating ATAC-seq data with RNA-seq data can be used to analyze transcription factor activity. Regions with increased accessibility and expression would show a TF's role as an activator, while regions with increased accessibility and decreased expression would represent a TF's role as a repressor. For this, a tool exists, diffTF (Berest et al., 2019), which estimates differential TF activity based on expression and chromatin data. The ad-hoc methods are scientifically motivated and produce biologically relevant outcomes. However, these ad-hoc approaches have several drawbacks, such as the need for extensive exploratory investigation and the inability to be applied consistently to all projects. Furthermore, these methods are frequently applied to a maximum of two omics datasets.

0.0.1.2 Systematic approaches

Systematic methods developed for data integration in biological research can be classified into three types: supervised, unsupervised, and semi-supervised. Unsupervised data integration refers to methods that draw conclusions from input datasets without the use of labelled response variables (e.g., factorization, Bayesian, clustering, dimensionality reduction, coexpression and correlation). On the other hand, supervised data integration methods take into account phenotype labels (e.g., control or treatment) and use, for example, machine learning approaches to evaluate models (network-based and multi-kernel); semi-supervised data integration methods fall somewhere between supervised and unsupervised methods and use both labelled and unlabeled samples to develop a learning algorithm to infer information from unlabeled samples (S. Huang, Chaudhary, & Garmire, 2017). A few methods for integrating data from multivariate multi-omics experiments have been developed: Data Integration Analysis Using Latent Components for Biomarker Discovery (DIABLO) (Rohart, Gautier, Singh, & Lê Cao, 2017) - an extension to sparse generalized canonical correlation analysis (sGCCA) (Tenenhaus et al., 2014), a supervised approach which seeks for common information across different data types through the selection of a subset of molecular features; Multi-Omics Factor Analysis (MOFA) - an unsupervised method for decomposing the sources of heterogeneity in multi-omics data sets (Argelaguet et al., 2018, 2020) and; mixOmics (Rohart et al., 2017) - a package that hosts several supervised methods under its umbrella (DIABLO; sparse Partial Least Square—Discriminant Analysis, PLS-DA (Lê Cao, Boitard, & Besse, 2011) and Multivariate INTEGRative method, MINT (Rohart et al., 2017)).

Despite the fact that these tools have addressed critical challenges for multi-omics applications, they are not without limitations. One of the major limitations of most of these tools is that they can only incorporate data from the same omics experiments. This restricts users to utilize the same samples for all experiments, which is never the case in biological research due to the limited quantity of components of interest (DNA or RNA) and the high demand for these components for library preparation and sequencing. Another issue is that these methods assume that users have already normalized the data and corrected it for batches or confounding factors such as cage effect, sample composition, and time of day (Bruning et al., 2016). Based on the data types, methods should be able to optimally normalize the data and correct it for confounding factors using supervised learning, which could improve uniformity, robustness, and precision. Another disadvantage is that some of the existing methods necessitate exploratory analysis prior to data integration. Exploratory data analysis is a critical process in many data science projects that requires advanced coding skills and takes a significant amount of time. The method should be able to perform automated data cleaning functions, such as outlier clipping, missing value imputation, and zero-variance variables, depending on the data types. It should also aid in exploratory analysis by providing summary statistics and plots for visual examination. Finally, most existing methods are designed to work with linear relationships and thus cannot be applied to more complex datasets that are related in a non-linear manner.

To summarize, integrating diverse and huge omics data is both a conceptual and practical problem in data analysis. Data integration has been used in omics research since the beginning of the era, and it is both a challenge and an opportunity. It is surely growing more popular. We devised and implemented a more straightforward and scientifically motivated approach to integrate multi-omics datasets, involving more than two omics datasets, please see Methods and Chapter 1.

Short RNA data analysis

Short RNAs are a type of non-coding RNA that plays an important role in nearly all aspects of organismal development in both health and disease. Despite their importance in a wide range of biological processes, analyzing short RNA sequencing data remains a difficult task. Specifically, a number of computational challenges arise when analyzing short RNA sequencing:

- Even after size selection, we expect to sequence other RNA biotypes in addition to short RNAs.
- Short RNAs have a length of less than 200bp, often much shorter (e.g. miRNAs at

21-22bp), and can therefore often map to multiple positions. In this case, standard pipelines either map the reads randomly to one position, making it difficult to determine their true origin, or they are discarded, resulting in a large proportion of reads being lost.

- Many short RNAs are post-transcriptionally modified. For example, tRNAs receive a “CCA” sequence at the 3’ end (Hou, 2010), additionally, Histidine tRNAs receive a “G” sequence at the 5’ end (Cooley, Appel, & Soll, 1982), resulting in an unalignment problem.
- Because of their biogenesis, short RNAs often overlap multiple features. For example, mature miRNAs overlap with precursors, which can result in misassignment.
- rRNAs are typically masked in the genome, and the Ensembl (Howe et al., 2021) or GENCODE (Harrow et al., 2012) databases’ annotations are unavailable, which make it difficult to quantify for rRNAs fragments.

All the aforementioned issues could lead to misalignment and low mappability of the sequenced data. Several tools have been developed to analyze short RNA-seq data, however, they typically fall prey to some of these issues. In Chapter 3, we, therefore, developed a novel approach to address them.

Goals of thesis

We defined the following goals for my thesis:

1. Collaboration with all the members of the Mansuy lab to study epigenetic inheritance in the MSUS mouse model. Please see Chapter 3, Appendix A, and Appendix B for this part. I was responsible for:
 - Version controlled and reproducible data analysis for data generated from RNA-seq, sRNA-seq, ATAC-seq, WGBS, and RRBS
 - Determine the qualitative and quantitative effects of early life experiences on the epigenome, particularly in germ cells. This includes identifying the epigenetic changes, including DNAm, RNA expression, non-coding RNAs (ncRNAs), and chromatin accessibility, induced in germ cells due to early life experiences like traumatic stress in mice relating these changes to genome activity. This also includes determining whether and to what extent changes in the epigenome persist across life, and if they can be passed to the offspring and underlie epigenetic

inheritance, and further, whether they can result in changes in DNA sequence across generations

- Supervising and collaborating with colleagues who were performing data analyses themselves
- Analyses of datasets using different tools
- Multi-omics analysis of data from same cells/ tissues
- Figure generation and interpretation for the manuscripts

2. Development of data analysis pipelines. Please see the Methods, for this part. I was responsible for:

- Benchmarking methods for including in the pipeline
- Testing different parameters of the tools for including in the pipeline

3. Collaboration with Dr. Irina Lazar for studying changes in spermatogonial stem cells across development. Please see Chapter 2 for this part. I was responsible for:

- Computational design of the project
- Version controlled and reproducible analysis of the datasets
- Analyzing spermatogonial stem cells dataset generated from RNA-seq, sRNA-seq, RRBS, and ATAC-seq from the data generated in the lab
- Analyzing spermatogonial stem cells dataset generated from RNA-seq, WGBS, and ChIP-seq from the publicly available datasets
- Comparison of published datasets with the lab's
- Analysis of ID4 ATAC-seq data from the literature and comparison with lab's dataset
- Multi-omics data integration
- Figures outline, figures generation, and interpretation for the manuscript
- Writing methods, figure legends, and updating/ revising manuscript

4. Collaboration with Dr. Pierre-Luc Germain for the development of shortRNA, a tool for comprehensive analysis of small RNA-seq data in R. Please see Chapter 4 for this part. I was responsible for:

- Development and testing of the tool
- Data analysis using different methods for qualitative and quantitative comparison with shortRNA

ep•i•genet•ics: (*n*) ...

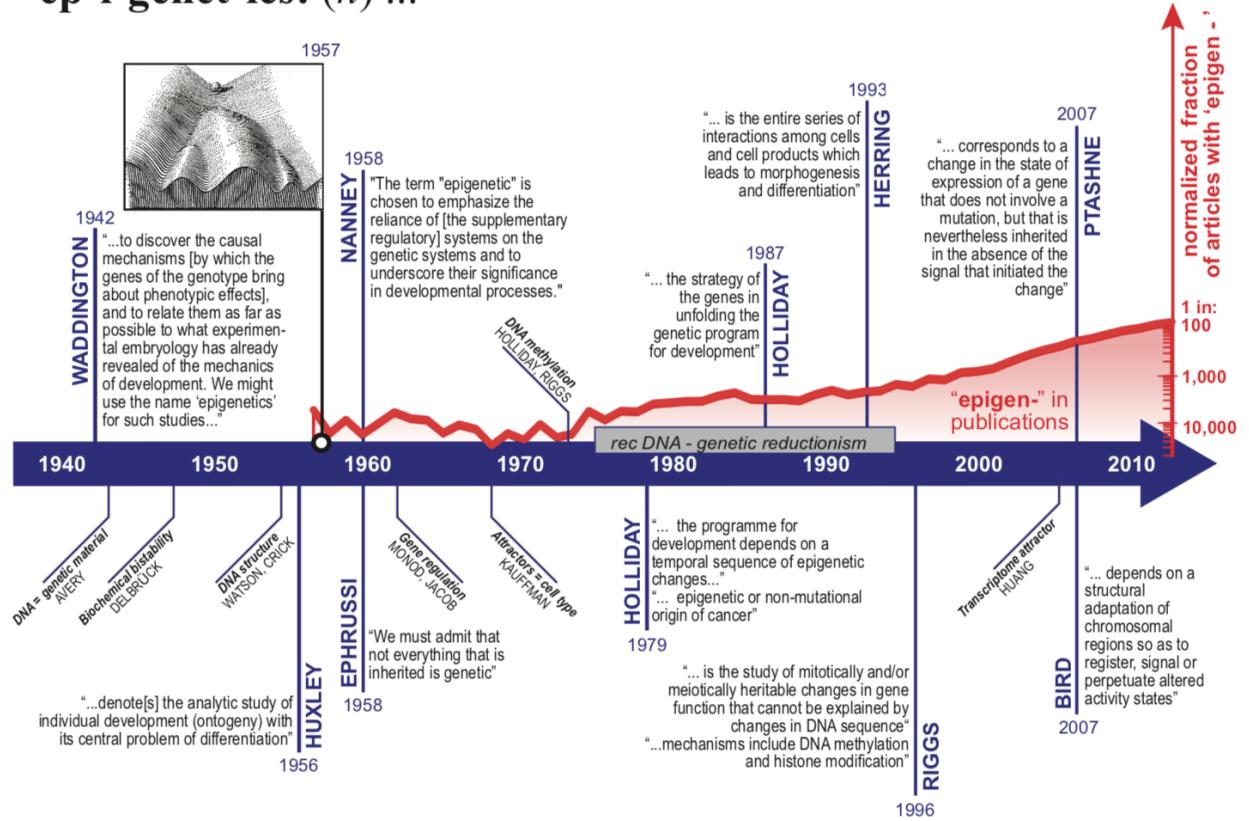


Figure 1: History of the term 'epigenetics'. The discovery of DNA methylation launched the associated modern (molecular) notion of 'epigenetics' (1975) and is reflected in the surge of the number of publications after that. Note the stagnation in the two decades after that DNA recombinant technology and genetic determinism dominated mindset in biology (grey horizontal bar). We are now seeing an enormous increase in the use of the term, as the post-genome era findings challenge the textbook gene-centred mindset. Textboxes display quotes, of selected authors (vertical text). The red curve indicates the number of Medline entries containing the wildcarded search term 'epigen*' per year, normalized for a number of publications that contain the generic keywords 'cell' and 'function'. Diagonal text boxes highlight key discoveries or findings in biology influencing the history of epigenetics but not directly related to it. Taken from (Susiarjo, 2016).

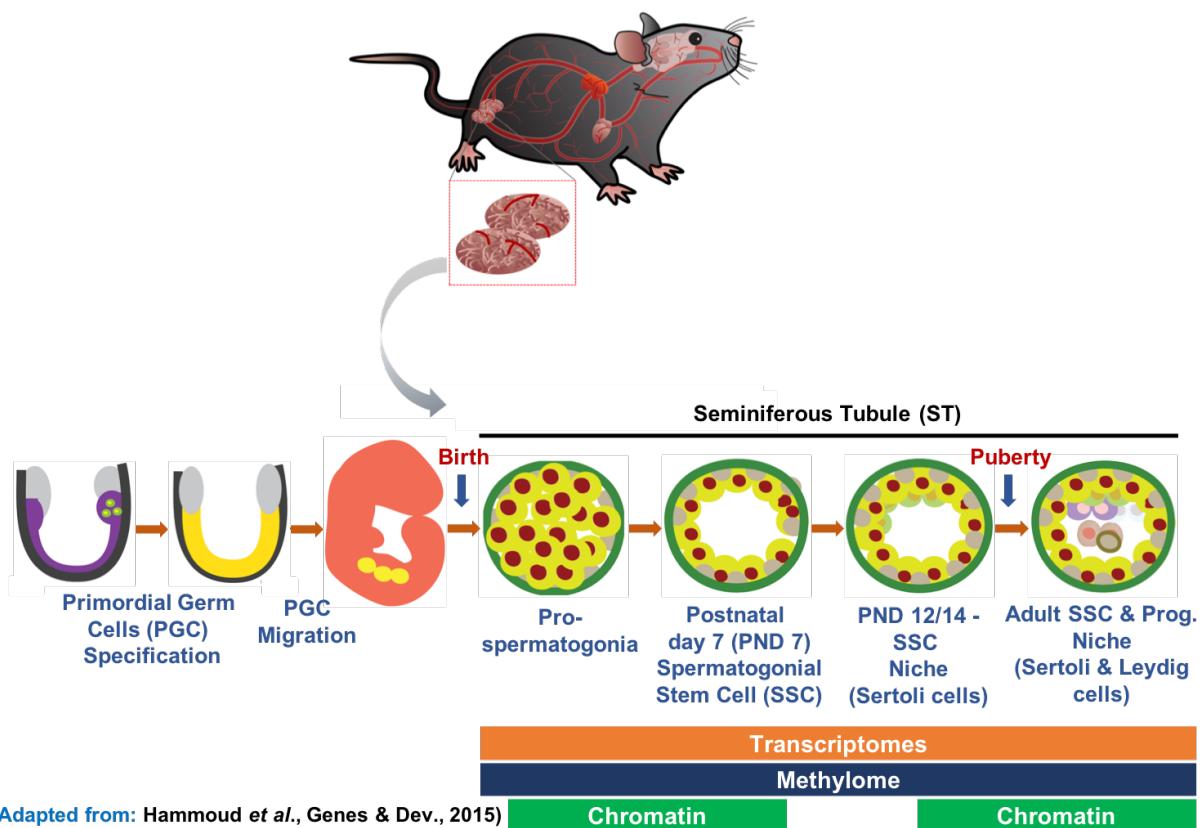


Figure 2: Graphical summary of the biology of germline stem cell specification and transitions. Adapted from (Hammoud et al., 2015)

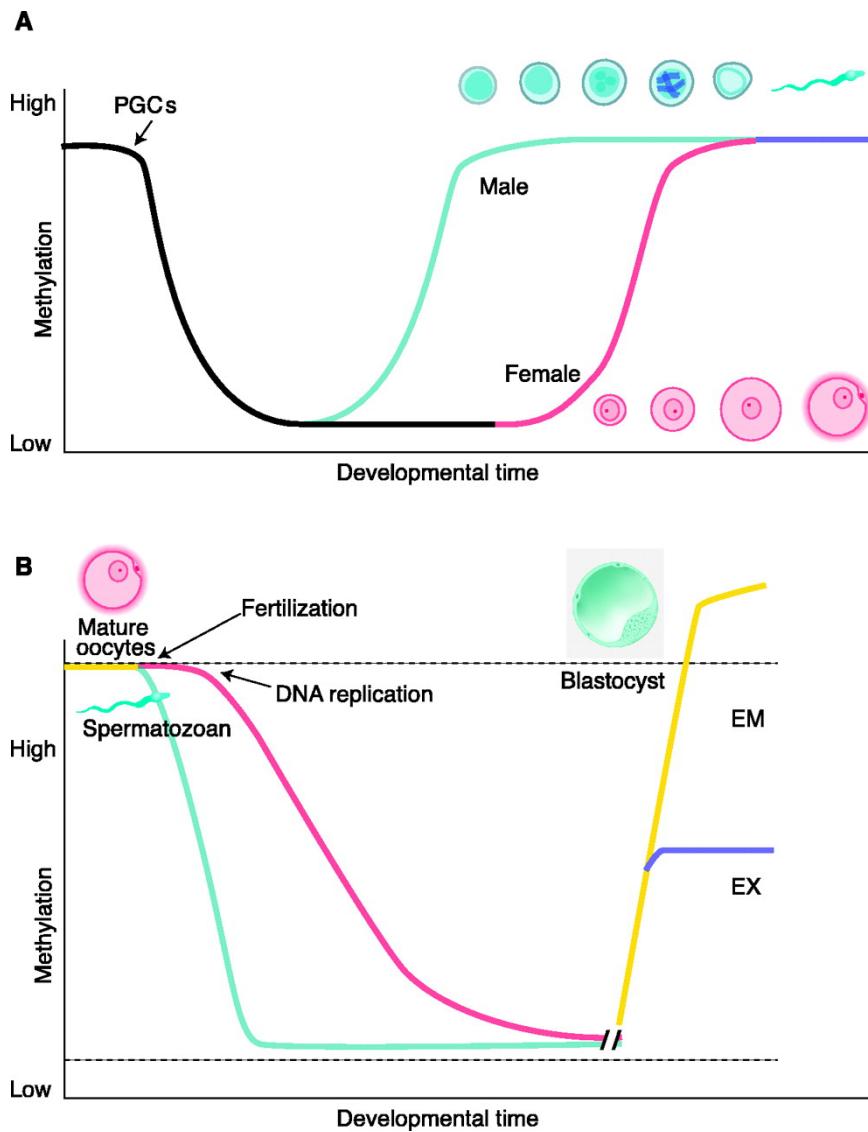


Figure 3: Methylation reprogramming in the germline.

A: Primordial germ cells (PGCs) in the mouse become demethylated early in development. Remethylation begins in prospermatogonia on E16 in male germ cells, and after birth in growing oocytes. Some stages of germ cell development are shown.

B: Methylation reprogramming in preimplantation embryos. The paternal genome (blue) is demethylated by an active mechanism immediately after fertilization. The maternal genome (red) is demethylated by a passive mechanism that depends on DNA replication. Both are remethylated around the time of implantation to different extents in embryonic (EM) and extraembryonic (EX) lineages. Methylated imprinted genes and some repeat sequences (dashed line) do not become demethylated. Unmethylated imprinted genes (dashed line) do not become methylated. Taken from (Reik et al., 2001).

Bioinformatics methods

Version controlled data analysis using git

Version control (also known as source control or revision control) is a software engineering process that involves managing changes to a computer program or document using a version control system (VCS). In graph theory, revisions form a directed acyclic graph that represents the growth line. The revisions to a computer program occur over time and can be tracked using revision numbers, timestamps, or descriptive text. While file copies have been used in the past to provide some form of version control, they are error-prone and time-consuming. The file is typically copied to another location, and unless the files are cleverly time-stamped, this simple approach may be error-prone.

Programmers developed a local VCS with a database that records all modifications to the files under revision control to keep track of the versions. A local VCS is an example of a revision control system (RCS). Since it is difficult to collaborate with a local VCS, a centralized VCS was developed, such as Apache Subversion (SVN). All changes to files are monitored under the centralized server using Centralized VCS. A centralized VCS, on the other hand, has significant drawbacks. No one can connect or save versioned changes to something they are working on if the server goes down for an hour. This leads to the development of the distributed VCS, such as git. In a distributed VCS, the repository can be completely mirrored, including the entire version history. A schematic of distributed VCS is shown in Figure 4, adapted from (Chacon & Straub, 2021).

GitHub (github.com) a distributed version control system, was used to manage data analysis projects and software development projects throughout this doctoral thesis work. One downside of using GitHub is that files greater than 50 megabytes cannot be hosted, which means that large files such as sequencing data, FASTQ files, and BAM files are not version managed. To solve this issue, git-fat was used, a version-controlled tool that offloads files to a local directory or another server (Brown, 2018).

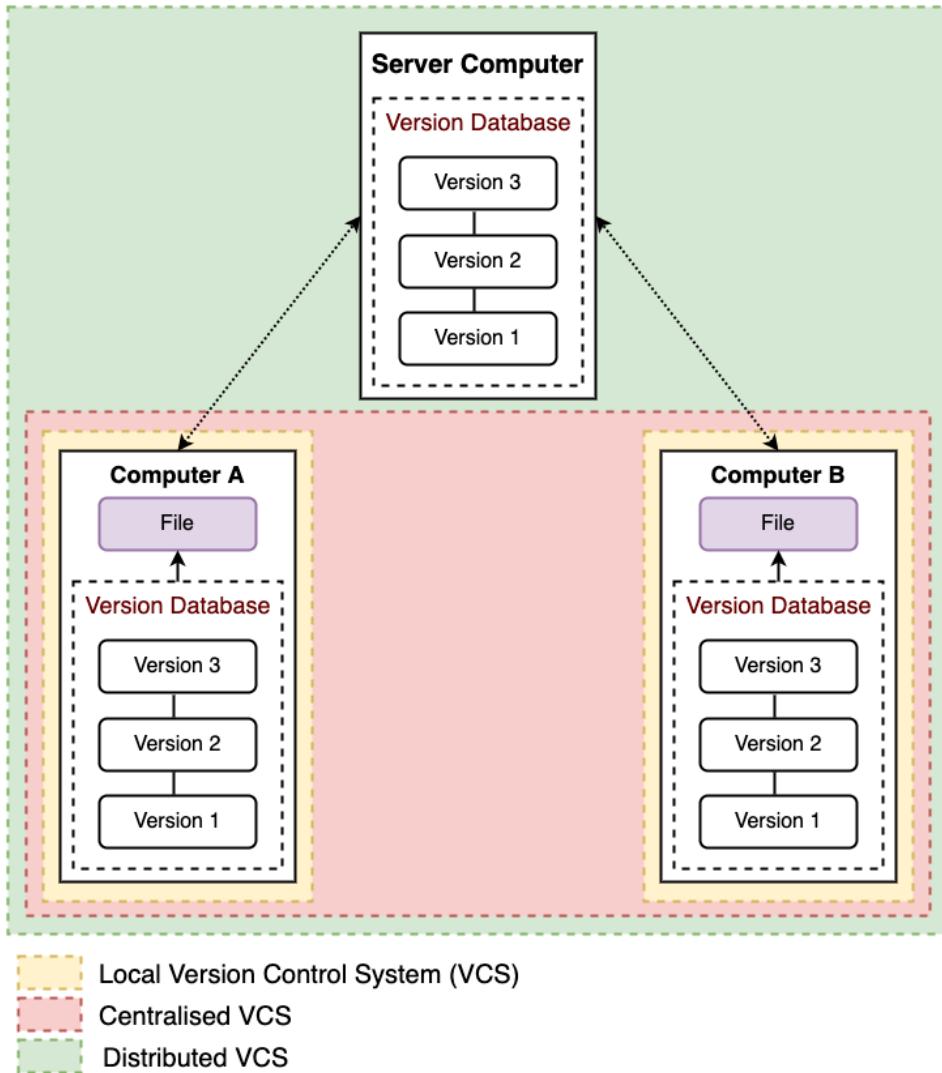


Figure 4: Overview of version control system (VCS). Adapted from (Chacon & Straub, 2021).

Data analysis directory organization

ScienceCloud service of the University of Zurich was used for computing. On ScienceCloud, a data volume is mounted on my instance (virtual computer). In the data volume, two directories were created, namely “Data” and “Project”. The “Data” repository stores raw sequencing files, FASTQ files, and “gitfat” objects, while the Project directory stores all analysis and code and is version controlled with git. A schematic representation of this workflow can be seen in Figure 5. This is done for each dataset analysed.

This organization of the data analysis projects ensures open and reproducible research. As all the analyses are version controlled, one can look at the analysis conducted back in time. Also, each analysis script has a log file, which stores the information of the tools version and

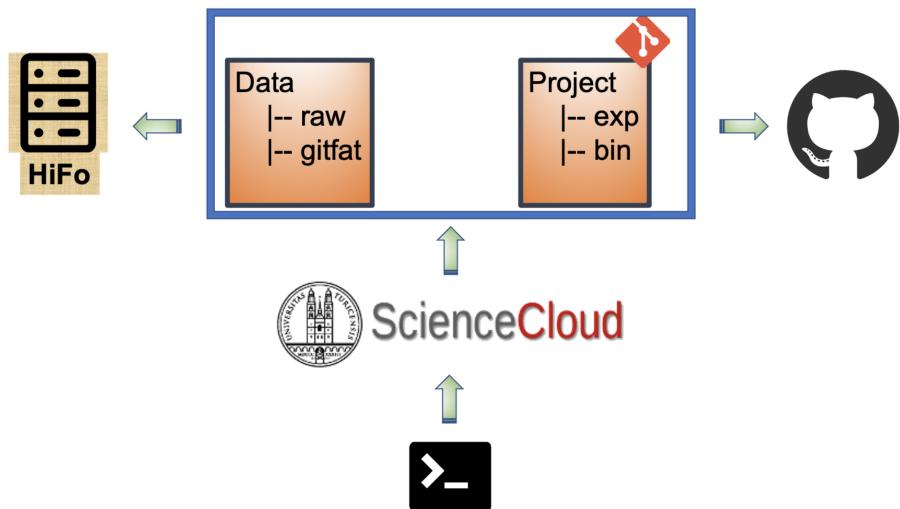


Figure 5: Organization of data analysis projects.

any warnings or errors encountered during the analyses. Further, at the end of the project, the GitHub repository can be made public and all the code for intermediate data analysis are available to the research community.

Pipelines for data analysis

I worked on multi-omics datasets for my doctoral research, which led to the development of data analysis pipelines for RNA sequencing (RNA-seq), Assay of Transposase Accessible Chromatin sequencing (ATAC-seq), small-RNA-seq, Whole-genome bisulfite sequencing (WGBS), Reduced representation bisulfite sequencing (RRBS), Chromatin immunoprecipitation followed by sequencing (ChIP-seq), and short RNA sequencing (sRNA-seq).

Quality check and trimming of sequencing datasets

The goal of data quality check and trimming is to find and clean any data quality issues in the dataset and possibly account for them. The sequencer determines the nucleotide bases in a DNA or RNA library during sequencing. A small sequence, known as a read, is created for each fragment in the library, which is a sequence of nucleotides. In a single experiment, sequencing technologies, such as Illumina, can generate a large number of sequence reads. The first step after receiving the sequenced data is to assess the data quality to determine the reliability of the sequenced reads. The FASTQ files are examined using FastQC (Andrews et al., 2012). FastQC provides a comprehensive view of the data, including the number of

sequenced reads, base sequence quality, “N” content (if a sequencer cannot call a base, due to the technical limitations, it enters “N”; of each sequencing platform.), sequence duplication levels, and adapter information. A Hypertext Markup Language (HTML) report generated by the software, for each sample, helps to understand the pre-processing steps required. The quality assessment is done using the quality scores (Q-scores) in the FASTQ files, which are translated to quality statistics in FastQC reports and plotted as boxplots. Q score is defined as the base-calling error probabilities and is calculated by the formula: $Q = -\log_{10} (P)$. For instance, if a nucleotide base is assigned a Q score of 30, this is equivalent to the probability of an incorrect base call of 1/1000 times, which means that the accuracy of base calling is 99.9%. A Q score of 30 represents perfect base calling with no errors and ambiguities and is considered a benchmark for quality in high-throughput sequencing (HTS) (Ewing & Green, 1998).

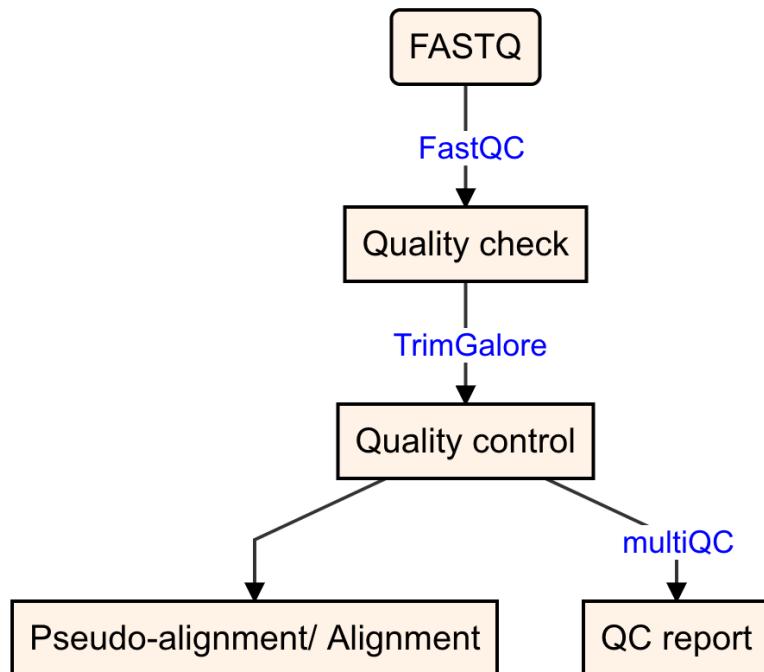


Figure 6: Workflow for quality check and quality control for sequencing data. QC: Quality Control.

Adaptors are artificial DNA oligonucleotides and are required for sequencing by platforms like Illumina. During the process of Illumina library preparation, adaptors are ligated to the short DNA sequences. Because of the adapters (attached at the end of reads), the sequencing Q score is usually low. Therefore, these are required to be removed before downstream processing of the data. Furthermore, the sequenced DNA reads could have trailing and leading ‘N’ (if a base caller fails to call a base at a genomic location, it will put ‘N’ instead of ‘A’, ‘T’, ‘G’ or ‘C’). Also, there is a chance that although long reads were sequenced, small

reads appeared after sequencing, or they became shorter after removal of trailing or leading ‘N’. It is always better to get rid of small reads below a certain length, as they would be mapped to multiple locations and will lead to misinterpretation. Trimming is the process of modifying the ends of reads. Trimming can help to increase the number of reads that the aligner or assembler can successfully use, lowering the amount of unmapped or unassembled reads. TrimGalore is used to trim adapters and inferior ends of reads with a Phred score of less than 30 (-q 30), to remove Ns from both sides of reads (-trim-n), and to discard trimmed reads that are too short to be informative (e.g. -length 30, or 15 for small RNA-seq data). After this, all results are used to generate an overview report with multiQC (Ewels, Magnusson, Lundin, & Käller, 2016), which is a reporting tool that parses summary statistics from results of other bioinformatics data analysis tools. Finally, it is critical to recognize, identify, and rule out issues that may affect downstream analysis interpretation.

The trimmed data is then aligned to the reference genome, indexed with the specific tool needed for alignment. For example, for aligning WGBS data with the reference genome, the reference genome is first indexed with Bismark. This workflow is shown in Figure 6.

Some library preparation kits use the unique molecular identifier (UMI), which are complex indices and are added to the sequencing libraries before the polymerase chain reaction (PCR) amplification step. UMIs were first implemented in the iCLIP protocol (König et al., 2010) but has been implemented for RNA-seq, small RNA-seq, single-cell sequencing, ChIP-seq, and WGS.

RNA-seq

RNA sequencing (RNA-seq) is a technique that uses high-throughput sequencing (HTS) to investigate the quantity and sequences of RNA. It examines the transcriptome to determine, which of our DNA-encoded genes are turned on or off, and to what degree. RNA-seq pipeline schematic is shown in Figure 7.

Pseudo-alignment and alignment

Sequence alignment is used to determine where the sequences are similar to the reference genome and how similar they are. We can estimate where a read came from by aligning or “mapping” it to a reference genome or transcriptome. There are two families of methods for mapping reads in RNA-seq data: alignment and pseudo-alignment. For alignment, aligners such as STAR or Rsubread are used. They align reads to a genome or transcriptome and perform spliced alignment. They return the reads’ base-level alignments. Transcript quantification tools such as Salmon (Patro, Duggal, Love, Irizarry, & Kingsford, 2017) and

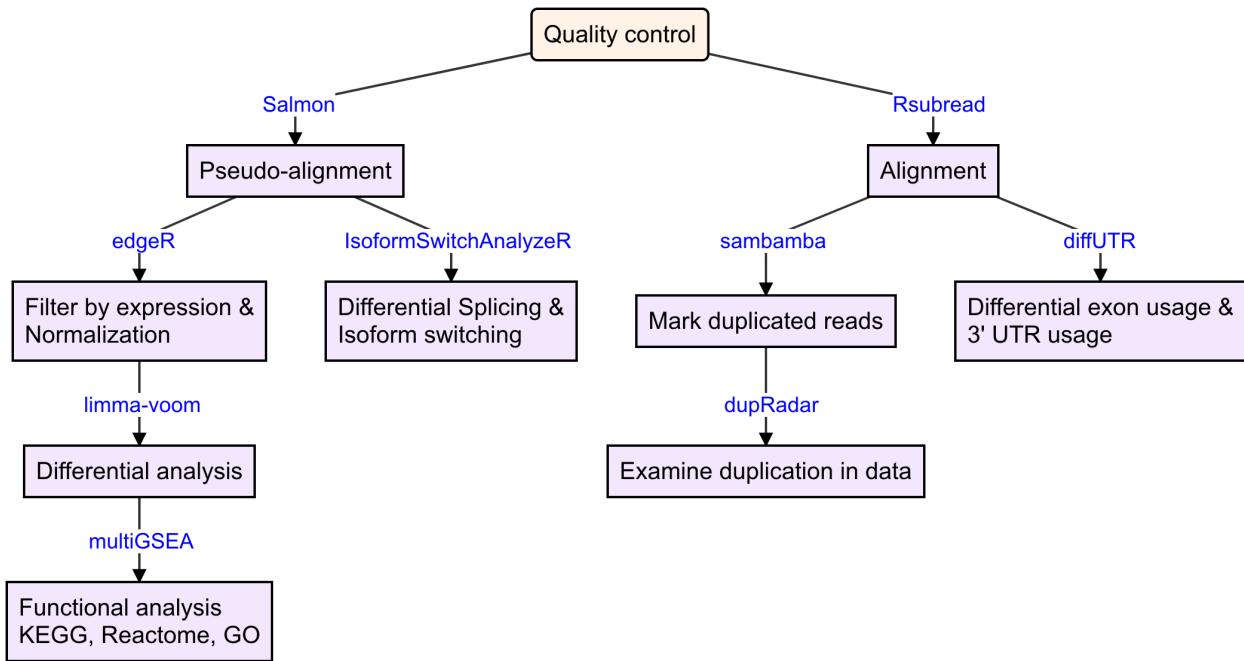


Figure 7: Workflow for RNA sequencing data analysis.

Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016) instead perform pseudo-alignment. These tools use statistical inference to determine transcript abundances by mapping reads to the transcriptome. Alignments from the aligners can then be used for a variety of purposes, such as for feeding transcript assembly tools, for variant calling pipelines, or for transcript quantification tools.

Quality-controlled reads were pseudo-aligned using Salmon with automatic detection of the library type (-l A), correcting for sequence-specific bias (-seqBias), and correcting for fragment GC bias correction (-gcBias) on a transcript index prepared from GENCODE, with additional piRNA precursors and transposable elements (concatenated by family) from Repeat Masker as in (K. Gapp et al., 2020). Downstream analyses including differential analysis, section ____; differential transcript usage and differential isoform switching, section ____; and functional analyses, section ____, are performed on the pseudo-alignment from Salmon.

Additionally, alignment is performed using the Rsubread package (Liao, Smyth, & Shi, 2019) using the subjunc function, specific for RNA-seq data. Analysis of duplicates; section ____; differential exon usage, section ____; and differential 3' usage, section ____, are performed on alignment obtained from Rsubread.

A direct comparison of genome alignment with transcriptome pseudo-alignment was performed in (Yi, Liu, Melsted, & Pachter, 2018) and the authors found that both approaches produce similar quantifications. However, (D. C. Wu, Yao, Ho, Lambowitz, & Wilke, 2018) suggested that it is not optimal to use alignment-free methods to analyze and quantify lowly

expressed genes and small RNAs.

Duplicated reads analysis

Duplicated reads are different copies of the exact same sequence. Unless a genomic region is highly expressed, most mRNA fragments are expected to be unique. A low degree of repetition may indicate that the target sequence is well-covered, whereas a high level of duplication is more likely to signal bias. Duplicate reads might be caused by PCR duplication or legitimately overrepresented sequences. Because PCR amplification is more efficient for some sequences than others, it can misrepresent the true proportion of sequences in the input, whereas truly overrepresented sequences are the result of very abundant transcripts in an RNA-Seq library and are an expected case and not of concern because they accurately represent the input.

Throughout the research, duplicated reads were marked in the aligned data using sam-bamba (Tarasov, Vilella, Cuppen, Nijman, & Prins, 2015), and further, the quality issues with PCR duplicates are analyzed using dupRadar (Sayols, Scherzinger, & Klein, 2016). dupRadar aids in separating the fraction of readings produced by artefacts from the fraction resulting from high expression.

Differential expression analysis

Differential expression analysis (at the transcript or gene level) is the statistical examination of normalised read count data to find quantitative differences in expression levels between experimental groups.

Read counts at the transcript level are obtained from pseudo-aligned data and are aggregated at the gene level. Low counts of candidate genes/transcripts across samples give minimal evidence for differential expression. When evaluating false discovery rates, they also add to the multiple testing burden, lowering the ability to find differentially expressed genes. A pre-filtering of counts is therefore performed using the filterByExpr function from edgeR (Robinson, McCarthy, & Smyth, 2010) with a design matrix (matrix containing data about multiple characteristics of several samples, such as biological group, age, and batch for library preparation) and requiring at least 20 counts (min.counts = 20) (Germain, Sonrel, & Robinson, 2020). Next, the counts are normalized for different sequencing depths between samples and to eliminate composition biases between samples, for example, if there are a few highly expressed genes dominating in some samples, leading to fewer reads from other genes. Normalization factors are obtained using the TMM normalization method (Robinson et al., 2010) from the edgeR package, using the calcNormFactors function. TMM stands for

Trimmed Mean of M values, in which the counts for the samples are scaled using a weighted trimmed mean of the log expression ratios. Differential expression analysis at gene and transcript level is performed using limma-voom (Law, Chen, Shi, & Smyth, 2014) pipeline from limma (Ritchie et al., 2015) package.

Exploratory data analysis

Exploratory data analysis of counts (using principal component analysis, PCA), differential analysis results (volcano plot, MA plot, and heatmap), and assessment of library composition is performed using plgINS, a versatile R package, (github.com/ETHZ-INS/plgINS), under development by Dr. Pierre-Luc Germain. The package also comes with a shiny application.

Differential transcript usage and differential isoform switching

Differential transcript usage (DTU) analysis looks for proportional differences in the makeup of a gene's transcripts between conditions.

DTU is performed using IsoformSwitchAnalyzeR (Vitting-Seerup & Sandelin, 2017, 2019). Isoforms are annotated via integration of a wide range of (predicted) annotations: Pfam was used for prediction of protein domains, CPC2 (version 2.0) for calculation of the coding potential, SignalP (version 5.0) for prediction of Signal Peptides, and IUPred2A was used to predict Intrinsically Disordered Region (IDRs) and Intrinsically Disordered Binding Regions (IDBRs). The results from Pfam, CPC2, IUPred2A, and SignalP are used with IsoformSwitchAnalyzeR for annotation and prediction of functional consequences for the identified isoform changes.

Differential exon usage and differential 3' untranslated regions usage

Differential exon usage (DEU) analysis looks for variations in exon usage between experimental conditions. mRNAs' 3' untranslated regions (3' UTRs) are best recognized for regulating mRNA-based functions such as mRNA localization, stability, and translation (Mayr, 2019). Differential UTR usage looks for variations in 3' UTR usage between experimental conditions.

DEU and differential 3' UTR usage are performed using the diffUTR R package (Gerber, Schratt, & Germain, 2021).

Functional analysis

The output of RNA-seq differential expression analysis is a list of significant differentially expressed genes (DEGs). The goal of functional analysis is to provide biological insight of DEGs. Functional analysis on Kyoto Encyclopedia of Genes and Genomes (KEGG)

(Kanehisa, Furumichi, Sato, Ishiguro-Watanabe, & Tanabe, 2021), Reactome (Jassal et al., 2020), Gene Ontology (GO) (Consortium, 2021) and curated pathway list from the Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) databases is performed using multiGSEA (github.com/lianos/multiGSEA), which facilitates the analysis with GOseq (Young, Wakefield, Smyth, & Oshlack, 2010), fGSEA (Korotkevich et al., 2016), Correlation Adjusted MEan RAnk gene set test (CAMERA) and camera pre-ranked (CAMERA-PR) (D. Wu & Smyth, 2012), ROtation testing using MEan Ranks (ROMER) and fry (Ritchie et al., 2015), and ROtAtion gene Set Testing (ROAST) (D. Wu et al., 2010). An advantage of using multiGSEA is that it has standardized input and output and provides a Shiny app for exploratory data analysis (github.com/lianos/multiGSEA.shiny). With multiGSEA, it is easy to specify to use testing relative to a threshold (TREAT) method (McCarthy & Smyth, 2009), which takes a user-specified log fold change cut-off and recalculates the moderated t-statistics and p-values. Using TREAT only affects enrichment tests that first threshold the genes in the experiment as “significant” or not, like GOseq and not tests like the camera.

ATAC-seq

The Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) is widely used in studying chromatin biology to study chromatin accessibility and to identify open chromatin regions (OCRs). The nucleosome is the most basic component of chromatin. Each nucleosome is made up of about two turns of DNA wrapped around a set of eight proteins called histones known as histone octamer. With the help of nucleosomes, the genome is securely packed and structured, known as chromatin. The organization and accessibility of DNA are influenced by a number of factors, including chromatin structure, nucleosome location, transcription factors, and histone modifications. As a result, these factors have a role in the activation and inactivation of genes. The genome is treated with a hyperactive version of the Tn5 transposase to insert sequencing adapters into OCRs (Buenrostro, Wu, Chang, & Greenleaf, 2015). Because it is easier, faster, and uses fewer cells than competing approaches like FAIRE-Seq and DNase-Seq, ATAC-Seq has become popular for discovering accessible regions of the genome. The schematic of the ATAC-seq pipeline is shown in Figure 8.

Alignment, post-alignment quality control, and peak calling

Alignment of the ATAC-seq data is performed using Bowtie2 (Langmead & Salzberg, 2012) with the following parameters: fragments up to 2 kb were allowed to align (-X 2000), entire read alignment (-end-to-end), suppressing unpaired alignments for paired reads (-no-mixed), suppressing discordant alignments for paired reads (-no-discordant) and minimum acceptable

alignment score with respect to the read length (`-score-min L,-0.4,-0.4`). Tn5 inserts adapters separated by 9bp when it cuts an accessible chromatin region (Kia et al., 2017). This indicates that reads on the positive strand should be adjusted 4 bp to the right and reads on the negative strands should be shifted 5 bp to the left in order for the read start site to reflect the centre of where Tn5 binds as done in (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013), which is important for the footprint analysis. Using alignmentSieve from deepTools (Ramirez et al., 2016), aligned data (BAM files) are adjusted for the read start sites to represent the centre of the transposon cutting event (`-ATACshift`), and filtered for reads with a high mapping quality (`-minMappingQuality 30`). In addition, because the mitochondrial genome is devoid of chromatin packaging (Bogenhagen, 2012) a large number of mitochondrial reads can be a concern in ATAC-seq. Some ATAC-Seq samples have been found to include 80% mitochondrial reads, prompting the development of wet-lab approaches to address the problem (Corces et al., 2017). Also, because the open chromatin regions of interest are typically found in the nuclear genome, mitochondrial reads are typically excluded from the analysis. Hence, reads mapping to the mitochondrial chromosome and ENCODE blacklisted regions (regions where genome assembly results in erroneous signal) were filtered using alignmentSieve. Because ATAC-seq does not require rigorous size selection during library preparation, it can identify nucleosome positions using fragments representing nucleosome-free fragments (NFFs) nucleosome monomers, and multimers. There is approximately 147 bp of DNA wrapped around a nucleosome, and in order to obtain NFFs, which are indicative of transcription factor bindings, the fragments less than 147 bp must be extracted. To extract NFFs, all aligned files are merged within experimental groups, reads are sorted by left-most coordinates and indexed using SAMtools (H. Li et al., 2009), and NFFs were obtained by selecting alignments with a template length between 40 and 140 inclusively (Goodnight et al., 2019; Jung et al., 2017). Peak calling is a statistical approach that leverages data coverage properties to identify locations that are enriched as a result of protein/ transcription factor binding. To find regions corresponding to potential OCRs in the nucleosome-free regions, we want to identify regions where reads have piled up greater than the background read coverage. Peak calling on the NFFs is performed using MACS2 (Y. Zhang et al., 2008) with genome size in base-pair (`-g genomeSize`) and specifying the paired-end BAM file format (`-f BAMPE`).

Peaks annotation and differential accessibility analysis

The goal of peak annotation is to map peaks to the regulatory elements. The peaks are annotated with transcript, and the distance to the nearest transcription start site, based on overlap with gene transfer format (GTF) file obtained from the GENCODE. The number of extended reads overlapping in the peak regions is calculated using the csaW package (Lun

& Smyth, 2015). Peak regions that did not have at least 15 reads in at least 40% of the samples were filtered out. Normalization factors were obtained on the filtered peak regions using the TMM normalization method (Robinson et al., 2010) and differential accessibility analysis is performed using the Genewise Negative Binomial Generalized Linear Models with Quasi-likelihood (*glmQLFit*) test from the *edgeR* package. *edgeR* has recently been benchmarked and recommended for differential chromatin accessibility analysis (Gontarz et al., 2020).

Functional analysis

GO analysis is performed on differentially accessible regions (DARs) with the rGREAT package (github.com/jokergoo/rGREAT), which is a wrapper around the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010). Transcription factor motif enrichment analysis is performed using the *marge* package (Amezquita, 2018), which is a wrapper around the Homer tool (Heinz et al., 2010).

Differential accessibility analysis at transposable elements

Transposable elements (TEs) GTF file is obtained from http://labshare.cshl.edu/shares/mhammelllab/www/data/TEtranscripts/TE_GTF/mm10_rmsk_TE.gtf.gz. The GTF file provides hierarchical information about TEs: **Class** (level 1, eg. LTR), **Family** (level 2, eg. LTR L1), **Subtype** (level 3, eg. LTR L1 L1_Rod), and **Locus** (level 4, eg. LTR L1 L1_Rod L1_Rod_dup1). TE loci are annotated based on overlap with GENCODE as described above for ATAC-seq peaks. Filtered BAM files (without reads mapping to blacklisted or mitochondrial regions) are used for analyzing TEs. Mapped reads are assigned to TEs using featureCounts from the R package *Rsubread* (Liao et al., 2019) and were summarized to Subtypes (level 3), allowing for multi-overlap with fractional counts, while ignoring duplicates. The number of extended reads overlapping at the TE loci is obtained using the *csaw* package (Lun & Smyth, 2015). Subtypes that did not have at least 15 reads, and loci that did not have at least 5 reads in at least 40% of the samples, were filtered out. Normalization and differential accessibility analysis is performed as described above for the peaks. GO and motif enrichment analysis is performed as described above for the peak regions.

ChIP-seq

A substantial part of gene expression regulation is controlled by protein-DNA interactions. Proteins like transcription factors and histones influence how much and where genes are expressed. ChIP, followed by sequencing, can be used to investigate the interaction between a

DNA sequence and a protein. There are two types of ChIP: cross-linked ChIP (X-ChIP) and native ChIP (N-ChIP). Formaldehyde is utilized to crosslink histones to DNA, and ultrasound sonication is used to fragment the DNA in X-ChIP, whereas the native covalent interaction between protein and DNA is exploited in N-ChIP (Park, 2009). The fundamental purpose of ChIP is to map global binding sites for any protein of interest, such as transcription factors (TF) and histones, in order to investigate specific modifications. *In vivo*, the approach generates a library of DNA locations bound to the protein of interest. Following ChIP on samples, library preparation, PCR amplification, and ultimately deep sequencing are performed. The schematic of the ChIP-seq pipeline is shown in Figure 9.

Alignment and peak calling

Alignment is performed using Bowtie2. Reads with more than 3 mismatches are removed from the aligned data, as suggested in (Royo, Stadler, & Peters, 2016), and reads with low mapping quality (`-minMappingQuality 30`) or mapping to the mitochondrial chromosome or aforementioned blacklisted regions are also filtered out. Peak calling is performed using MACS2 with mouse genome size in base-pair (`-g genomeSize`) and the BAM file format (`-f BAM/PEBAM`), together with the ChIP input sample.

Peaks annotation, differential analysis, and functional analysis

These analyses are performed as described above for the ATAC-seq peaks.

Bisulfite sequencing

DNA methylation (DNAm) involves the addition of methyl groups to the DNA molecule. It most normally occurs at the C5 position of cytosines inside CpG dinucleotides, where it is stable during mitosis and meiosis. Methylation can change the activity of DNA without changing its sequence. DNAm is a reversible, cell-type-specific DNA modification that is usually persistent through cell division. It is not equally distributed across the genome, but rather is linked to CpG density. High-throughput bisulfite sequencing is one of the most reliable methods for measuring DNA methylation. In bisulfite conversion, unmethylated cytosine residues are converted to uracil, but methylated cytosines remain intact (Frommer et al., 1992), allowing methylation and unmethylated cytosine bases to be distinguished. Following PCR amplification, these uracils are converted to thymines. This is followed by the WGBS. WGBS is considered a gold-standard technology for DNA methylation detection as it provides a qualitative, quantitative, and efficient technique to identify 5-methylcytosine at single base-pair resolution. The cost of sequencing an entire genome is typically quite high.

As a result, despite its use on huge genomes like the human genome, significant numbers of individual samples are rarely sequenced. For this reason, RRBS was developed, in which the bisulfite process occurs but only around 1% of the genome is sequenced. This makes it possible to sequence the genomes in greater numbers. The idea behind RRBS is to enrich the sequencing library with CpG-dense sections of the genome to boost the sequencing depth of cytosine-rich areas (Meissner et al., 2005). This is done by using CpG-specific restriction enzymes like MspI, which identifies the sequence 5'-CCGG-3'. MspI cleaves the complete genome in RRBS, leaving at least two cytosines per cleaved DNA fragment (one cytosine at either end of the MspI-MspI fragment). Sequencing libraries are made from size-selected MspI DNA fragments, followed by bisulfite conversion and sequencing of the converted library. The schematic of the BS-seq pipeline is shown in Figure 9.

Alignment and methylation calling

Mismatches between the reads and the reference genome are introduced during bisulfite conversion, resulting in slow and imprecise mapping. Further, the DNA code's complexity is reduced. Hence, specialized tools are required for aligning the bisulfite sequencing data. Alignment of the bisulfite sequencing data is performed using Bismark (Krueger & Andrews, 2011) on a genome index built using `bismark_genome_preparation`. Methylation information for individual cytosines (methylation calls) is extracted using the `bismark_methylation_extractor` tool from the Bismark package.

Differential methylation analysis

Differential analysis can be performed at the loci level (differentially methylated loci, DML) or at the region's level (differentially methylated regions, DMRs). For the identification of DML and DMRs, the data is first filtered by the coverage among samples. DML are identified using edgeR. Identification of DMRs is not so straightforward, as defining regions boundaries is not easy the tool has to take into account correlation across loci of a region. Hence, during a course, we benchmarked different tools for detecting DMRs dktanwar.github.io/sta426-project-dmr-comparison and found dmrseq to outperform all other methods that we compared. DMRs from RRBS data are identified using the parameters recommended by Dr. Keegan Korthauer, author of dmrseq (github.com/kdkorthauer/dmrseq/issues/14).

CNV analysis

The phenomenon of copy number variation (CNV) occurs when regions of the genome are repeated and the amount of repeats differs between individuals. CNV is a type of structural

variation that occurs when a duplication or deletion event occurs that impacts a large number of base pairs. To infer and visualize CNV from WGBS data, we use the CNVkit (Talevich, Shain, Botton, & Bastian, 2016).

DML and DMR annotation, and functional analysis

These analyses are performed as described above for the ATAC-seq data.

Short RNA-seq

The small RNA-seq pipeline is not described in this section, as a method for analyzing small RNA-seq data has been developed and is discussed in detail in Chapter 3.

Multi-omics analysis

Using different experiments, followed by sequencing, on the same samples or cells/tissues, such as RNA-seq, ATAC-seq, ChIP-seq, and BS-seq, results in multi-dimensional omics datasets that can be used to investigate relationships between different biological processes, such as gene expression, chromatin accessibility, histone modifications, and DNA methylation. Integrating multi-omics datasets is most challenging and is not straightforward to perform, as there are no standard recipes.

We present a straightforward and biologically useful approach for performing multi-omics analysis on RNA-seq, ATAC-seq, ChIP-seq, and BS-seq datasets. First, we overlap the differentially accessible regions (DARs) with ChIP-seq peaks enrichment, RNA-seq expression of overlapping genes with DARs, and methylation values of DARs overlapping loci from BS-seq data into one matrix, overlapMatrix. The DARs are initially divided into proximal (less than +/- 2.5 kb from a TSS) and distal (more than +/- 2.5 kb from a TSS) regions, as per the ENCODE guidelines (Harrow et al., 2012; Thurman et al., 2012); the proximal regions are further divided into active and inactive (proximal active and proximal inactive) groups based on nearby gene expression (if nearby gene expression is detected) (**step 1**). Next, these three groups are separated based on whether the biological groups are more or less accessible (**step 2**). This classification would reveal six separate clusters of DARs, each with its own biological profile. Further, proximally active genes are categorised based on their up/ down expression (**step 3**). Moreover, we categorised DARs depending on whether they had H3K4me3, which is a critical marker for the active transcription and developmentally stable genes (L. Fu & Zhang, 2019); H3K27ac, which is associated with the higher activation of transcription and therefore defined as an active enhancer mark (T. Zhang, Zhang, Dong, Xiong, & Zhu, 2020); or dual H3K4me3/H3K27ac enrichment (Beacon et al., 2021); as well as

H3K27me3 enrichment, which is involved in transcriptional silencing (Bogliotti & Ross, 2012) (**step 4**). The schematic is shown in Figure 11. Furthermore, gene ontology and transcription factor (TF) motif enrichment analysis can be done on these categorised DARs. For a more in-depth understanding, all of the regions can be further classified based on their methylation state (high or low). Please see Chapter 2 for an application of this multi-omics analysis.

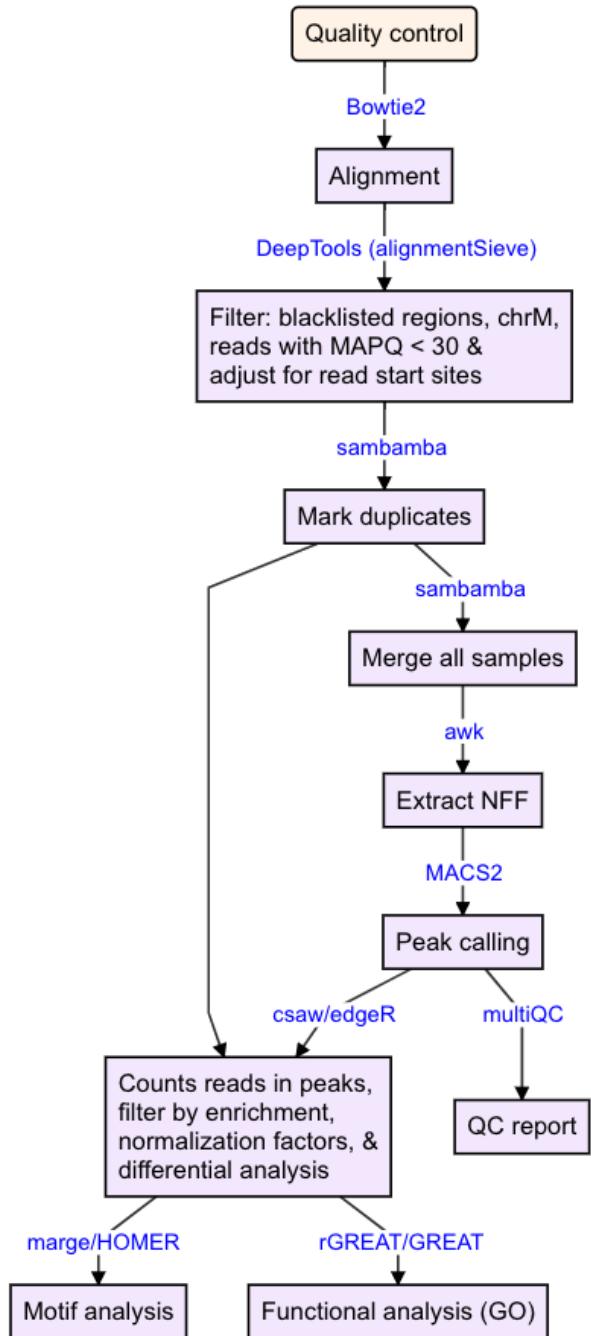


Figure 8: Workflow for ATAC sequencing data analysis. QC: Quality Control, NFF: Nucleosome Free Fragments.

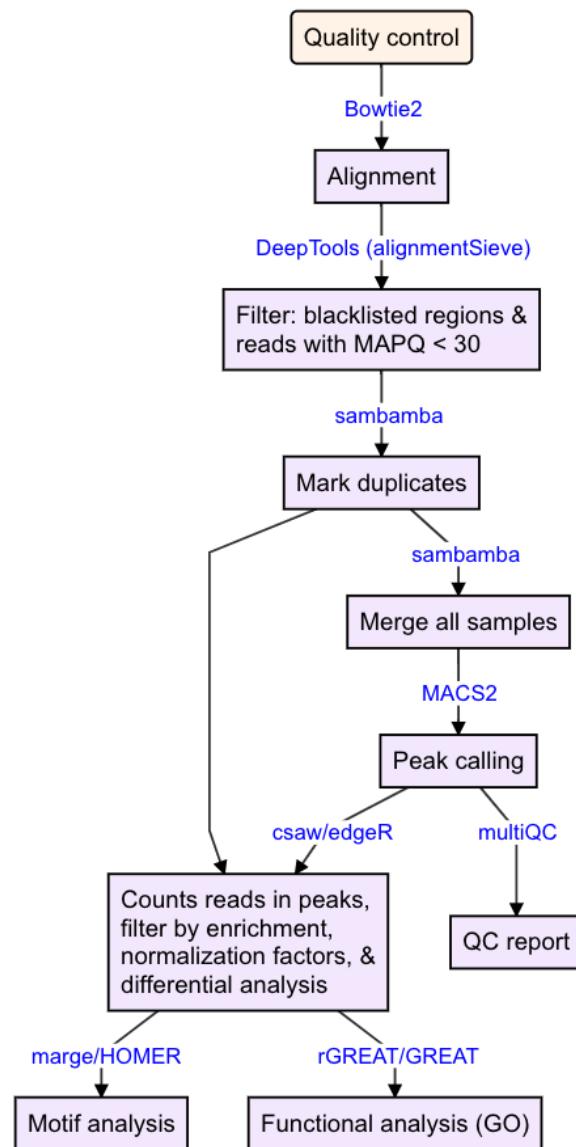


Figure 9: Workflow for ChIP sequencing data analysis. QC: Quality Control.

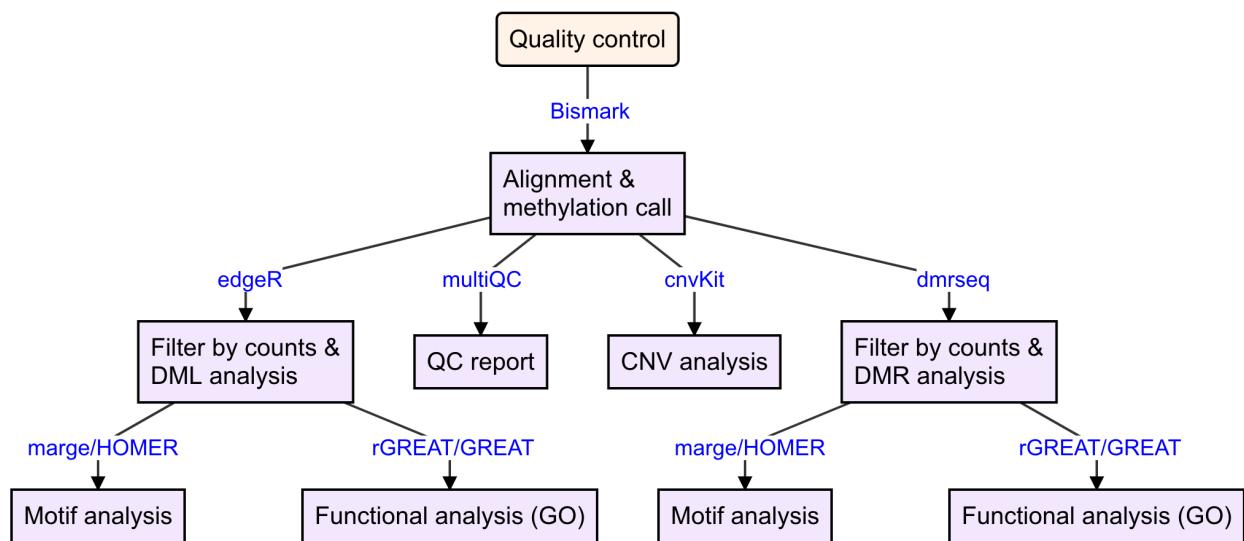


Figure 10: Workflow for Bisulfite sequencing data analysis. QC: Quality Control, DML: differentially methylated loci, DMR: differentially methylated regions, CNV: copy-number variation.

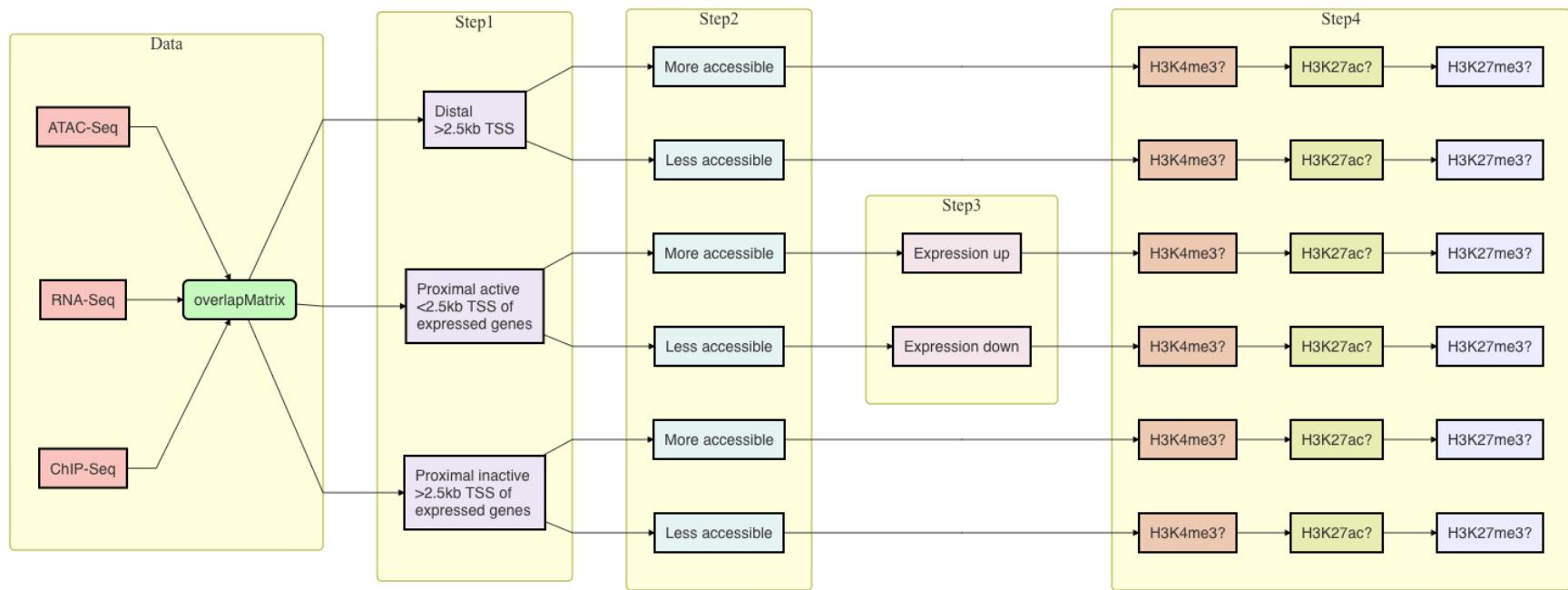


Figure 11: Workflow for multi-omics data integration.

Visualization of these regions is an essential step. We used EnrichedHeatmap and ComplexHeatmap packages for the visualization of categorized DARs.

ATAC-seq: ATAC-seq data is the starting point and we first find the middle of each DARs. After that, we obtain the coverage for 1kb upstream and downstream from the middle of DARs. We use the normalizeToMatrix function from the EnrichedHeatmap package and normalized the signal using the weighted mean between the intersected parts and un-intersected parts(`mean_mode = "w0"`) in a window of 50bp.

$$v_{w0} = \frac{v_w}{W + W'}$$

$\{W\}$ is the sum of width of the intersected parts $\left(\sum_i^n w_i\right)$ and W' is the sum of width for the non-intersected parts.

Moreover, we also plot the log2fold-changes of DARs using ComplexHeatmap.

ChIP-seq: ChIP-seq data coverage for 1kb upstream and downstream from the mid of DARs is obtained. We use the normalizeToMatrix function from the EnrichedHeatmap package and normalized the signal using the weighted mean between the intersected parts and un-intersected parts (`mean_mode = "w0"`) in a window of 50bp.

BS-seq: BS-seq data methylation values for 1kb upstream and downstream from the mid of DARs is obtained. We use the normalizeToMatrix function from the EnrichedHeatmap package and normalized the signal using the mean of all the signals (`mean_mode = "absolute"`) in a window of 50bp.

$$v_a = \frac{\sum_i^n x_i}{n}$$

RNA-seq: For RNA-seq data, we used ComplexHeatmap to plot the log2-fold-changes of expression values between the groups, and also the log2-fold-changes obtained after the differential analysis.

Chapter 1

Dynamic chromatin accessibility in spermatogonial cells for transcriptional programings from early postnatal to adult stages

Deepak K Tanwar^{1,2,3,†}, Irina Lazar-Contes^{1,2,3,†}, Pierre-Luc Germain^{1,2,3,4}, Isabelle M Mansuy^{1,2,3,#}

¹Laboratory of Neuroepigenetics, Brain Research Institute at the Medical Faculty of the University of Zurich.

²Institute for Neuroscience of the Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland.

³Zurich Neuroscience Center, ETH and University of Zurich, Zurich, Switzerland.

⁴Statistical Bioinformatics Group, Swiss Institute of Bioinformatics, Zurich, Switzerland.

†Equal contributions

#Corresponding author

Contributions: *I performed data analysis, generated figures, wrote computational methods and figure captions, revised manuscript.*

1.1 Update on the new data analysis

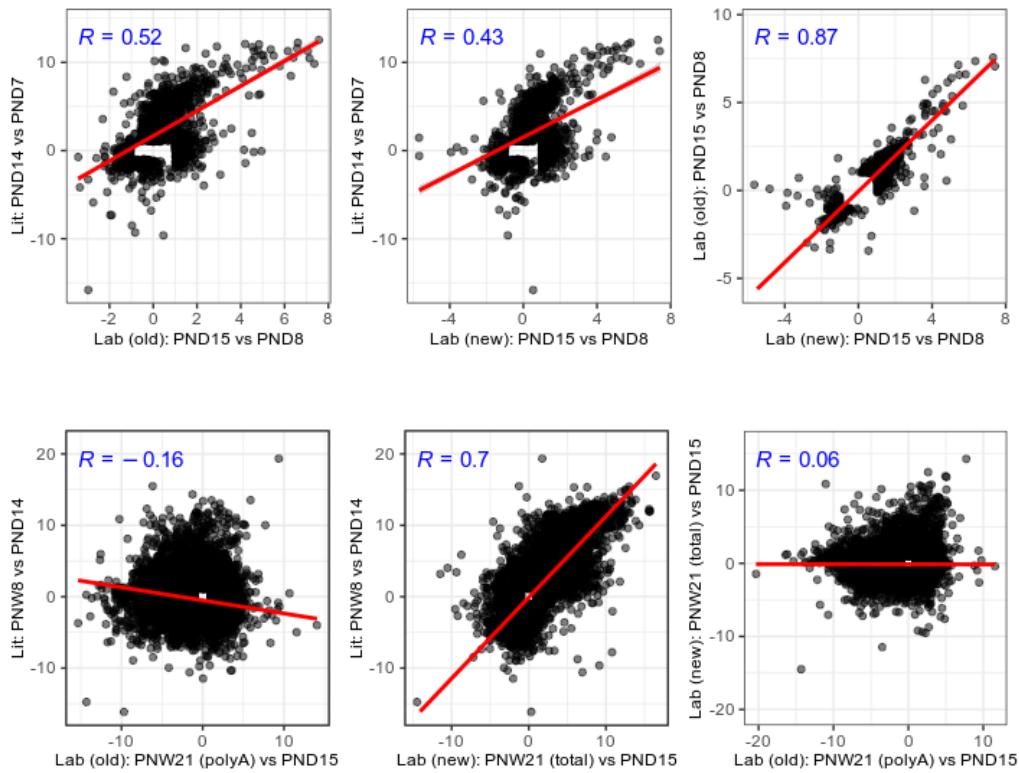
This project has undergone two major changes as a result of the replacement of one of the spermatogonial cells (SCs) RNA-seq datasets by a better one.

1.1.1 Inclusion of RNA-seq data from the literature

Adult SCs RNA-seq data from the lab was excluded since it was obtained from polyA RNA-seq, as opposed to pups' RNA-seq, which was Total RNA-seq. In the preprint version of the manuscript ([lazar-contes2020?](#)), we used RNA-seq data generated in the lab from SCs at the developmental stages postnatal day 8 (PND8), PND15, and postnatal week 21 (PNW21; also termed as adults in the manuscript). The RNA-seq datasets from PND8 and PND15 were generated using total RNA libraries, whereas the datasets from PNW21 were generated using polyA RNA libraries. Because polyA and total RNA-seq datasets are not directly comparable ([zhao2018?](#)), we obtained RNA-seq data from the PND7, PND14, and PNW8 stages from the literature (Hammoud et al., 2014, 2015) and compared it to the lab's RNA-seq data.

First, we compared the gene expression of lab PND8 samples to PND7 from literature and lab PND15 samples to PND14 from literature and discovered a strong correlation; however, data from PNW8 and PNW21 were not well correlated (result not shown). Following that, we performed differential analysis on data from the lab: PND15 vs PND8 and PNW21 vs PND15, as well as data from the literature: PND14 and PND7, and PNW8 and PND14. We compared the log₂ fold-changes from PND14 vs PND8 with PND15 vs PND8, and PNW8 vs PND14 with PNW21 vs PND15 after differential analysis. We discovered that the correlation between PND14 vs PND8 and PND15 vs PND8 was 52% ($R = 0.52$, Pearson method) and that most of the genes were changing in the same direction, whereas the correlation between PNW8 vs PND14 and PNW21 vs PND15 was negative 16% ($R = -0.16$; Pearson method) and that around 50% of the genes were changing in the opposite direction, as shown in Figure 1A. As a result, we incorporated the PND14 and PNW8 datasets into our project to better compare developmental stages. We used differential analysis between PNW8 and PND14 rather than PNW8 and PND15 because the PND14 and PND15 samples were sequenced separately and could have unaccountable batch effects. These analyses were then updated in the manuscript and appear in Chapter 1.

(ref:dnc1) **Log₂ fold-change comparison of RNA-seq data from the literature and lab (old and new).**

**Figure 1.1:** (ref:dnc1)

1.1.2 New RNA-seq data from lab

Because of these limitations, we decided to generate new SCs data in the lab. Previously, the “Takara total RNA pico input mammalian kit v2” was used to generate libraries, but this kit is no longer available. The new RNA-seq dataset was obtained from the total RNA libraries prepared using the “Takara total RNA pico input mammalian kit v3.” Furthermore, PND8 and PND15 libraries were previously sequenced on the Illumina HiSeq 4000 platform, which is no longer available for sequencing at the Functional Genomics Center Zurich (FGCZ). To avoid any technical/ batch effects, we therefore re-sequenced the total RNA libraries from PND8 and PND15 on the Illumina NovaSeq platform using the “Takara total RNA pico input mammalian kit v3” along with the PNW21 total RNA libraries.

First, we compared the newly sequenced data to data generated previously in the lab and from the literature. We discovered that gene expression from newly generated PND8 data was highly correlated with data generated in the lab previously and data from the literature. The data from the PNW21 stage is also well correlated with the literature (result not shown).

(ref:dnc2) **Volcano plot of differentially expressed genes (adjusted $P \leq 0.05$ and**

absolute $\text{Log}_2\text{FC} \geq 1$).

A: PND15 and PND8. **B:** PNW21 and PND15. We processed the RNA-seq data using the

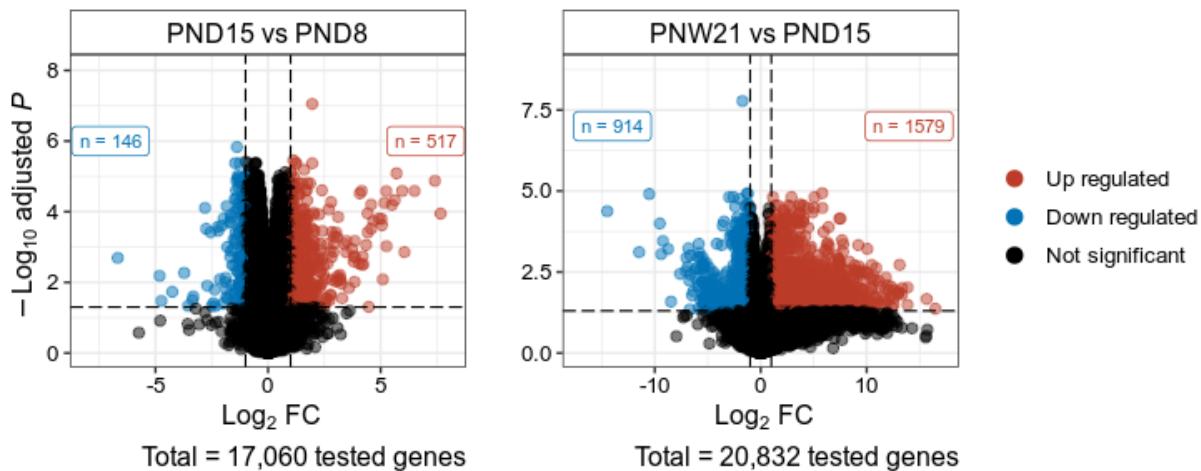


Figure 1.2: (ref:dnc2)

methods described in Chapter 1 and Methods. Following that, we conducted a differential analysis between PND15 and PND8, as well as PNW21 and PND15. Figure 2 shows that we found 663 genes that were differentially expressed between PND15 and PND8, and 2,493 genes that were differentially expressed between PNW21 and PND15.

We compared log2 fold-changes from new data PND15 vs PND8 to literature data PND14 vs PND8, and new data PNW21 vs PND15 to literature data PNW8 vs PND14. We discovered that the correlation between PND14 vs PND8 and PND15 vs PND8 was 43% ($R = 0.43$, Pearson method) and that most of the genes were changing in the same direction, and that the correlation between PNW8 vs PND14 and PNW21 vs PND15 was 70% ($R = 0.7$; Pearson method) and that most of the genes were changing in the same direction, as shown in Figure 1C, 1D. Furthermore, we compared the lab's old data with the new data and discovered an 87% correlation ($R = 0.87$; Pearson method) and nearly all the genes changing in the same direction for PND15 vs PND8, whereas the correlation for PNW21 vs PND15 was 6% ($R = 0.06$; Pearson method) and nearly 50% genes changing in the opposite direction, Figure 1E, 1F. This is due to the fact that the previous PNW21 data were obtained from polyA libraries, whereas the new PNW21 data were obtained from total RNA libraries.

1.1.2.1 Identification of clusters of genes with distinct expression profiles

We discovered five clusters of genes that were differentially expressed. We have used the union of differentially expressed genes between PND15 and PND8, as well as PNW21 and PND15, and ran k-means clustering ($k = 2$ to 10) on the log₂ fold-changes of normalized expression values, using PND8 samples as the baseline. Based on visual inspection, we select 7 clusters ($k = 7$) and merge the clusters with visually observed similar expression profiles to obtain 5 clusters, Figure 3.

(ref:dnc3) K-means clustering ($k = 5$) of genes with differential expression in postnatal versus adult spermatogonial cells. Each row depicts a gene and each column depicts a sample. Gene expression fold-change is calculated by subtracting mean log₂ counts per million (CPM) of PND8 from PND15 and PNW21. Samples are clustered using Ward's method and genes are ordered by the "MDS_angle method" using `seriation` (R package).

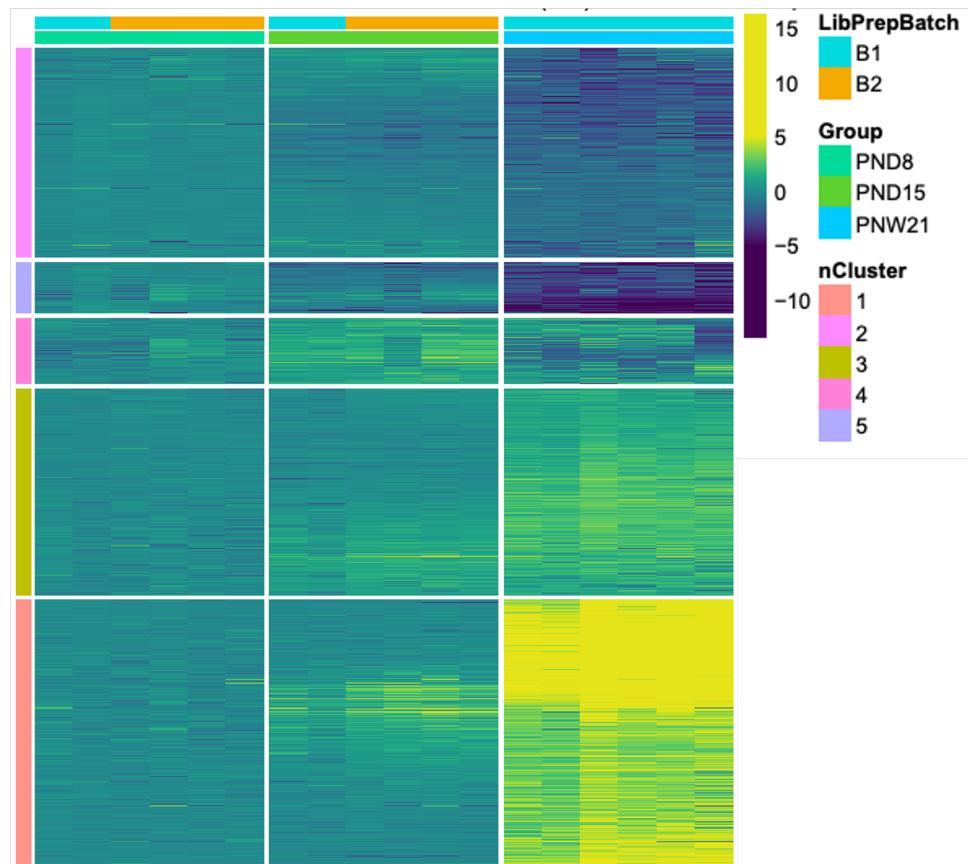


Figure 1.3: (ref:dnc3)

1.1.2.2 Multi-omics data integration

After combining data from ATAC-seq, RNA-seq, and ChIP-seq, we were able to generate clusters of distinct genome activity. We combined these datasets, with ATAC-seq as the starting point, Methods. We classified the differentially accessible regions (DARs) based on ATAC-seq and RNA-seq data into distal (> 2.5 kb of the transcription start site, TSS) and proximal (2.5kb of TSS). Following that, distal DARs were divided into two groups based on increased or decreased accessibility in PNW21, Figure 4,6,8, while proximal DARs were divided into six groups based on increased or decreased accessibility in PNW21 and up or down-regulation of nearby gene expression in PNW21 Figure 5,7,9. Category 1: increased chromatin accessibility and upregulation of neighbouring genes; Category 2: increased chromatin accessibility and downregulation of neighbouring genes Category 3: decreased chromatin accessibility and downregulation of neighbouring genes, Category 4: decreased chromatin accessibility and upregulation of neighbouring genes, Category 5: increased chromatin accessibility with no detectable gene expression, and Category 6: decreased chromatin accessibility with no detectable gene expression. All such clusters had also been identified in the previous analysis (see Chapter 2 and (lazar-contes2020?)), however one of the original categories (Category 5 from Figure 5) disappeared in the new analysis. This category is related to TSS-proximal DARs which, against expectations, increase in expression despite a decrease in accessibility. Since expression implies promoter accessibility, it is safe to assume that this category was attributable to lack of comparability of the two RNAseq protocols.

To conclude, the results obtained from the newly generated RNA-seq data from the lab are quite similar to the results generated previously. We have obtained similar results in terms of differential analysis and multi-omics data integrations. However, a number of analyses are still to be performed, for instance, gene-ontology analysis on differentially expressed genes, identified clusters of gene expression, and clusters obtained from multi-omics data integration; transcription factor motif analysis of clusters obtained from multi-omics data integration; and revision of figures and manuscript.

Figures 4,5,6,7 are enriched heatmaps that show different clusters for distal and proximal regions. Each row represents a peak region, which is ordered by the ATAC signal. The mid-x-axis corresponds to the centre of a peak region and is extended to $+/- 1$ kbp. The ATAC heatmap's colour key represents the signal. For RNA-seq, the fold-change in gene expression is calculated by subtracting the mean \log_2 CPM of PND8 from all samples.

(ref:dnc4) Enriched heatmap of distal differentially accessible regions. (ref:dnc5)
Enriched heatmap of proximal differentially accessible regions. (ref:dnc6) Enriched
heatmap of distal differentially accessible regions with ChIP-seq and BS-seq

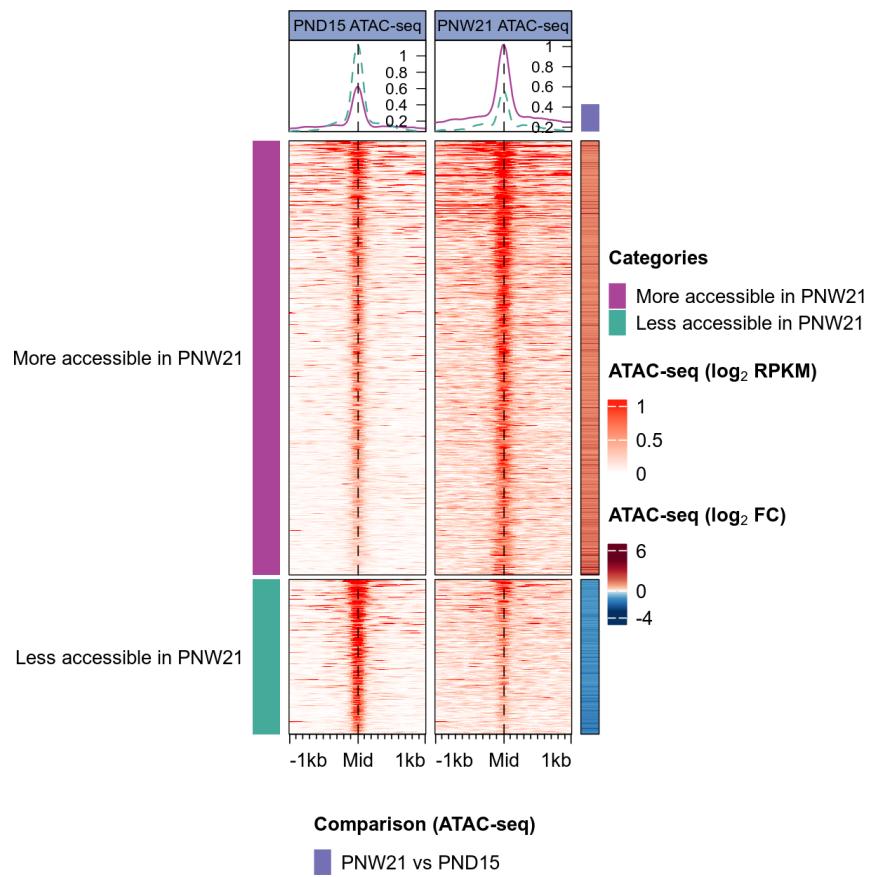
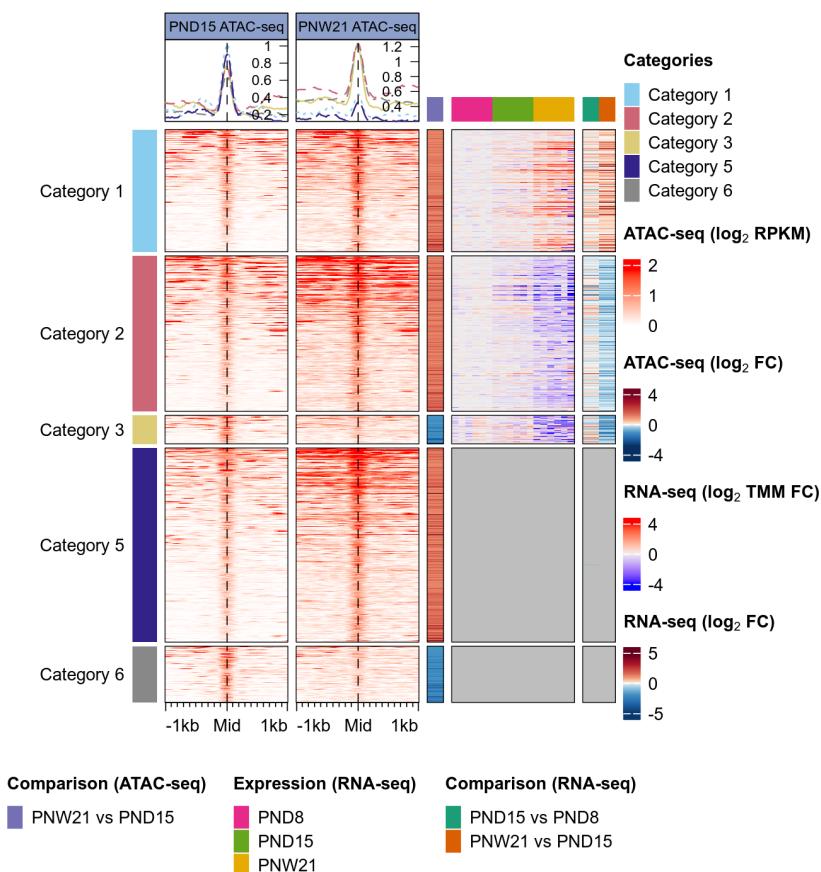
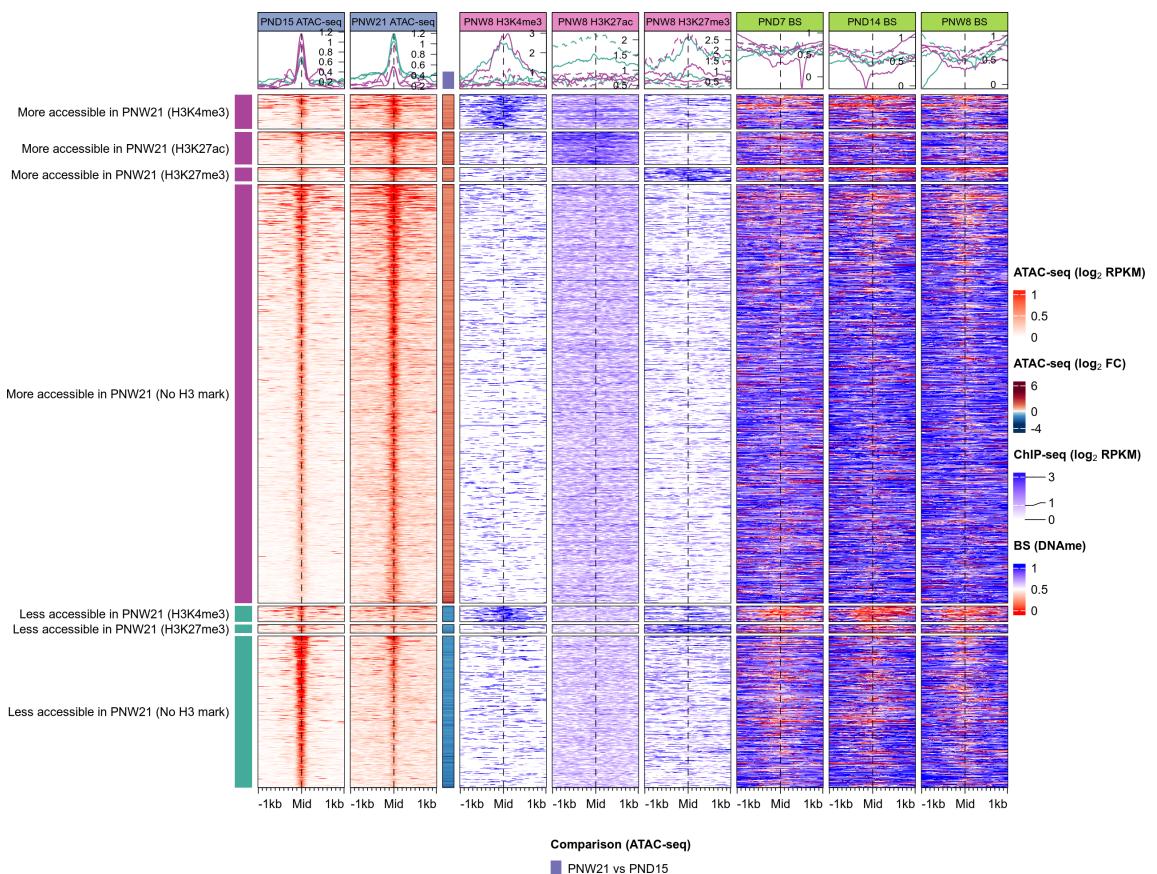


Figure 1.4: (ref:dnc4)

data. (ref:dnc7) Enriched heatmap of proximal differentially accessible regions with ChIP-seq and BS-seq data.



**Figure 1.6:** (ref:dnc6)

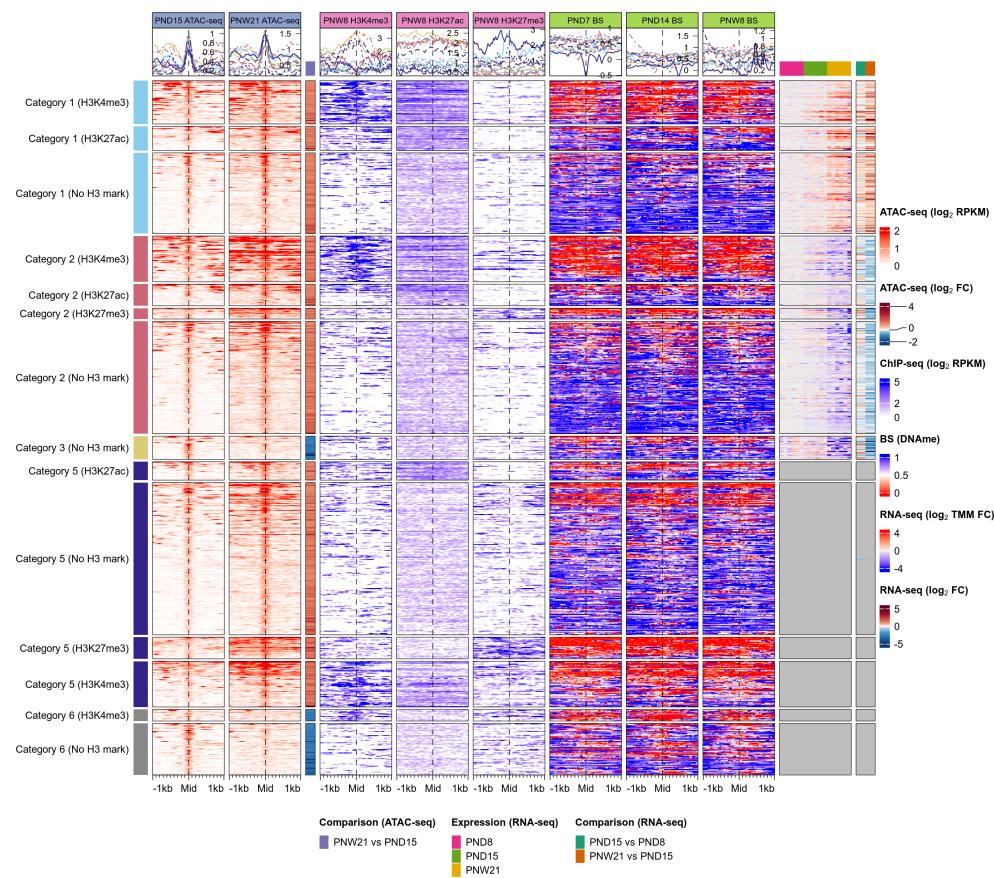


Figure 1.7: (ref:dnc7)

Chapter 2

shortRNA: a flexible framework for the analysis of short RNA sequencing data

2.1 Abstract

Short RNA are important molecules that play key functional roles in the regulation of the genome. Several classes of short RNAs, such as miRNA, tRNA and tRNA fragments, and piRNAs have been characterized and have complex biogenesis. Short RNA sequencing is becoming increasingly relevant in the research of regulatory mechanisms in a wide range of biological functions. From an analysis point of view, each type of RNA has its own features, and hence, specialized methods have been developed focused on particular types, which not only multiply the work needed for a comprehensive analysis but potentially create misassignment problems. In addition, methods are typically divided into genome-based methods that do not deal adequately with post-transcriptional modifications, and transcript-based methods that are blind to unannotated features. Finally, there are several outstanding issues in the analysis of short RNAs, which are critical in the analysis of some samples. There is, therefore, a need for an analysis framework that is: sufficiently tailored to consider the specificities of different classes of short RNA and their biogenesis, and sufficiently general and exhaustive to perform global analyses. We developed a user-friendly, highly flexible and comprehensive R package for a thorough end-to-end analysis of short RNA sequencing data (planned submission to Bioconductor). Our package is appropriately adapted to take into account the differences between different types of short RNA and their biogenesis and is seamlessly expandable to include additional annotation. We use a customized genome

annotation with artificial chromosomes to account for post-transcriptional modifications, and a flexible rule-based approach to assign reads along a tree of hierarchically organized features. This enables the systematic querying, exploration, and differential expression analysis of short RNAs at various degrees of granularity, from specific sequences to RNA classes. We also include various normalization and experimental-bias correction methods. The shortRNA R package is developed to perform all the steps from within R and is platform-independent.

2.2 Introduction and background

Previous decades have revealed a number of RNAs that are distinct from the messenger RNAs. Short-RNAs are non-coding RNA molecules that are fewer than 200 nucleotides long and play an important role in genome regulation. Several types of short RNAs, of different lengths, have been discovered including microRNA (miRNA; 18-24 nt); small interfering RNA (siRNA; 21-27 nt); small nucleolar RNA (snoRNA; 60-170 nt); small nuclear RNA (snRNA/ U-RNA; 10-300 nt); Small temporal RNA (stRNA; 18-24 nt); tRNA-derived small RNA (tsRNA), including tRNA halves (tsRNA; 28-36 nt) and tRNA fragments (tRFs; 14-22 nt); small rDNA-derived RNA (srRNA; 18-30 nt); repeat-associated small interfering RNA (rasiRNA; 24-27 nt); and Piwi-interacting RNA [piRNA; 26-31 nt, present in animals in the germline cells and also in the brain (Zuo, Wang, Tan, Chen, & Luo, 2016)]. Some short RNAs, for instance, miRNA and siRNA, are known to suppress gene expression by sequence-specific interactions with regulatory regions during transcription, RNA processing and translation (J. Zhu et al., 2016), and by forming the core with RNA-induced silencing complex (RISC) (Pratt & MacRae, 2009). In addition, short-RNAs have been related to cancer, Parkinson’s disease, Alzheimer’s disease, and prion disease (Gong, Liu, Liu, & Liang, 2005), and hence, are used as biomarkers.

Several methods have been developed for the analyses of short RNA sequencing data. The development of high-throughput sequencing (HTS) techniques allows researchers to study short RNAs in diverse tissues or cells. HTS not only allows for the quantification of known short RNAs but also for the identification and quantification of novel short RNAs. The small RNA sequencing (sRNA-Seq) bears challenges and biases that researchers need to be informed of in order to properly analyze the data. First, most of the methods concentrate on a specific type of short RNA, such as MINTmap (Loher, Telonis, & Rigoutsos, 2017) focus on analyzing tRFs, TAM 2.0 (J. Li et al., 2018), Prost! (Desvignes, Batzel, Sydes, Eames, & Postlethwait, 2019), Chimira (Vitsios & Enright, 2015), and mirTools 2.0 (J. Wu et al., 2013) focus on miRNAs. Next, a few tools concentrate on most types of short RNAs (not all), such as ncPRO-seq (C.-J. Chen et al., 2012), sRNAtoolbox (Rueda et al., 2015), SPAR (Kuksa

et al., 2018), Threshold-seq (Magee, Loher, Londin, & Rigoutsos, 2017), sRNAAnalyzer (X. Wu et al., 2017), UEA sRNA Workbench (Stocks et al., 2012), Sports1 (Shi, Ko, Sanders, Chen, & Zhou, 2018), and Oasis 2 (Rahman et al., 2018). Tools that concentrate on many types of short RNAs, such as Oasis2 and Sports1, perform sequential mapping, for example, they perform sequential mapping by first assigning the reads to miRNAs, then assigning the leftover to tRNAs, and then the remaining ones to mRNAs, which can create a misassignment problem as short reads can map to multiple locations. But this information is lost because only unmapped reads are aligned against the next annotation. In addition, although the libraries for sequencing are tailored to short RNAs, other long RNAs are also often detected in the sequencing data, and also because short RNAs often overlap with other features, prioritization between the biotypes is crucial. Further, there are quite a high number of sequenced reads left that map to the reference genome but are not assigned to the known features. Moreover, although most of the tools do not consider assignment rules for reads assignment, few tools such as Prost! have a set of defined rules for reads assignment. But, the rules are only for the miRNAs. Furthermore, the multi-mapping reads are either excluded or randomly assigned to the multi-mapping positions in the genome. As well as, not all the tools deal with the post-transcriptional modifications, such as the addition of "CCA" towards the 3' end of tRNAs (Barraud & Tisné, 2019; Hou, 2010; Ibba & Soll, 2000), which is important for the recognition of tRNA by enzymes and translation (Green & Noller, 1997; Sprinzl & Cramer, 1979); and the addition of "G" towards the 5' end in histidine tRNAs (Cooley et al., 1982; Cozen et al., 2015), this is critical for histidyl-tRNA synthetase (HisRS) recognition (Fromant, Plateau, & Blanquet, 2000), which is responsible for the integration of histidine into proteins (Freist, Verhey, Rühlmann, Gauss, & Arnez, 1999). On top of these limitations, most of the tools are written in different programming languages, are platform-dependent and depend on external tools, which can create a barrier for installation for the user; and the tools are either web-based, which restrict the user for additional downstream analysis or command-line based, which creates a restriction for the non-computational researchers. Besides, the available tools use feature-based counting (number of reads, counts, associated with each feature), except for seqpac (Skog et al., 2021), a recently published tool, which uses sequence-based counting (number of reads, counts, associated with each unique sequence). Sequence-based counting would prevent the same sequence from being annotated multiple times within and across samples, hence increasing the efficiency for alignment and reads assignment. Also, it directly enables looking at specific variations in sequences or their boundaries. However, due to this complexity, it can be challenging to explore data at the level of individual sequences, which calls for methods that can afford different levels of granularity. Hence, specialized tools are required with specialized pipelines to analyze sRNA-seq data.

We developed shortRNA, an R tool that addresses all of the issues listed above for processing sRNA-seq data. Our tool is cross-platform (it may run on any operating system) and built on the Bioconductor framework, in particular making use of efficient data structures (e.g., DataFrame, FactorList, and TreeSummarizedExperiment) enabling interoperability with other Bioconductor packages. Users can use shortRNA to conduct a thorough analysis of their data, from quality control (trimming, adapter removal, UMI compressing) and alignment to quantification and downstream analysis. ShortRNA also has a set of customizable criteria for assigning reads to various types of short RNAs and seamlessly enables querying features at the level of individual fragments. We have tested our tool using a published mouse dataset from sperm, simulated data from sperm, and human data from peripheral blood mononuclear cells (PBMC; unpublished).

2.3 Methods

2.3.1 Development and testing environment

The shortRNA package is developed and tested on the Linux (GNU/Linux 4.4.0-210-generic x86_64) operating system, Ubuntu (version 16.04.7 LTS) with 16 processors and 124 GB of RAM, using R 3.6.3 and R 4.0.5. As the shortRNA tool is developed as an R package, it is platform-independent (can easily be installed and run on other operating systems).

2.3.2 Backbone data structures of shortRNA

The shortRNA package is developed around four main data structures from Bioconductor: phylo, FactorList, DataFrame, and TreeSummarizedExperiment. These four data structures store the data and results in the most efficient manner and enable fast computation.

2.3.2.1 phylo

The phylogenetic tree is a branching diagram used to depict evolutionary relationships. In R, the phylo class stores the phylogenetic relationship. In shortRNA, we save the relationship between RNA biotypes, features, and reads as a phylo object, section ____.

2.3.2.2 FactorList

In the case of a long vector of repeated characters, R takes more memory to store them and also the computation could be quite slow. These repeated character vectors could be stored as factors, where levels are provided to each unique character in the

vector (<https://datascience.stackexchange.com/questions/12018/when-to-choose-character-instead-of-factor-in-r>). However, when factors are stored in a list, the levels are traditionally defined for each element of the list, which can be very inefficient memory-wise. The FactorList, a class from the IRanges Bioconductor package, instead stores lists of factors as a single factor vector with added list membership information. In addition, like other AtomicList objects, it enables list operations without iteration. This is helpful for efficiently saving the object in R memory and for fast computations.

2.3.2.3 DataFrame

Rectangular data can be stored as data.frame class object in R. DataFrame functions from S4Vectors Bioconductor package and behaves similar to data.frame in terms of construction and subsetting. An advantage of using DataFrame is that it can store any type of object in a column, for example, a FactorList, or even another DataFrame, while retaining all the methods and functionalities of traditional data.frames.

2.3.2.4 TreeSummarizedExperiment

TreeSummarizedExperiment is a Bioconductor package extending the classical SummarizedExperiment (Morgan, Obenchain, Hester, & Pagès, 2020) and with additional hierarchy data and operations (R. Huang et al., 2020). The SummarizedExperiment class holds rectangular matrices of experiment data, accompanied by row and column annotation data. There are two classes in the SummarizedExperiment package: SummarizedExperiment and RangedSummarizedExperiment. Instead of a DataFrame of features, RangedSummarizedExperiment objects represent genomic ranges of interest. Figure 2.1A depicts the structure of the SummarizedExperiment. The rectangular data matrices are stored as assays, rowData stores annotation for corresponding rows in the assays, colData stores annotation for corresponding columns in the assays, ranges are described by a GRanges or a GRangesList object, which are stored as rowRanges, and metadata can be used to store additional data-related information.

In extension to the SummarizedExperiment, TreeSummarizedExperiment has rowTree, which stores the hierarchical structure of rows of assays; colTree, which stores hierarchical structure of columns of the assays; rowLinks, which stores the link information between the rows of the assays and rowTree; colLinks, which stores the link information between the columns of the assays; and referenceSeq, which stores the reference sequence for the features, depicted in Figure 2.1B.

2.3.3 Pipeline

The shortRNA pipeline includes three major data analysis steps: preprocessing of data, alignment and reads assignment with the customisable assignment rules, and statistical analysis and visualization. Figure 2.2 shows the outline of the pipeline. First, the data goes through the preprocessing steps that consist of filtering and alignment of unique sequenced reads. The raw data undergo quality assessment and then the quality control steps are performed. The data is then aligned to a custom genome containing, in addition to the reference genome, extra pseudo-chromosomes for handling post-transcriptional modifications. Aligned reads are then overlapped with the features annotation and reads are assigned to the features using customisable assignment rules, which is an important part of this pipeline, after that, a features tree is formed. After the reads assignment, the data goes through the statistical analysis steps that consist of normalization and differential analysis. The quality controlled data are normalized and then the test for differential expression of short RNAs is performed in the user-defined biological groups.

2.3.3.1 Quality control and trimming

Quality control (QC) and trimming is a vital step in HTS data analysis. To assess and control the quality of sRNA-Seq data, we adapted the functionalities from two R packages: Rfastp (Wang & Carroll, 2020) and seqTools (Kaisers, 2020). Rfastp is used for trimming the low-quality reads and adapter trimming. Further, with the summary files from Rfastp, the functions in our tool could be used for making a table of comparison for before and after QC, plots for duplicated reads, reads quality plot, and base-content plots. We benchmarked Biostrings (Pagès, Aboyoun, Gentleman, & DebRoy, 2020), qrqc (Buffalo, 2020), ShortRead (Morgan et al., 2009), and seqTools for reading and storing FASTQ files quality data in R, using an 863 MB FASTQ file. seqTools outperformed the other three tools in terms of reading and storing the FASTQ file quality data, as shown in Figure 2.3. Hence, we used it for reading the FASTQ files quality data, before and after QC, to make reads length distribution plots. To summarize, the QC functions provide information about sequencing depth, reads quality, possible adapter sequences, duplicated reads, and sequence length distribution, both before and after QC. The interactive plots provide users with greater insight into their data. Exemplary plots and tables from the QC report of a sample are shown in Figure 2.4.

Unique Molecular Identifiers (UMIs) are complex indices of 8-16 nucleotide lengths that are introduced to sequencing libraries before PCR amplification steps. Researchers can use UMI to evaluate the efficiency with which they collect input molecules, identify sampling bias, and, most importantly, identify and compensate for the effects of PCR amplification

bias (Y. Fu, Wu, Beane, Zamore, & Weng, 2018). Preprocessing of data sequenced with UMIs is required for reads deduplication and correction, as well as the creation of consensus sequences from each UMI. In shortRNA, we have adapted the UMIc tool for UMIs collapsing (Tsagiopoulou et al., 2021), which takes into account the frequency and the Phred quality of nucleotides and the distances between the UMIs and the actual sequences for collapsing the sequences.

Schematic of quality checks and quality control is shown in Figure 2.5.

2.3.3.2 Sequence count matrix

From the trimmed FASTQ files, we extract the unique reads sequences and make a sample by sequence count matrix. This matrix represents the counts of sequences in each sample, keeping only sequences occurring more than once. This is done by the fastq2SeqCountMatrix() function from shortRNA. Further, we export these sequences as a fasta file. An overview of this step is shown in Figure 2.5. The two main outputs from this step are sequenceFasta, which would be used for alignment, and countsMatrix.

2.3.3.3 Annotation preparation

Sequencing generates nucleotide sequences with unknown functions, while sequence annotation provides descriptive information about sequenced DNA sequences. Several databases are available for the annotation of the sequences. In shortRNA, we used miRBase (Griffiths-Jones, Saini, van Dongen, & Enright, 2008) for miRNAs; GtRNADb (P. P. Chan & Lowe, 2016) for tRNAs; mitoRNADb (Jühling et al., 2009) for mitochondrial tRNAs; for mouse, we use piRNA precursors from (X. Z. Li et al., 2013), rRNAs from SILVA (Quast et al., 2013), and mRNAs and other biotypes from Ensembl (Howe et al., 2021). It is possible for users to have additional databases of their choice.

The miRBase database is a collection of miRNA sequences and annotations. In miRBase, miRNAs are named as, for example, org-mir-20a. The first 3 letters signify the organism, “20” tell us that it was an early-discovered family (20th family that was named), “20a” tell us that there is possibly another related miRNA, for example, org-mir-20a. Here, org-mir-20a is a precursor. The mature miRNA species may be derived from both the 5’ and 3’ arms of the precursor duplex, and are called the miRNA-5p and -3p species, respectively. In this example, it would be org-mir20a-5p and org-mir20a-3p. All the miRNAs overlapping between Ensembl and miRBase were removed from Ensembl and the remaining one from Ensembl, not overlapping, were renamed to match the nomenclature of miRBase. For example, Mir7679 is renamed to org-miR-7679. Further, if the length of Ensembl miRNAs were more than 25bp,

we labelled them as precursors.

A miRNA cluster is a group of two or more miRNA hairpins that are transcribed from miRNA genes that are physically adjacent, transcribed in the same direction, and are not separated by a transcription unit or a miRNA in the opposite orientation. There are mostly 2-3 mature miRNAs in a miRNA cluster, but there is an existence of a bigger miRNA cluster, which is found in humans on chromosome 13: miR-17 to miR-92, involved in tumour formation, and development of heart lungs and immune systems (Lai & Vera, 2013). Researchers group the miRNAs to form a miRNA cluster by the distance between them. For example, (Baumgart et al., 2017; Griffiths-Jones et al., 2008) call a group of miRNA as clusters if miRNAs are within 10kb and (Yuan et al., 2009) used a distance of 50kb. In shortRNA, we use a distance of 10kb for clustering miRNAs as more than 40% of experimentally validated human miRNA cluster genes have been identified within 10kb (Griffiths-Jones et al., 2008; Lai & Vera, 2013).

The genomic tRNA database (GtRNADb) is a database of tRNA gene predictions created by tRNAscan-SE (P. Chan, Lin, Mak, & Lowe, 2021) on whole or almost complete genomes. In GtRNADb, tRNAs symbols consist of 5 parts, separated by a “-”, for example, tRNA-Ala-AGC-9-2. In this example, tRNA (prefix) represents tRNA genes that are high scoring and not predicted as pseudogenes. If it was a pseudogene, it would have been represented as “tRX”. At the second position, we have three-letter amino acids (isotype), which stand for tRNA isotype. At the third position, we have anticodon detected in the gene sequence. At the fourth position, there is a unique ID (transcript ID) of a tRNA transcript or “isodecoder” with a particular isotype and anticodon. The fifth position represents the gene locus ID and for tRNA genes with multiple identical copies, this gene locus ID represents the particular gene copy in the genome. For the mitochondrial tRNAs, we obtained the sequences from mitotRNADb (Jühling et al., 2009) using the tRNADbImport (GM, 2018) R package. Further, we removed all the duplicated sequences and renamed the mitochondrial tRNAs as per the naming convention of GtRNADb, adding “mt” as a prefix. Mature tRNAs receive a post-transcriptional addition of “CCA” (Barraud & Tisné, 2019), hence, we added “CCA” to all the tRNA/ mt-tRNA sequences at the 3’ end. All histidine tRNAs of known sequence are one nucleotide longer at the 5’ end than are other tRNA species(Cozen et al., 2015),hence we added “G” at the 5’ end of the sequences of Histidine tRNAs/ mt-tRNAs to account for this post-translational modification, as also done in (Cozen et al., 2015; Shi et al., 2018). Because of these post-transcriptional modifications, we add the sequences of tRNA into the genome FASTA file as pseudo chromosomes. For example:

```
>pseudoChr_tRNA-Thr-CGT-1-1
GGCGCGGTGCCAACGTGGTAAGGCGTCGGTCTCGTAAACCGAACGATCACGGTTCGAACCCCGTCCGTGCCTCCA
```

```
>pseudoChr_mt_tRNA-His-GTG-1-1
GGTGAATATAGTTACAAAAAACATTAGACTGTGAATCTGACAACAGGAAATAAACCTCCTGTTACCCCCA
```

All the tRNAs overlapping between Ensembl and GtRNADb were removed from Ensembl and the remaining tRNA from Ensembl, not overlapping, were labelled as pseudo tRNAs.

SILVA is an rRNA database, which provides extensive, quality-checked, and regularly updated datasets of aligned short (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA sequences. rRNAs are repeated sequences and are masked in the genomes in the Ensembl database, and are not listed in the features annotation. Hence, we add them as well as pseudo chromosomes to the genome FASTA file, similar to tRNAs.

`getDB()`, Figure 2.6, function retrieve the databases, account for the post-translational modifications of tRNAs, alter mitochondrial tRNA database, and create miRNA clusters and assign miRNAs and miRNA precursors to miRNA clusters. The output from `getDB()` is then parsed to the `prepareAnnotation()` function, which prepares features annotation as `GRanges()`, save the genome and pseudo-genomes and index that for alignment. Features annotation are then converted to `FactorList` using `featuresAnnoToFL()`. The main three outputs from the annotation preparation are features as `GRanges()` and `FactorList()`, and indexed customGenome.

2.3.3.4 Alignment

Alignment is necessary to know where in the genome reads belong. The alignment of short RNA remains a persistent under-recognized problem. We align the FASTA file of sequences with the custom genome prepared in the annotation step. For alignment, we use the `Rsubread` R package (Liao et al., 2019) with the index of custom genome generated on the genome and pseudo-chromosomes (`index = "customGenome"`), with the GTF from Ensembl database obtained from `getDB()` of shortRNA (`GTF = exonsBy(db$ensdb)`), and allowing to report the maximal number of equally-best mapping locations (`nBestLocations = 16`). Please refer to the schematic in Figure 7.

2.3.3.5 Reads annotation and assignment

After alignment, the aligned file (BAM file) is overlapped with the features annotation (`featuresGR`) using the `overlapWithTx2()` function. It is possible that a read overlaps with multiple features. Next, the overlaps are parsed with the `assignReads()` function with the assignment rules, `defaultAssignRules()`, for validating the overlap and finding assignment of multi mapping reads. Please refer to the schematic in Figure 7.

2.3.3.6 Assignment rules

A sequence read can be mapped to multiple locations in the genome. It is critical to ensure that the reads are properly mapped to the feature and that they are assigned to one or more features. In the shortRNA package, we defined a set of customizable rules for this purpose. The output of the function `defaultAssignRules()` is described below:

We consider reads with at least 50% overlap to the features to be valid read assignments.

```
$overlapBy  
[1] 0.5
```

By default, we do not prioritize the assignment based on the size of the overlap between reads and features (when a read overlaps multiple features); however this option is available to users.

```
$prioritizeByOverlapSize  
[1] FALSE
```

We prioritize overlap in the same strand, but enable overlaps from the opposite strand if there is no known feature on the same strand.

```
$sameStrand  
[1] "prioritize"
```

We give priority to known features in our assignments.

```
$prioritizeKnown  
[1] TRUE
```

Next come validation rules specific to RNA types. For example, in order to be assigned to primary piRNAs, the read should be 26-32 nucleotides long, with the first nucleotide being a T.

```
$typeValidation$primary_piRNA  
$typeValidation$primary_piRNA$fun  
function(src, allowRevComp=FALSE, length=26:32){  
  length <- as.integer(length)  
  valid <- src$length >= min(length) & src$length <= max(length)  
  if(length(w <- which(valid))==0) return(valid)
```

```

seqs <- as.character(src$seq[w])
valid[w] <- sapply(strsplit(seqs,""),FUN=function(x){ x[[1]]=="T" })
if(allowRevComp){
  revcomp <- as.character(reverseComplement(DNAStringSet(seqs)))
  valid[w] <- valid[w] |
    sapply(strsplit(seqs,""),FUN=function(x){ x[[1]]=="T" })
}
valid
}

```

If the read cannot be identified as a primary piRNA, it is referred to as a piRNA precursor.

```
$typeValidation$primary_piRNA$fallback
[1] "piRNA_precursor"
```

Secondary piRNA sequences should be 26 to 32 nucleotides long, with A in the 10th position (Brennecke et al., 2007).

```
$typeValidation$secondary_piRNA
$typeValidation$secondary_piRNA$fun
function(src, allowRevComp=FALSE, length=26:32){
  length <- as.integer(length)
  valid <- src$length >= min(length) & src$length <= max(length)
  if(length(w <- which(valid)) == 0) return(valid)
  seqs <- as.character(src$seq[w])
  valid[w] <- sapply(strsplit(seqs,""),FUN=function(x){ x[[10]]=="A" })
  if(allowRevComp){
    revcomp <- as.character(reverseComplement(DNAStringSet(seqs)))
    valid[w] <- valid[w] |
      sapply(strsplit(seqs,""),FUN=function(x){ x[[10]]=="A" })
  }
  valid
}
```

If a read cannot be identified as secondary piRNA, it is referred to as a piRNA precursor.

```
$typeValidation$secondary_piRNA$fallback
[1] "piRNA_precursor"
```

For miRNAs, the length should be between 19 and 24 nucleotides, and the read should overlap the feature by at least 16 bp. The maximum number of non-overlapping nucleotides allowed is three.

```
$typeValidation$miRNA
$typeValidation$miRNA$fun
function (src, length = 19:24, minOverlap = 16L, maxNonOverlap = 3L)
{
  src$length %in% length & src$overlap >= minOverlap & (src$length -
  src$overlap) <= maxNonOverlap
}

$typeValidation$miRNA$length
[1] 19 20 21 22 23 24
```

If a read cannot be assigned to a mature miRNA, it is assigned to a miRNA precursor.

```
$typeValidation$miRNA$fallback
[1] "miRNA_precursor"
```

Reads that are assigned to tRNAs (or pseudo tRNAs) undergo additional classification. tRNA 5p fragments are less than 30 bp long and begin in the feature's 5bp. tRNA 3p fragments are less than 50 bp long, with a distance to the feature end of 5bp. Starts at ± 1 bp of the feature and has a length of 30 to 34 nucleotides for 5p half. The 3p half begins at ± 1 bp of the feature, has a length of 34 to 50 nucleotides, and ends with CCA.

```
$reclassify
$reclassify$tRNA
function(srcts, rules=list(
  "tRNA_internal_fragment"=function(x){ rep(TRUE, nrow(x)) },
  "tRNA_5p_fragment"=function(x){ x$startInFeature < 5L & x$length < 30L },
  "tRNA_3p_fragment"=function(x){ x$distanceToFeatureEnd < 5L & x$length < 50L },
  "tRNA_5p_half"=function(x){ x$startInFeature %in% -1:1 & x$length %in% 30:34 },
  "tRNA_3p_half"=function(x){ x$distanceToFeatureEnd %in% -1:1 & x$length >= 34L &
    x$length <= 50L & grepl("CCA$",x$seq) }
)) {
  valids <- vapply(rules, FUN.VALUE=logical(nrow(srcts)), FUN=function(fn){
    fn(srcts)
  })
```

```

if(nrow(srcs) == 1) valids <- t(valids)
factor(apply(valids, 1, FUN=function(x) max(which(x))),
       seq_len(ncol(valids)), colnames(valids))
}

```

We also assign priorities to different types of RNA. Priorities can be changed by increasing or decreasing the numbers; for example, in the case of only miRNA sequencing, users can set higher priorities for miRNAs by changing the number from 1 to 2 or higher.

\$priorities					
miRNA	tRNA	tRNAP	Mt_tRNA	snRNA	1
snoRNA	antisense	primary_piRNA	secondary_piRNA	precursor	
1	1	1	1	1	-1
long_RNA	longRNA				
-1	-1				

2.3.3.7 Features tree

We organized the features and reads in the form of a rooted phylogenetic tree in a hierarchical fashion. This is done by the function `assignReadsToTree()` in `shortRNA`, Figure 12. From the root, we have two main branches: shortRNA and long RNA, Figure 8C. We keep the features that are longer than 200 bp in long RNA and the features that are shorter than 200 bp in short RNA. Each branch is then further divided into RNA types. The miRNAs and tRNAs are further divided into sub-branches to account for the multi-mapping problems, Figure 9 and Figure 10. Sequence quality and a follow up trimming, may result in one nucleotide shorter sequence. Because of the hierarchical organization, unique reads mapping to a feature are kept in the hierarchy, Figure 8. If the user would like to include an additional database for short RNAs, an additional branch can be added to the tree while preparing annotation, `prepareAnnotation()`.

2.3.3.8 Reads assignment ambiguity

Reads can align to multiple locations in the genome, and when this is the case, typical workflows will either randomly align to one location or not report any alignment. We allow for multi-mapping during alignment, and address ambiguities when we assign reads to features. Specifically, we assign it to the parent of all the features to which the read maps. Figure 11A depicts a read that can be mapped to multiple tRNAs, Figure 11B. While assigning reads to the features tree, we assign this read as an ambiguous read to the parent of all the features it is mapping to; in this case, tRNA-Leu-CAG, Figure 11C.

2.3.3.9 Construction of TreeSummarizedExperiment object

The feature tree and the assigned reads DataFrame is then used for the construction of the TreeSummarizedExperiment object, which is then used for all the downstream analysis, including differential analysis. This is depicted in Figure 12.

2.3.3.10 Differential analysis

The hierarchical structure of the features implies that differential expression analysis could be performed at different levels of the hierarchy, which would however increase the multiple testing problem. We, therefore, rely on the method for dynamic testing of hierarchical hypotheses implemented in treeclimbR (R. Huang et al., 2021) for differential analysis of features.

Following normalization and differential analysis, users can use the TreeHeatmap, ggtree, and castor packages for exploratory data analysis and figure creation. The shortRNA makeTracks() function can be used to create genomic tracks, as shown in Figures 15, 19, and 20. We intend to expand our package by writing wrappers for these packages in order to generate plots from within shortRNA.

2.3.3.11 Table of important functions

Table 1 shows some of the most important functions of the shortRNA package. In the table, the functions are organized in the chronological order of data analysis.

Function name	Function description
generateQCreport	Perform quality check, quality control, provide trimmed files and generate an HTML report
fas tq2SeqCountMatrix	Generates a sequence by sample count matrix
getDB	Obtain databases for annotation of the desired genome
prepareAnnotation	Using the databases provided, create a features GRanges object and index the genome
alignShortRNA	Align the sequences with custom genome index built using prepareAnnotation
overlapWithTx2	Find overlaps between read alignments and a feature annotation
assignReads	Assign overlapped reads from overlapWithTx2 as per the assignment rules
d efaultAssignRules	Display default reads assignment rules
featuresAnnoToFL	Convert the features GRanges object to a FactorList

Function name	Function description
addReadsToTree	Assign reads to the FactorList of features as per output of assignReads
makeTracks	Make a genomic track of aligned data for a given genomic location

2.4 Results

2.4.1 Datasets used for testing the pipeline

We used a previously published sperm sRNA-seq dataset from (Katharina Gapp et al., 2021) (GEO accession: GSE162112), in which we additionally spiked simulated reads. We also used an unpublished PBMC human dataset from the Schratt Lab (<https://schrattlab.ethz.ch/>) at the ETH Zurich, on which qPCR were additionally performed for 5 miRNAs (miR-30b-5p, miR-30e-5p, miR-30d-5p, miR-499-5p, and miR-1248). We used these human data to correlate quantifications from shortRNA, seqPAC, Sports1, and Oasis2 with delta Ct values from qPCR. We analyzed both of these datasets with shortRNA, seqPAC, Sports1, and Oasis2 tools to compare the quantification of miRNAs. Further, we used the simulated reads from miRNAs to test the reads assignment and to assess differential analysis.

For both datasets, we used miRBase (miRNA), GtRNADB (tRNA), MttrNADb (mitochondrial tRNA), SiLVA (rRNA), and Ensemble. For the sperm dataset from the mouse, we also used piRNA precursors.

2.4.2 Comparison of quantification and identification of miRNAs

2.4.2.1 shortRNA quantification is positively correlated with Oasis2 and Sports1

We used SPORTS1, seqpac, Oasis2, and shortRNA with the default parameters to analyze the sperm dataset. Following the data analysis, we limited our assessment to miRNAs for comparison across tools. We found 1970 miRNAs using Oasis2, 1137 using Sports1, 1024 using seqpac, and 503 using shortRNA. We began by examining the quantification with all four tools. Figure 13 shows a comparison of the quantification of miRNAs from SPORTS1, Oasis2, and seqpac with shortRNA. We discovered that the quantifications of shortRNA and Oasis2, as well as shortRNA and Sports1, are highly correlated. In contrast, the correlation of quantification between shortRNA and seqpac was negative. We investigated this discrepancy and discovered a large number of short reads (length: less than 15 nucleotides) mapping to miRNAs and assigned by seqpac, but which were considered invalid overlaps in light of our

assignment rules.

2.4.2.2 shortRNA accurately identify miRNAs

Following that, we examined the overlaps of identified miRNAs between four tools. Despite the fact that all of the tools use miRBase, we discovered that only 44 miRNAs were commonly detected by all of the tools (Figure 14). One possible explanation is that the miRBase for mouse genome contains approximately 2000 mature miRNAs and approximately 1300 miRNA precursors, and seqpac and Sports1 only use miRNA precursors in the data analysis pipeline, whereas Oasis2 and shortRNA use both mature and miRNA precursors. Figure 14 shows that there were 231 miRNAs that were commonly detected by Oasis2, Sports1, and seqpac, but not by shortRNA. Next, we looked into these miRNAs and discovered that three were removed from the most recent version of miRBase; for the other 228, we did not find alignment in the genome using shortRNA, despite accepting mismatches and soft-clipping. Based on our findings, we believe that these miRNAs are the result of misalignment or sequential alignment by the other three tools. When we examined the sequences mapping to these miRNAs according to seqpac, we discovered that all of the reads mapping to these miRNAs were shorter than 15 bp, and as a result, are very unlikely to represent functional miRNAs. In conclusion, we discovered that shortRNA can accurately identify miRNAs.

2.4.3 Data simulation to check the alignment and reads assignment by shortRNA

We validated the alignment and reads assignment by spiking the reads in the sperm dataset (Appendix 3). In brief, we simulated reads for a miRNA cluster containing mmu-miR-125a-5p, mmu-miR-99b-5p, mmu-let-7e-5p, mmu-let-7e-3p, mmu-miR-99b-3p, and mmu-miR-125a-3p, as well as reads from a miRNA precursor, mmu-miR-144. To test for differential analysis, the sequences from these miRNAs were spiked while also creating differences between groups. Three invalid sequences were also simulated for the precursor to test the accuracy of read assignment and were properly aligned to miRNA precursors, as shown in Figure 15.

Further, the reads were assigned accurately to miRNAs, Table 2.

Sequence	Ov erlap	Transcript ID	Transcript type	Tra nscrip t		Valid miRNA
				l en g th	val id	
AAGTTTG	21	MI0000168	miRNA_precursor	66	TRUE	FALSE
TGATGA-						
GACACTAC						
ACAGGTGA	22	miR-125a-3p	miRNA	22	TRUE	TRUE
GGTTCTTGGGAGCC						
CAAGCTCG	22	miR-99b-3p	miRNA	22	TRUE	TRUE
TGTCT-						
GTGGGTCCG						
CACCCGTA	22	miR-99b-5p	miRNA	22	TRUE	TRUE
GAACC-						
GAC-						
CTTGGC						
CTATACGG	22	let-7e-3p	miRNA	22	TRUE	TRUE
CCTCC-						
TAGCTTC						
CTGTAA	20	MI0000168	miRNA_precursor	66	TRUE	FALSE
GTGGT-						
GATGA-						
GAC						
GG	23	miR-144-5p	miRNA	23	TRUE	TRUE
ATATCAT						
CATATAC						
TGTAAGT						
GGCTGGGA	50	MI0000168	miRNA_precursor	66	TRUE	FALSE
TATCAT-						
CATATACT						
GTAAGTTTGT-						
GATG						
AGACACT						
ACAGTAT						
TACAGT	20	miR-144-3p	miRNA	20	TRUE	TRUE
ATAGATG						
ATGTACT						

Sequence	Ov erlap	Transcript ID	Transcript type	Tra nscrip t length	Valid	miRNA
TCCCTGAGAC24		miR-125a-5p	miRNA	24	TRUE	TRUE
CCTT-						
TAACCT-						
GTGA						
TGAGGTAG	22	let-7e-5p	miRNA	22	TRUE	
GAGGTTG-						
TATAGTT						

2.4.4 Differential expression analysis

We performed differential analysis using the treeClimbR package. Through differential analysis, candidate proposal, multiple testing correction, and candidate evaluation, treeclimbR integrates the observations with a tree that reflects the hierarchical relationship between entities and finds an appropriate resolution on the tree to interpret the association. During simulations, we spiked the mature miRNAs miR-99b-5p and miR-99b-3p to be different. After differential analysis, because both miR-99b-5p and miR-99b-3p were simulated to be different, we found that miR-99b was called to be differentially expressed, Figure 17. Looking at the cluster, miRNACluster_17:17830188-17830879, the full cluster would have been differentially expressed if one read mapping to miR-125a-5p, in grey, would also have been significantly different.

2.4.5 Validation using human data and quantitative real-time PCR

To validate the Oasis2, Sports1, seqpac, and shortRNA quantifications, we used an unpublished PBMC human dataset and corresponding quantitative real-time PCR (qPCR) quantifications. The PBMC dataset was analyzed using the default parameters from Oasis2, Sports1, seqpac, and shortRNA. Then, we correlated the qPCR quantification of five miRNAs with the quantification using all four tools, Figure 18. Because Sports1 and seqpac only report miRNA quantification at the level of precursors, the precursor of mature miRNAs quantification was used for the correlation of these tools with the qPCR. When we looked at the overall correlation of miRNAs with qPCR quantifications, we discovered that the _____ tool had a higher Pearson correlation of 90% ($R=.9$), Figure 18A, and a more robust correlation of _____. Overall, we found a low correlation of quantifications between four tools and qPCR. We

found that Oasis2 quantification was negatively correlated with qPCR Ct values for miRNAs miR-30d-5p and miR-30e-5p, as shown in Figure 18B. The quantification of miRNAs using short RNAs was found to be positively correlated with the quantification of miRNAs using qPCR. In addition, we performed a visual inspection of the reads mapping to these miRNAs by creating genomic tracks, as shown in Figures 19 and 20. As a result, while shortRNA quantifications were visually observed to correctly count reads, they were not substantially correlated with qPCR, but were generally better correlated with qPCR quantifications than the other methods.

2.4.6 Qualitative comparison with other tools

There are several tools published for analyzing the sRNA-seq data. In Table 2, we made a qualitative comparison of 20 other tools.

2.5 Discussion & Outlook

Due to the non-unique genomic origin, short length, and numerous post-transcriptional modifications of short RNA species, data processing of sRNA-Seq has proven to be difficult. Moreover, from a bioinformatics point of view, there are several challenges in analyzing data generated from small RNA sequencing techniques, including alignment, reads assignment, normalization, and differential analysis. Importantly, quantifying small RNAs correctly is a computational challenge. To a large extent, different tools provide varying quantification. This is startling, and it highlights the need for further research and better tools in the field.

We present shortRNA, an R package, a novel tool, for analyzing sRNA-seq data, which can reduce the likelihood of false discoveries from the sRNA-seq data. shortRNA is available from github.com/mansuylab/shortRNA and is planned to be submitted to the Bioconductor. The shortRNA R package is developed around the existing frameworks of Bioconductor: phylo, FactorList, DataFrame, and TreeSummarizedExperiment. Hence, it is expected to be able to interact seamlessly with many other tools, such as ggtree, castor, and TreeHeatmap. shortRNA performs preprocessing, alignment, and downstream analysis of the sRNA-seq data. Moreover, shortRNA is able to deal with the sequencing data with UMIs using UMIc. Further, if an adapter sequence is unknown, it is possible to detect the adapter sequence using shortRNA.

We tested the package on various human and mouse datasets and demonstrated it using the mouse sperm dataset, simulated mouse sperm dataset, and human dataset from PBMCs. We compared the quantifications of miRNAs from shortRNA to those of three other

tools: Sports1, Oasis2, and seqpac and found them to be well correlated with Sports1 and Oasis2. We demonstrated shortRNA’s ability to assign reads to features in a meaningful way by customisable assignment rules. shortRNA uses a considerable proportion of reads, multi-mapping reads, that are either discarded by other methods or aligned randomly. We demonstrated that shortRNA correctly annotates reads to features using simulated reads from the sperm dataset. In addition, we demonstrated that shortRNA correctly identifies differentially expressed reads. Finally, using the human PBMC data and corresponding qPCR quantifications from five mature miRNAs, we demonstrated that shortRNA quantification is better correlated with qPCR data than the other three tools.

In general, however, we observed a very low correlation of all methods against qPCR in terms of the relative miRNA expression across samples. There are a number of potential explanations for this discrepancy. First, the lack of normalization of miRNA quantifications with U6 expression. In several investigations, including in this dataset, RNU6B (U6) is used to normalize circulating miRNA data (Donati, Ciuffi, & Brandi, 2019). Although qPCR is used to measure mature miRNA levels, it’s great sensitivity necessitates proper normalization to account for non-biological variance. This is naturally different from the RNA sequencing analysis, where the normalization is based on the whole distribution of RNAs. This discrepancy could therefore be one reason for the low correlations. On the contrary, it has been advised that U6 should not be utilized for data normalization of circulating miRNAs (Benz et al., 2013; Donati et al., 2019). Next, the lack of UMIs in the data. UMIs make it possible to eliminate PCR amplification biases(Y. Fu et al., 2018). The lack of UMIs in human PBMC data might have resulted in technical duplication, and hence affecting the correlation with the qPCR data.

Most tools for the sRNA-seq data analysis, including Sports1 and Oasis2, use feature-based counting, whereas seqpac uses sequence-based counting for sRNA-seq data analysis. The benefit of sequence-based counting is that it preserves sequence integrity and allows users to BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) candidate sequences to confirm classification and obtain additional information. In addition, sequence-based counting allows for the examination of sequence-specific variations or their boundaries. Further, when using sequence-based counting, users can choose to remove sequences that have low evidence and do not replicate across samples in their data. ShortRNA, on the other hand, has the advantage of both sequence-based and feature-based counting due to the hierarchical organization of features and mapping sequences. Because of the short lengths of the reads and their proclivity to originate from higher-copy number regions of the genome, multi-mapping reads are much more common in sRNA-seq data. Commonly used sRNA-seq alignment methods for multi-mapping reads are either very low precision (choosing an alignment at random and

make it difficult to determine their true origins) or sensibility (ignoring multi-mapping reads, which result in losing a high proportion of reads). ShortRNA allows these reads to be aligned multiple times and then assign them, using a set of customisable rules, either to a specific feature or to a position in the feature tree which makes it usable while preserving knowledge of eventual ambiguities.

shortRNA makes use of miRBase, GtRNADb, MtTRNADb, SILVA and Ensembl databases, but it is possible to manually add databases and extend the list. We used adaptable assignment rules to identify the correct assignment of a read. For example, if the data comes only from miRNAs, the user can set higher priority for mapping the reads with miRNAs by changing the number from 1 to 2 or higher. Further, we used a phylogenetic tree structure to organize the features and to assign the reads to features. If a sequence is overlapped with multiple features, we assign the read to the parent of both features. This helps in solving the problem of multi-mapping.

2.6 Limitations

shortRNA has not been thoroughly tested across multiple platforms. Although we anticipate that the shortRNA will be platform independent, it has not yet been tested on other operating systems. We intend to do so at a later date. In addition, we also intend to extensively benchmark shortRNA with existing tools and strategies for reads assignment.

Simulations for other types of RNA are necessary to benchmark the read assignment. We have only simulated reads from miRNAs and miRNA precursors and demonstrated the proper assignment of reads using shortRNA. Simulations of reads mapping to tRNAs, reads mapping to piRNAs and reads mapping to rRNAs will be performed to better assess the reads assignment.

The ambiguous assignment of multi-mapping reads to the parent of all mappable features is not always biologically meaningful. If a read is assigned to multiple tRNAs, Figure 11, it is still meaningful to assign it to the parent of all mapping tRNAs. In contrast, if a read maps to both a tRNA and a protein-coding gene, assigning it to the parent of both protein coding and tRNA provides less biological insight. Hence, we plan to compare shortRNA to other read assignment strategies for multi-mapping reads in depth. Other strategies for assigning multi-mapping reads exist (Handzlik, Tatsoglou, Vlachos, & Hatzigeorgiou, 2020; Johnson, Yeoh, Coruh, & Axtell, 2016), and a benchmark with simulations would be required to assess the shortRNA assignment of multi-mapping reads. Instead of assigning reads to the parent of all mappable features, this would aid in the proper assignment and provide more biological insight for the reads mapping to multiple biotypes.

The current implementation of UMIc for dealing with sRNA-seq data based on UMIs is extremely slow. There are several for-loops for running analysis, as well as a non-parallel implementation of the code by the authors of UMIc. We intend to optimize the code for speed by parallelizing it and replacing for-loops with parallel apply functions.

shortRNA has limited user-friendly plotting capabilities. Users can currently use data visualization packages developed for TreeSummarizedExperiment, such as TreeHeatmap, directly. However, we intend to create wrappers for other packages in order to facilitate the visualization of data and results. Also, we will be creating a Shiny app for exploratory analysis of the data and results.

IsomiRs analysis could be implemented in shortRNA. IsomiRs are miRNAs that vary slightly in sequence, which result from variations in the cleavage site during miRNA biogenesis (5'-trimming and 3'-trimming variants), nucleotide additions to the 3'-end of the mature miRNA (3'-addition variants) and nucleotide modifications (substitution variants). While in principle our hierarchical and sequence-based approach is ideal for their investigation, tailored functions would make this more straightforward for the users. We plan to extend our package to include isoMiRs analysis by adapting functionalities of the isomiR R package (Lorena Pantano [Aut, Cre], Georgia Escaramis [Aut], ChristosArgyropoulos [Aut], 2017).

shortRNA is developed to work with single-end sRNA-seq data. We plan to extend our package to work with paired-end sRNA-seq data.

2.7 Conclusion

In conclusion, we developed an R package, platform-independent, for analysis of the sRNA-seq data. This package provides a great overview of the quality, performs QC, aligns the data to custom made genome, performs reads annotation and assignment with adaptable assignment rules, organizes reads and features as a phylogenetic tree, performs differential analysis, and provides interactive plots for exploratory data analysis. shortRNA currently works with single-end sequencing data.

2.8 Data and code availability

shortRNA is available from: github.com/mansuylab/shortRNA

No sequencing data were generated during this study. Code for simulating reads is presented in Appendix 4.

Data analysis steps for sRNA-seq data with shortRNA is available in Appendix 3.

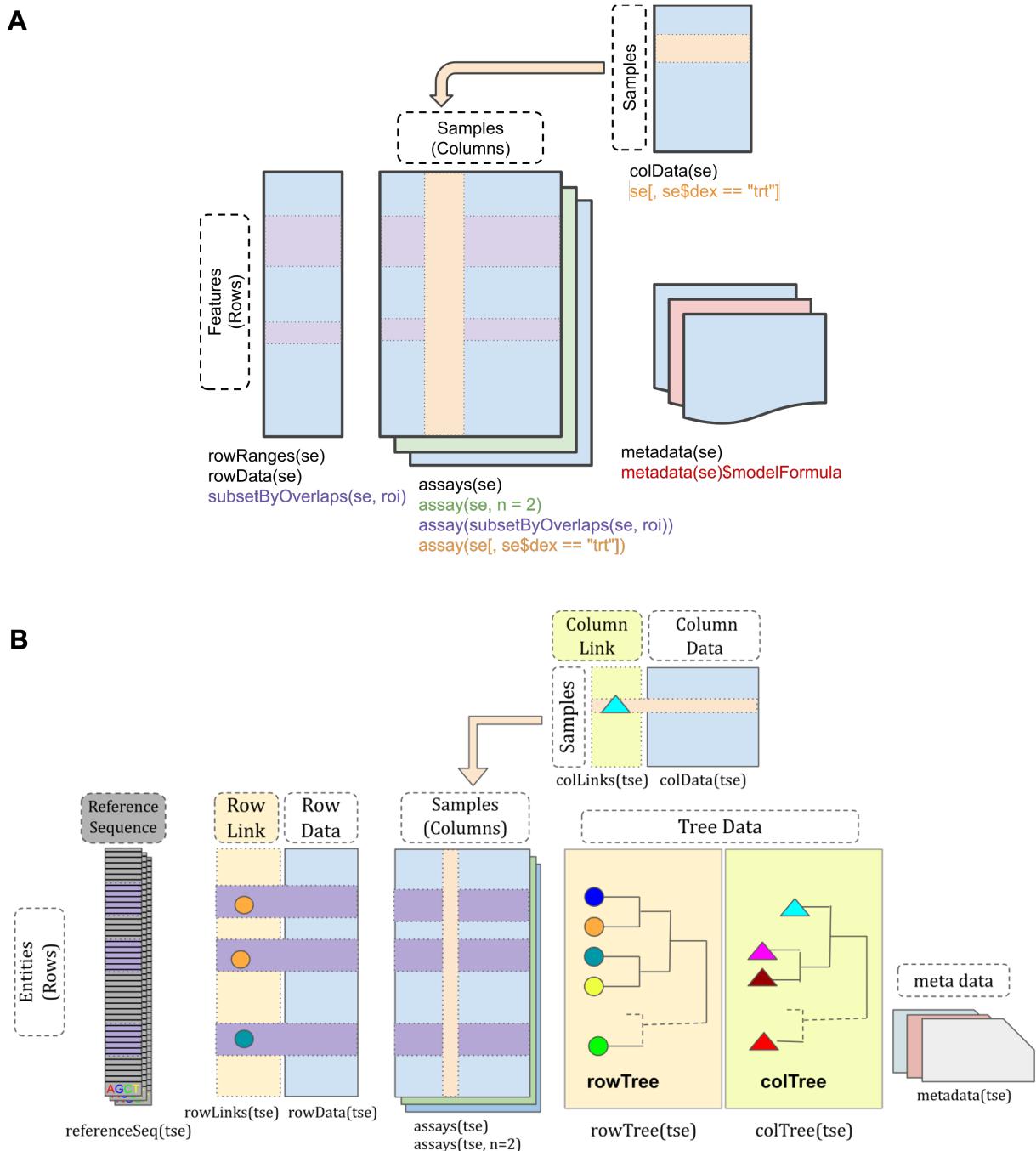


Figure 2.1: Schematic of A: `SummarizedExperiment` (Morgan et al., 2020),
B: `TreeSummarizedExperiment` (R. Huang et al., 2020)

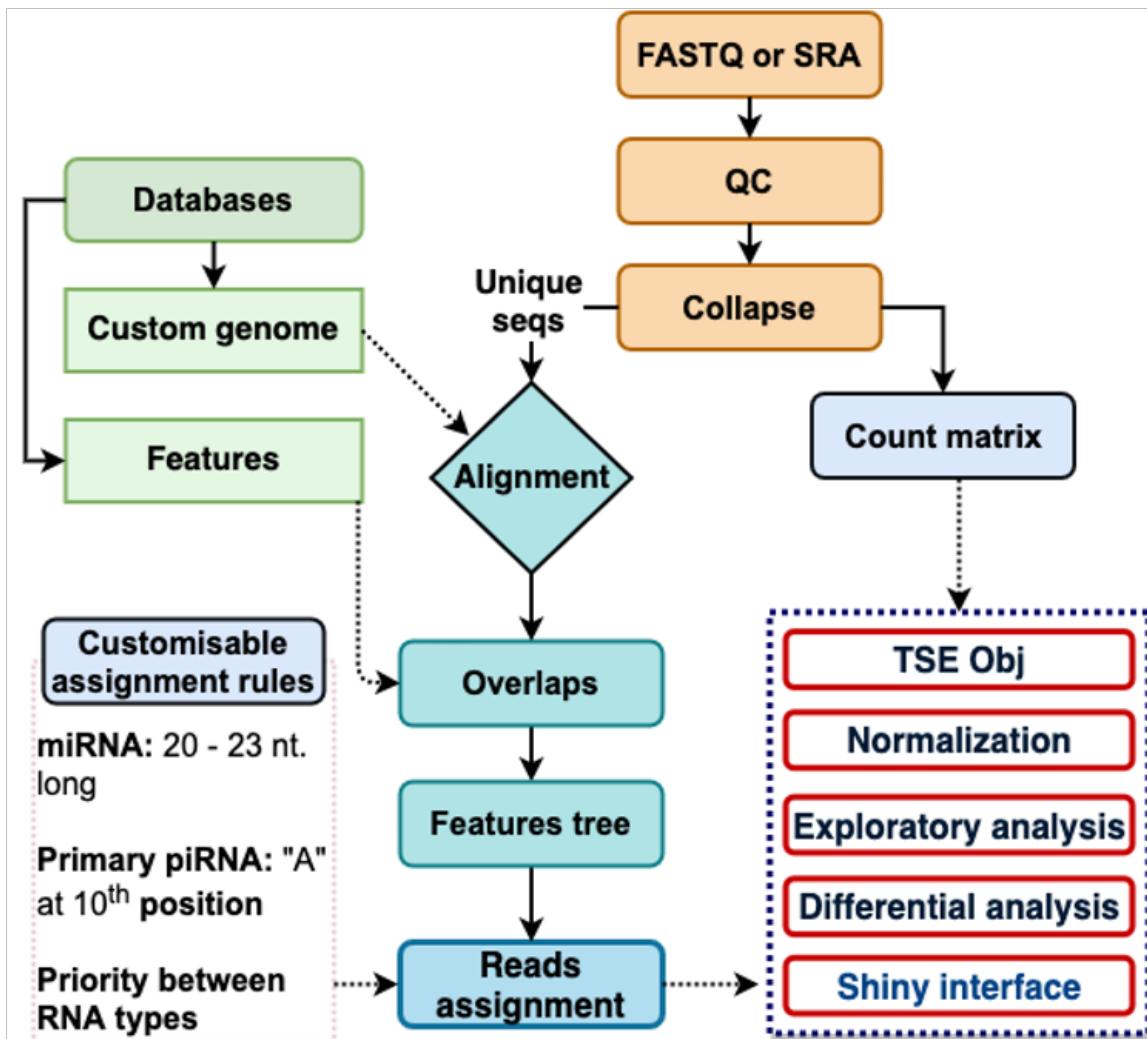


Figure 2.2: Overview of shortRNA pipeline

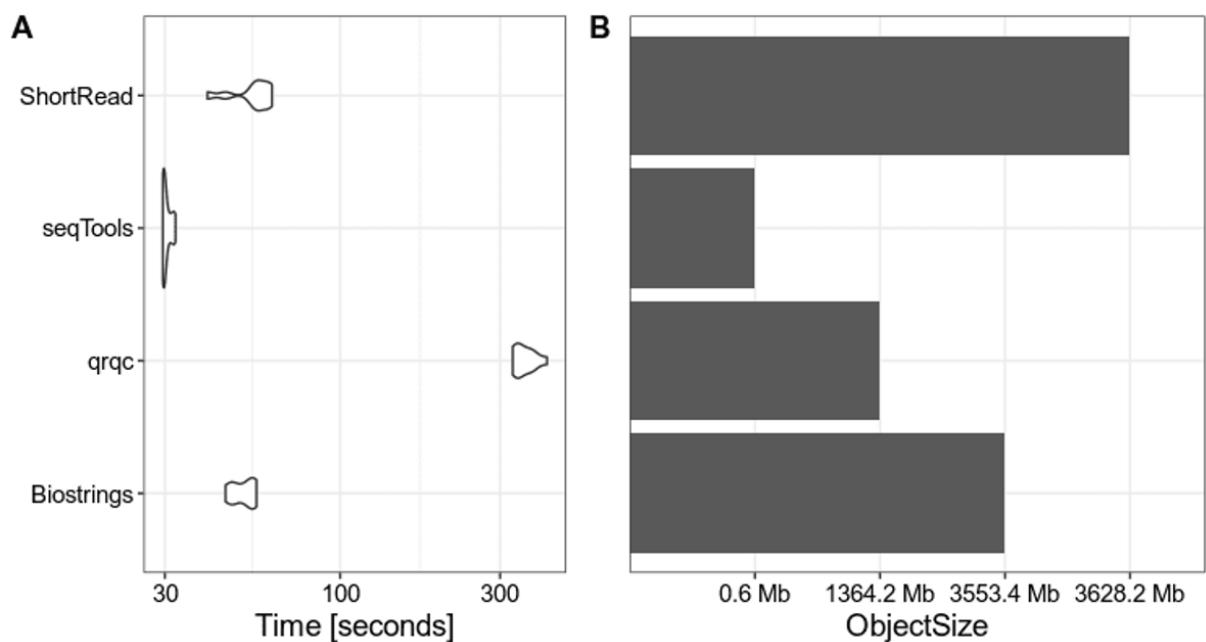


Figure 2.3: Benchmark of four different Bioconductor tools for reading and storing FASTQ file quality data. Performance of four tools is measured in **A**: the amount of time they take for reading the file, **B**: the amount of memory they take to store the data.

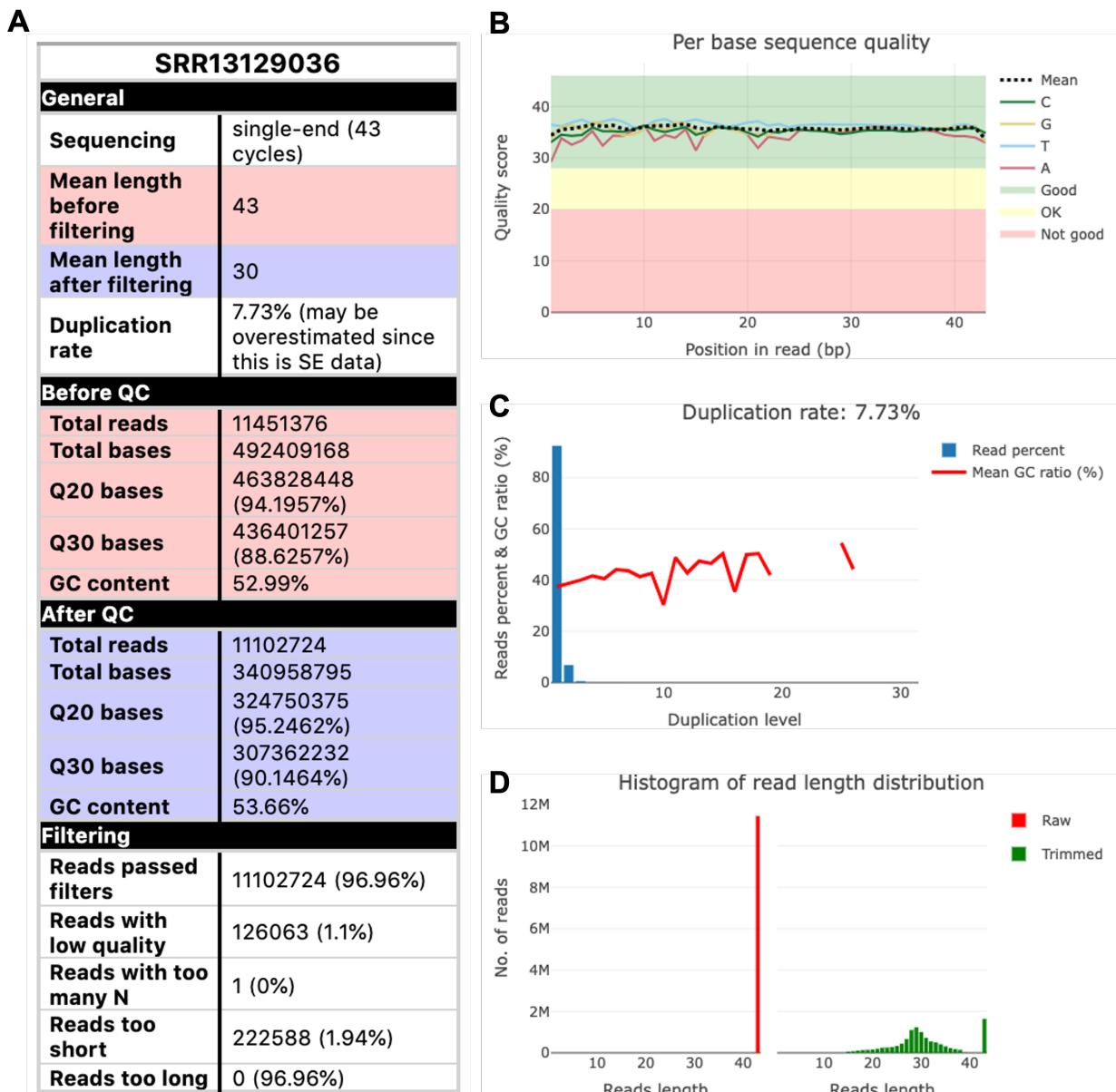


Figure 2.4: Exemplary QC report plots of a sperm sample from (Katharina Gapp et al., 2021). **A:** Quality control table, **B:** Quality score lines plot at each nucleotide in the sequencing reads, **C:** Barplot of reads duplication level with the mean GC ratio represented as lines, **D:** Histogram of reads distribution, before and after quality control

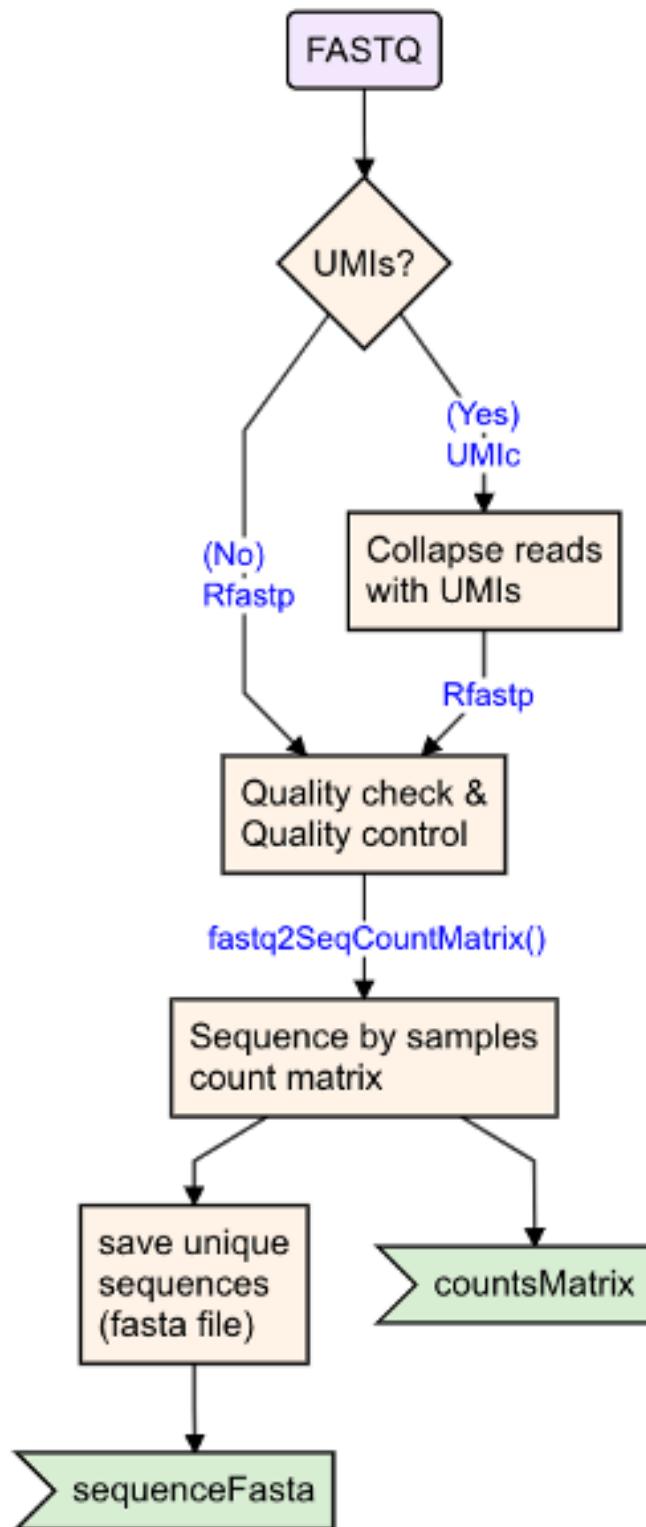


Figure 2.5: Schematic of quality check and quality control of FASTQ files, UMIs collapsing (if present), and collapsing FASTQ files to obtain unique sequences for alignment and making a sequence by samples counts matrix

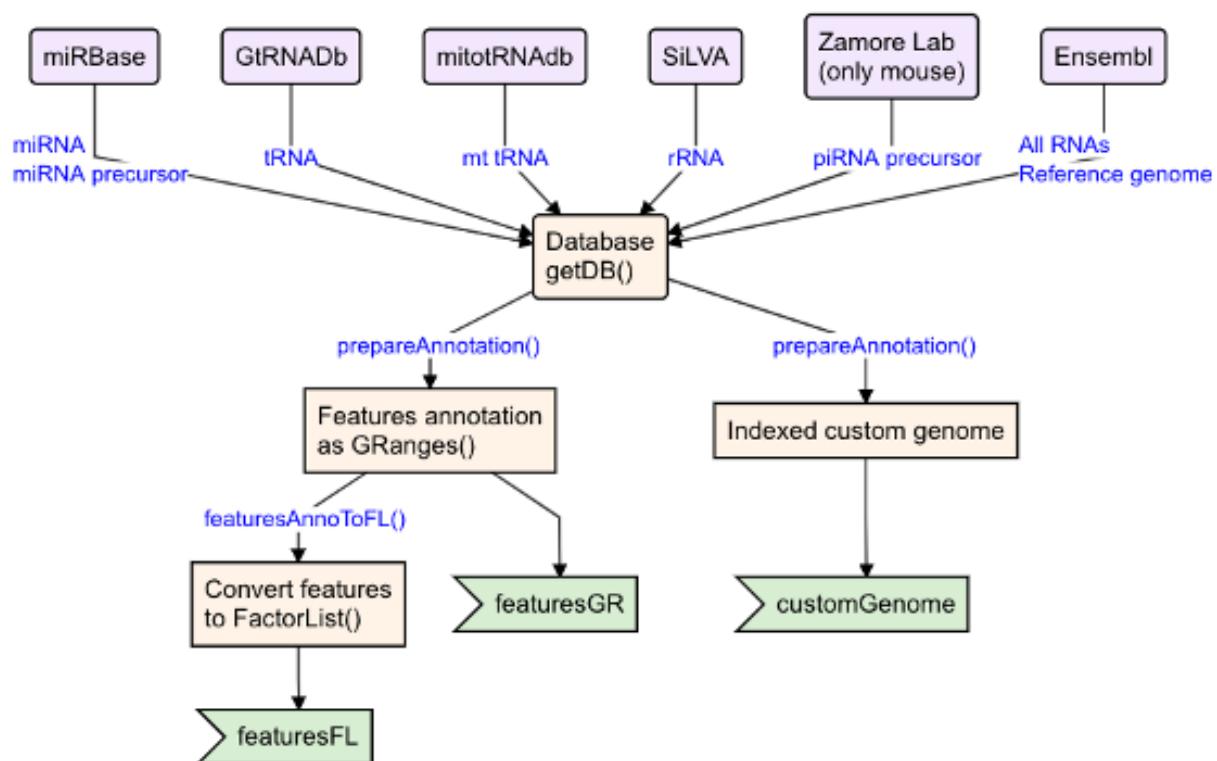


Figure 2.6: Schematic of obtaining database, preparing annotations and custom genome indexes for alignment

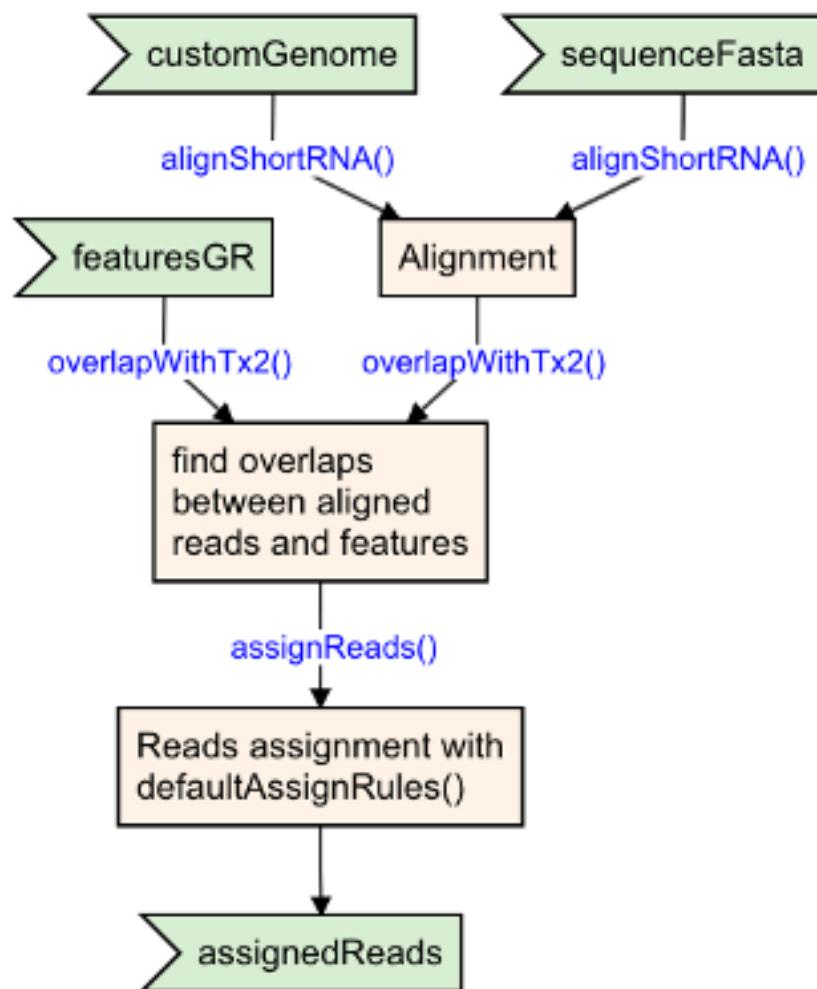


Figure 2.7: Schematic of alignment, overlapping features and aligned sequences, and assigning reads to features using assignment rules

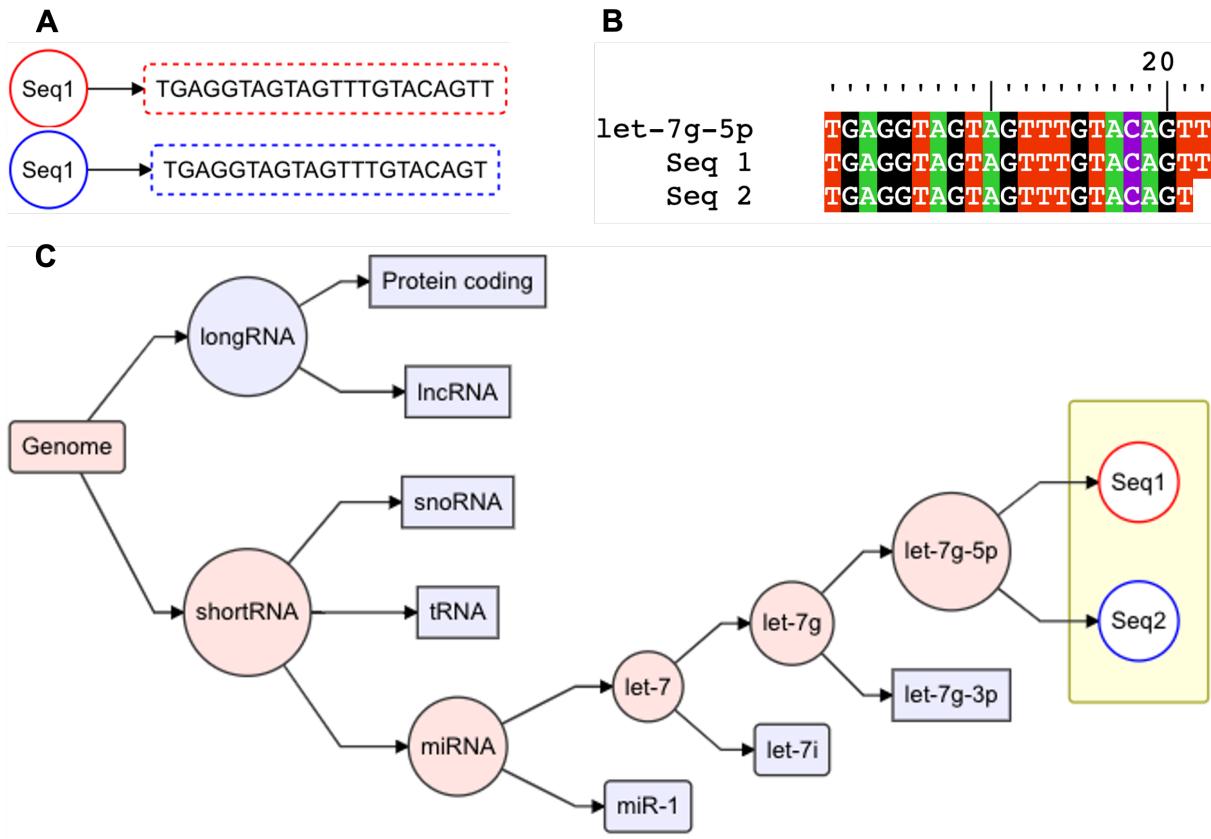


Figure 2.8: General framework of feature tree and assignment of sequences to the tree. **A:** Example of two sequences. **B:** Alignment of both the sequences to miRNA. **C:** General outline of the feature tree and assignment of both the sequences to the tree

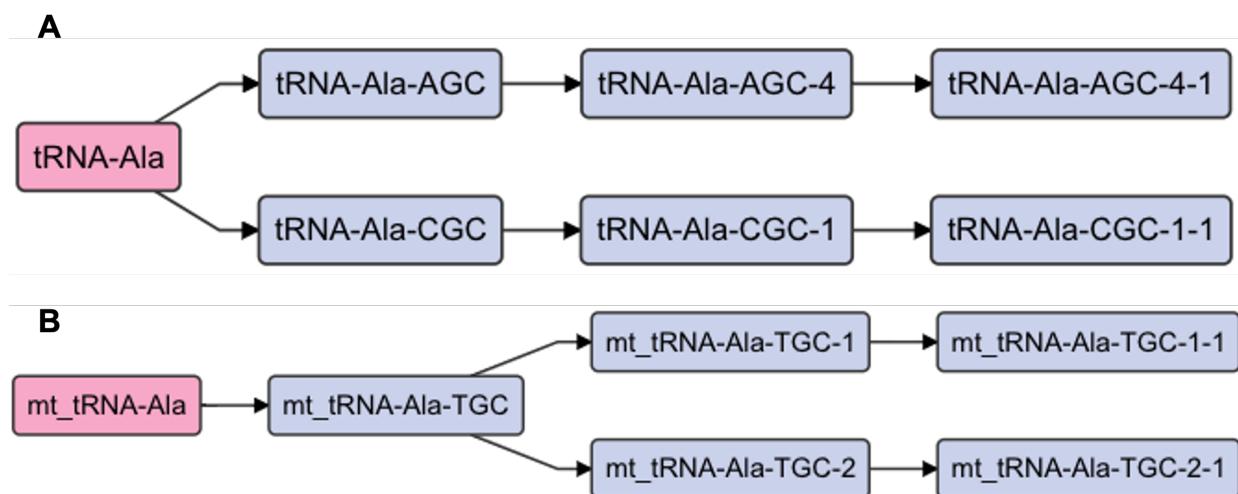


Figure 2.9: Hierarchical organization of **A:** tRNA features and **B:** mitochondrial tRNA features.

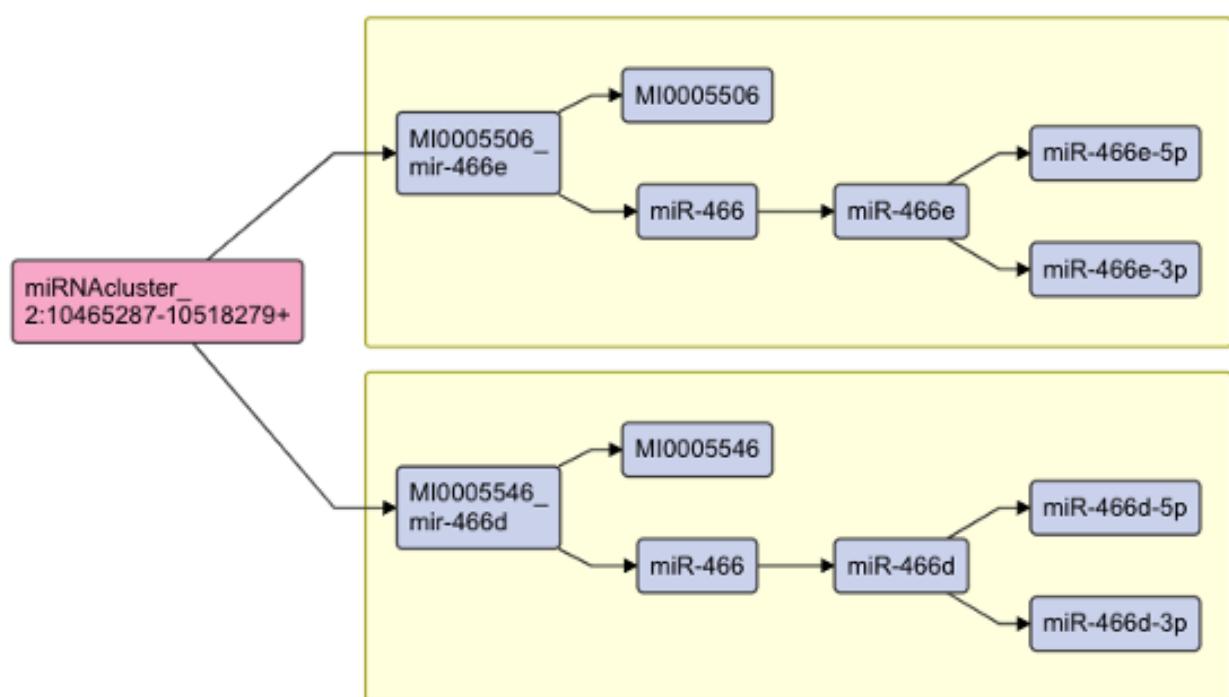


Figure 2.10: Hierarchical organization of miRNA features

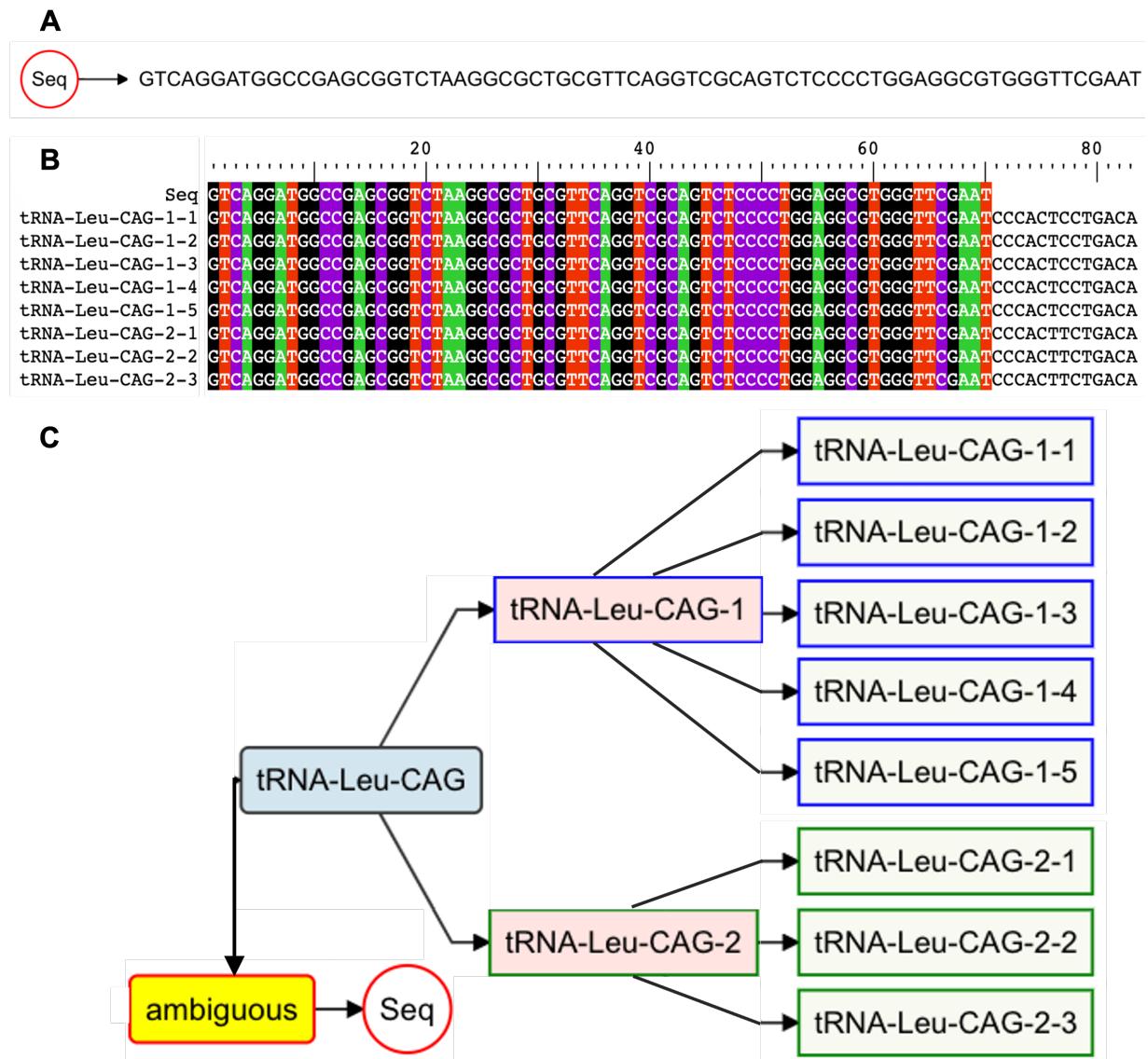


Figure 2.11: Multiple assignments of a sequence mapping to a tRNA. **A:** Exemplary sequence. **B:** Multiple sequence alignment plot of the sequence and the features to which it is mapping. **C:** Assignment of the reads to the tRNA feature tree

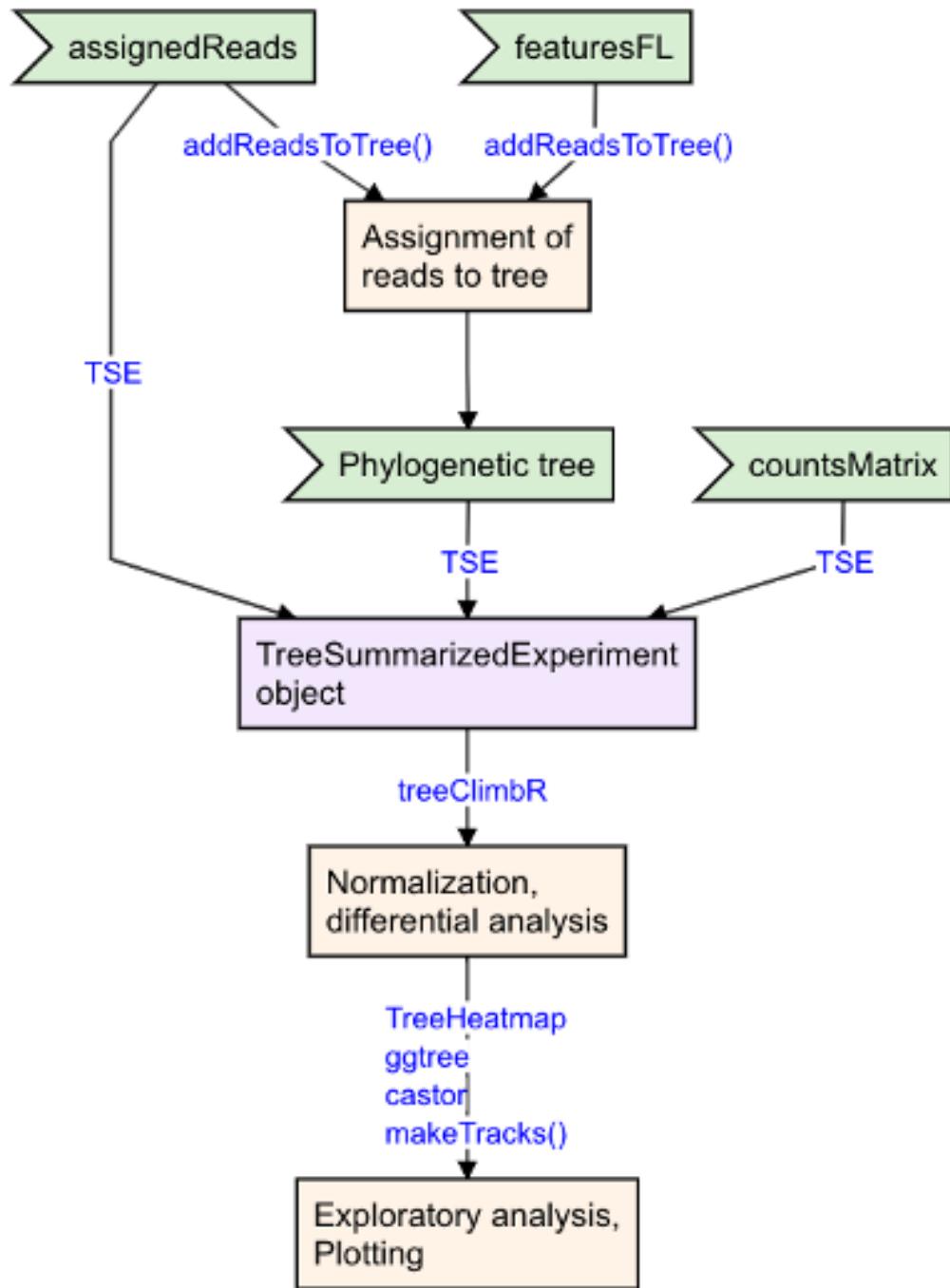


Figure 2.12: Schematic of assigning reads to the features tree of features, creation of TreeSummarizedExperiment object, and downstream analysis and plotting

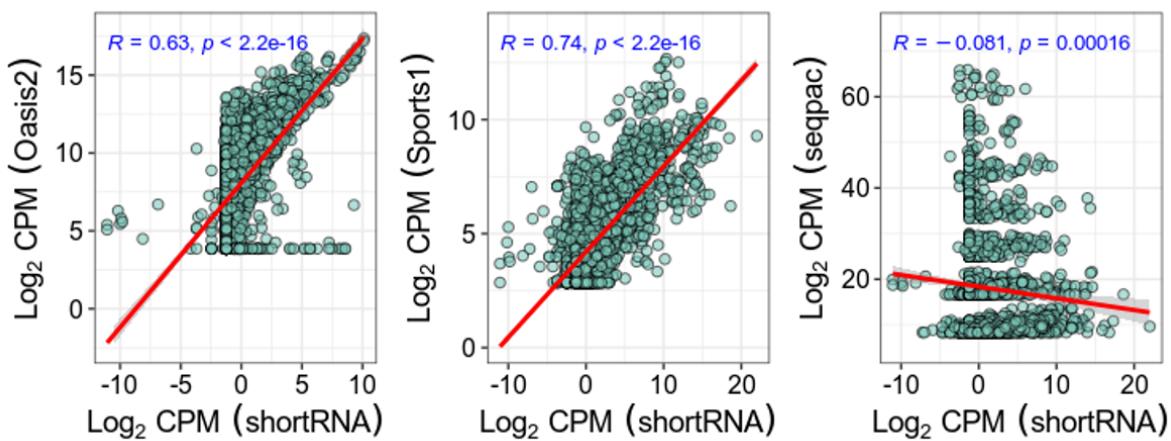


Figure 2.13: Comparison of Log2 TMM normalized quantification of miRNAs from shortRNA with Oasis2, Sports1, and seqpac

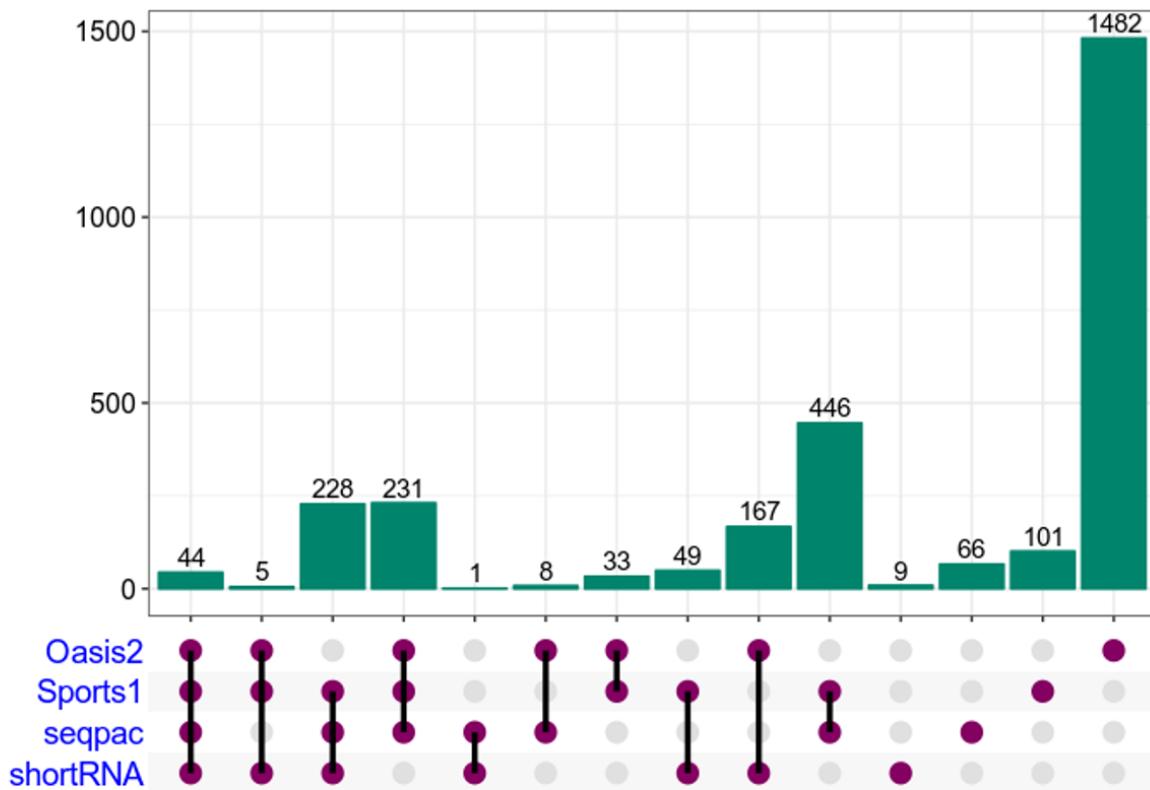


Figure 2.14: Upset plot of overlaps of detected features between four methods: shortRNA, Oasis2, Sports1, and seqpac

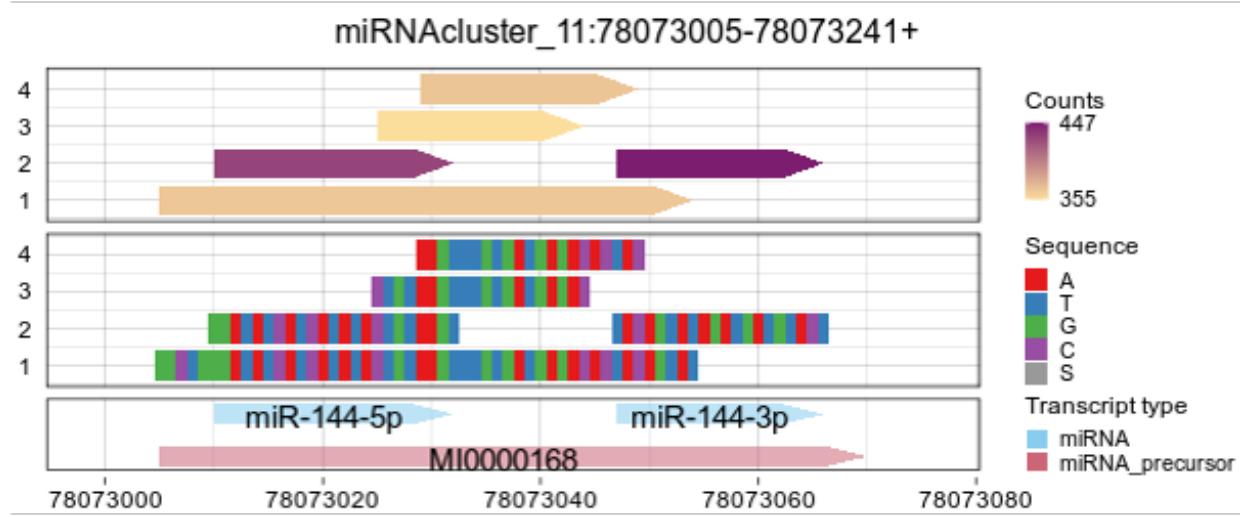


Figure 2.15: Genomic plots of simulated reads from miRNA precursor.

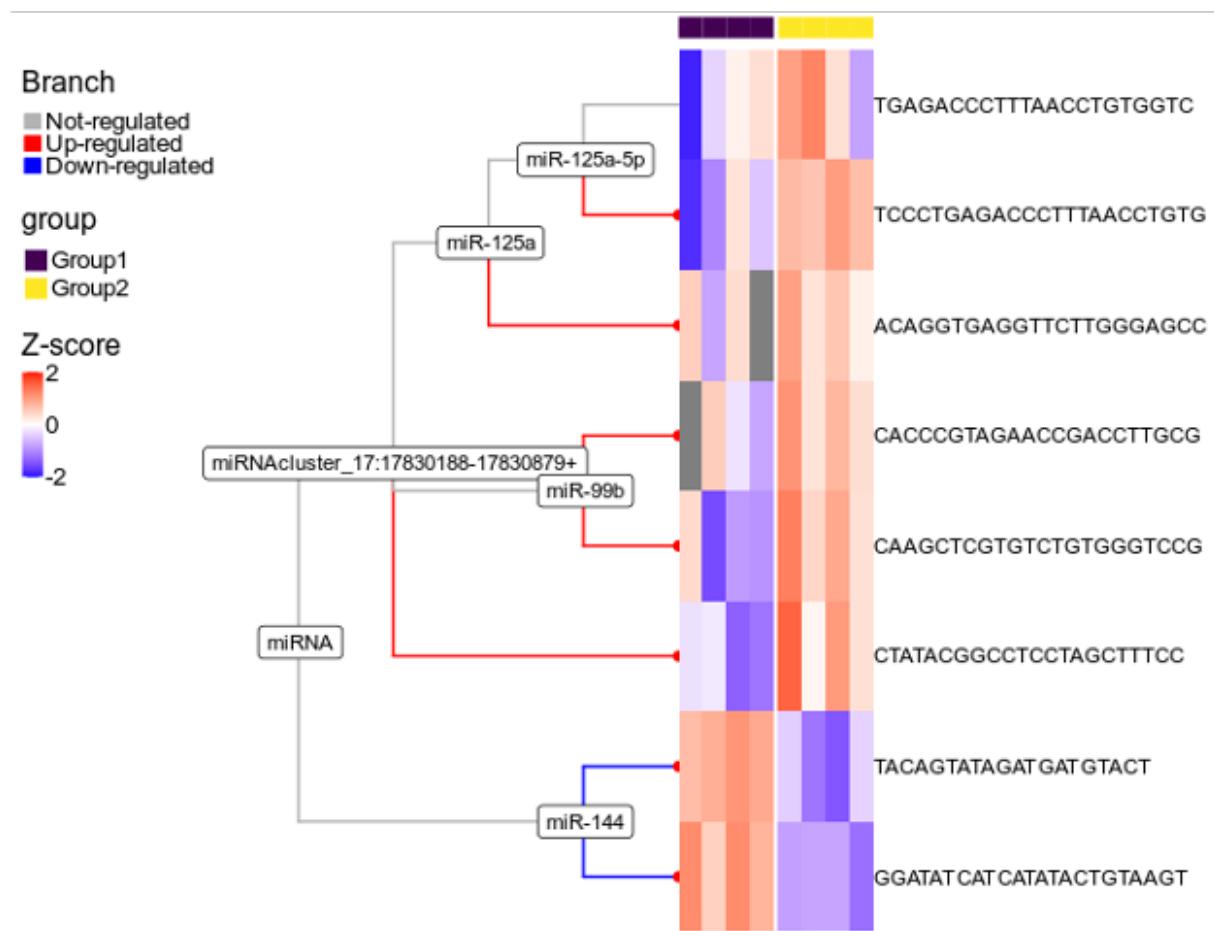


Figure 2.16: Hierarchical representation of features and assigned reads to them. Differential expression analysis of features from two miRNA clusters. Simulated read nodes are highlighted in orange colour

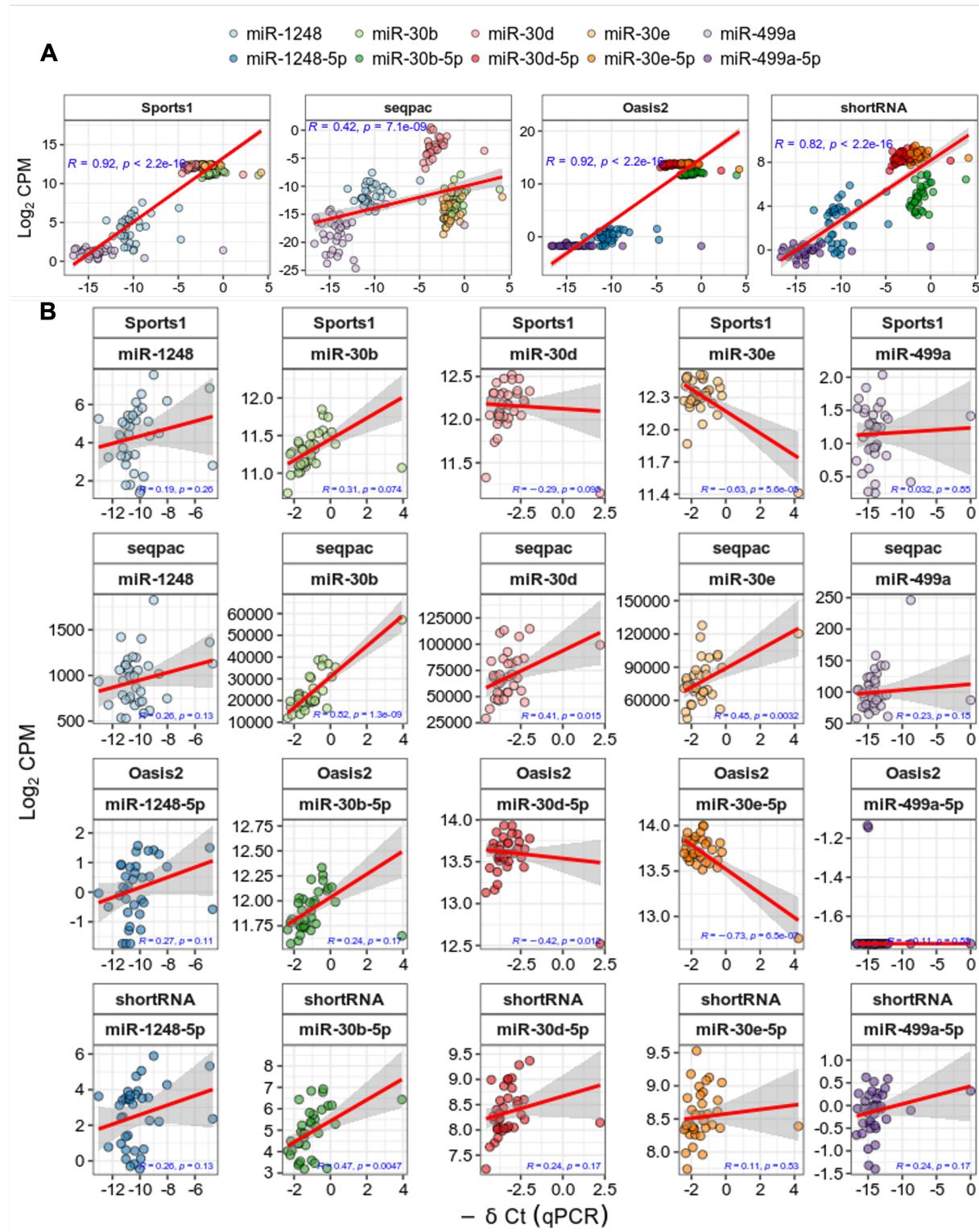


Figure 2.17: Comparison of quantifications of five miRNAs from four different tools with -Delta Ct values from qPCR. **A:** Overall correlation. **B:** Correlation at individual miRNA

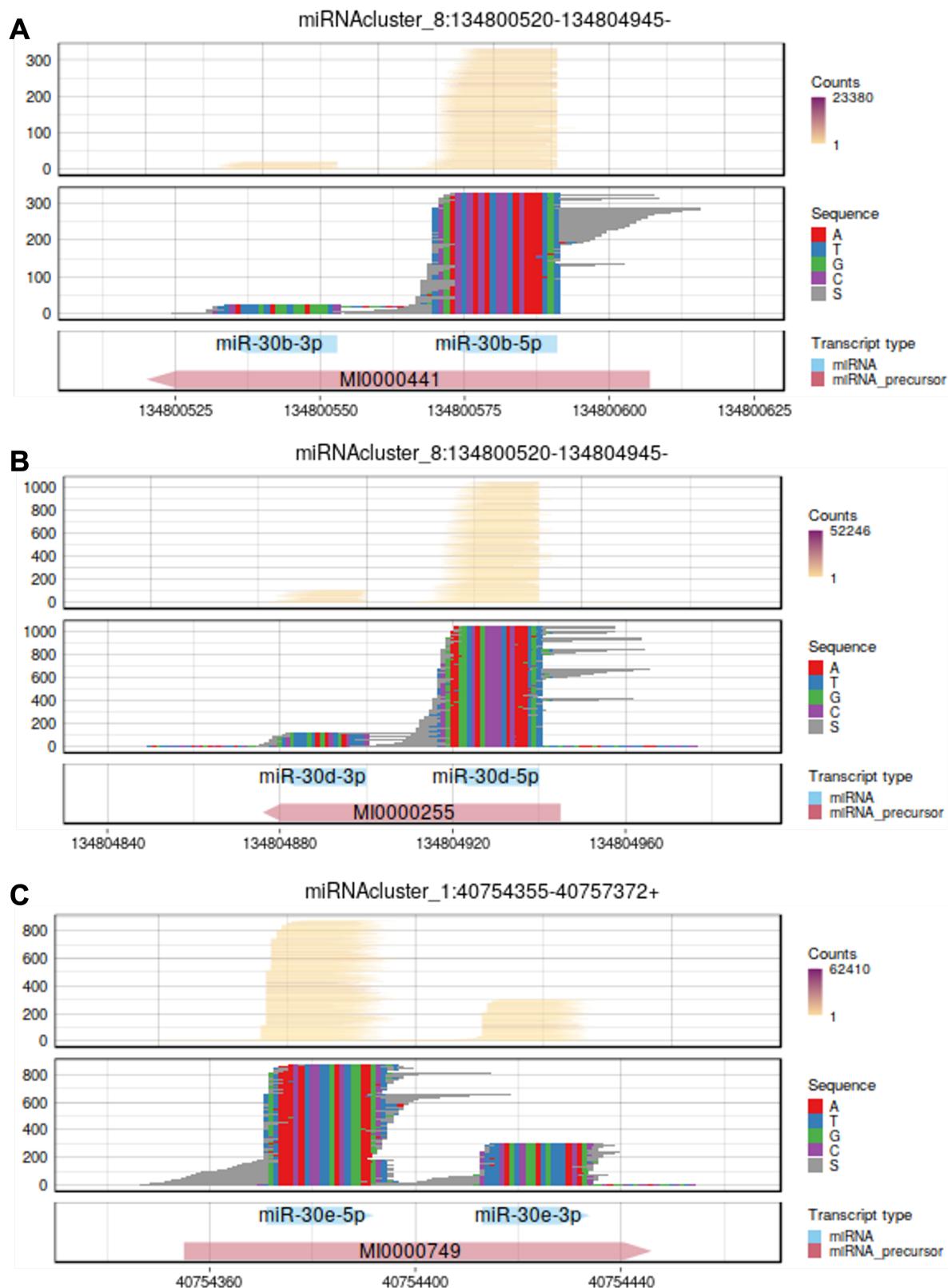


Figure 2.18a: Genomic plots of validated miRNAs

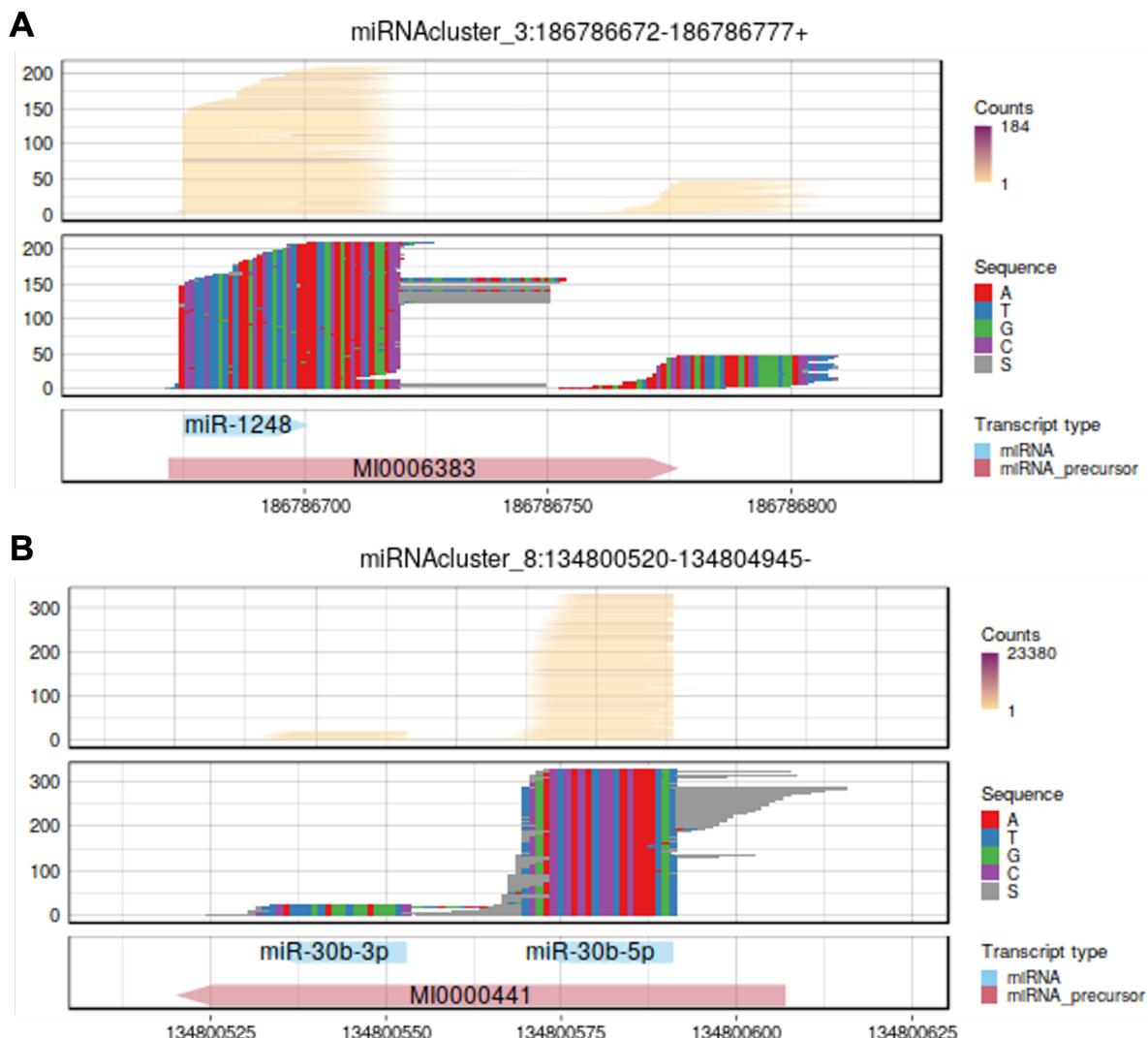


Figure 2.18b: Genomic plots of validated miRNAs

Conclusion and outlook

2.9 Overall conclusion

In this Doctoral thesis, first, the work is presented on the development of pipelines for multi-omics high-throughput sequencing data, Methods, and data analysis pipelines have been used to analyze multi-omics data from the lab and literature, Appendix 1. Next as summarized in Chapter 1, we compare the open chromatin landscape of developing and adult spermatogonial cells to explore the development of spermatogonial cells in postnatal development. We discovered and characterized genomic regions with differential genomic activity in spermatogonial cells by combining the data generated by the host lab with the data available from the literature. After that, as summarized in Chapter 2, we studied the impact of early postnatal stress on the transcriptome of epididymal extracellular vesicles. Our findings show that chronic stress in early postnatal life affects miRNAs in adult male reproductive tract extracellular vesicles, with consequences for mature sperm and zygotes. Finally, in Chapter 3 we created an R package for analyzing short RNA sequencing data that is platform-independent, provides a good overview of the quality, performs quality control, aligns the data to a custom-made genome, performs read annotation and assignment with adaptable assignment rules, organizes reads and features as a phylogenetic tree, performs differential analysis, and provides interactive plots for exploratory data analysis. However, there are a number of changes or further investigations that might be needed, which we will explore in this section.

MSUS-associated differences

The MSUS model is appropriate for examining epigenetic inheritance owing to environmental perturbations, but low effect sizes make it difficult to discover candidates that passed multiple testing corrections. The costs associated with sample creation and sequencing, makes it challenging to have a higher number of replicates in the experimental designs. After multiple testing correction, obtaining significant p-values requires either appropriate data distribution

assumptions or a high number of replicates. Clipper, a new statistical framework for reducing false discovery rates without relying on p-values or specific data distributions, was published recently (Ge et al. 2021). Clipper has been shown to outperform conventional methods for limiting false discoveries, thus pipelines can be updated to implement it.

RNA sequencing data analysis pipeline

Data analysis pipelines for RNA sequencing could be adapted. Salmon, a tool for obtaining quantifications from RNA-seq data, is used for pseudo-alignment while accounting for GC biases at the fragment level, sequence-specific biases, and position-specific fragment starts distribution. Salmon Anomaly Detection (SAD) tool can be incorporated into the RNA sequencing data analysis pipeline in addition to Salmon quantifications. SAD identifies probable misquantifications in Salmon’s RNA-seq transcript expression estimation. However, due to several external dependencies, implementing this tool may be difficult (<https://github.com/Kingsford-Group/sad/issues>), and this will almost certainly necessitate containerization from the authors, either via conda or Docker. Also, while evaluating and quantifying lowly-expressed genes and small RNAs with alignment-free pipelines, such as Salmon, potential pitfalls have been recognized, particularly when the small RNAs contain biological variability. As a result, when the analysis isn’t limited to simply protein-coding genes, an alignment-based technique like Rsubread or STAR can be used (Wu et al. 2018). Further, for differential transcript usage and isoform switching, we employed IsoformSwitchAnalyzeR. Under the hood, IsoformSwitchAnalyzeR employs DEXseq for differential transcript usage. However, DEXseq has the drawback of not being able to account for blocking variables. BANDITS could also be used instead of DEXseq in the IsoformSwitchAnalyzeR pipeline because it outperforms DEXseq (Tiberi and Robinson 2020).

ATAC sequencing data analysis pipeline

In recent years, ATAC-seq has grown in popularity as a tool for investigating chromatin accessibility. Despite improvements in protocols, bioinformatics analysis tools are still in their development, with no comprehensive analytical pipeline defined. The interpretation of ATAC-seq results is currently hampered by this. The majority of the tools used to analyze ATAC-seq data are from the ChIP-seq or DNA-seq fields. For example, MACS2, the peak caller we used for ATAC-seq, was designed originally for ChIP-seq data. HMMRATAC is a peak calling method for ATAC-seq data that combines the nucleosome-free and nucleosome-enriched signals from a single ATAC-seq dataset to find open chromatin regions (Tarbell

and Liu 2019). Our strategy was instead to use the nucleosome-free fragments (which are indicative of TF binding) for peak calling. Because we only looked into nucleosome-free fragments, HMMRATAC was not used. Also, HMMRATAC's authors advise against using it on ATAC-seq datasets that have undergone physical or computational size selection. But, HMMRATC could be employed for investigating outside the nucleosome-free fragments. The ATAC-seq pipeline can also be extended to include tools for nucleosome positioning and transcription factor footprinting analysis. However, nucleosome detection is problematic due to an inherent shortcoming of ATAC-seq data, which typically has inadequate read coverage beyond peaks (Yan et al. 2020). Based on empirical and computational estimates, the recommended minimum number of mapped reads for open chromatin detection and differential analysis in mammalian species is 50 million, and 200 million for TF footprinting (Yan et al. 2020). Due to the reduced read coverage outside the open chromatin regions, nucleosome detection in ATAC-seq data was shown to be more difficult than in MNase-seq data (Yan et al. 2020). Because we did not sequence this deep, we did not include TF footprinting in the pipeline, which would necessitate a thorough investigation of tools. These drawbacks necessitate a detailed study of the tools employed in ATAC-seq for normalization, peak calling on NFFs, and TF footprinting.

Updating manuscript from Chapter 1 in light of latest RNA-seq data

The study of spermatogonial cell development in early postnatal and adult was not appropriately designed. As a result, the data analysis and, as a result, the manuscript has to go through many revisions. Initially, it was only sought to investigate the changes at different developmental stages because of the MSUS paradigm. Later, the datasets were repurposed for this study. The study originally employed polyA RNA sequencing data from the PNW21 stage, while RNA sequencing data from the stages PND8 and PND15 came from Total RNA sequencing. Data from the PNW8 stage was taken from the literature due to minimal correlations between polyA and Total RNA sequencing data. It was determined to re-generate the data from the PNW21 stage in the lab in order to have more impactful research and accurate findings. However, because the same library preparation kit and sequencing platform were not available, data from PND8 and PND15 were decided to be re-sequenced. Because of these changes, the entire downstream analysis, including functional analysis, motif analysis, data integration, and re-creation of the figures, will have to be repeated. Some of these analyses have been completed and are included in Chapter 1 as an update.

Low correlations of quantification between shortRNA and qPCR

In terms of relative miRNA expression across samples, we found a low correlation between shortRNA and qPCR. We covered a variety of possible explanations for this mismatch in Chapter 3, including normalizing the data with U6 RNA expression and the lack of UMIs. We don't have an exact explanation for the poor correlations, but we want to look into them further using a variety of techniques, including testing the shortRNA tool with UMI-based data with corresponding qPCR quantifications and creating a simulation dataset.

Appendix A

Datasets analyzed from lab's MSUS model

Tissue	Cells	Model/Time/Age	HTS techniques	Generations	CTRL	MSUS	Data availability	DOI	GitHub
Sperm	Sperm	NA	RNA-Seq (Riboo)	F1	6	7	E-MTAB-5834	NA	mansuylab/sperm_kathi
Sperm	Sperm	NA	short RNA-Seq	F1	4	3	GSE50132	10.1038/mn.3695	dktanwar/sperm_GSE50132
Sperm	Sperm	NA	RNA-Seq (Riboo)	F1	10	10	NA	Not published	mansuylab/sperm_martin mansuylab/omnisperm
Sperm	Sperm	NA	WGBS	F1	10	10	NA	Not published	mansuylab/sperm_WGBS mansuylab/omnisperm
Sperm	Sperm	PPAR injected (24 hours)	RNA-Seq (polyA)	F1	4	4	GSE154369	10.15252/embj.2020104579	mansuylab/tesa_sperm_1 mansuylab/gretchen
Sperm	Sperm	PPAR injected (46 days)	RNA-Seq (Riboo)	F1	7	7	GSE154369	10.15252/embj.2020104579	mansuylab/tesa_sperm_2 mansuylab/gretchen
Testis	Spermatogonial Cells	PND8	RNA-Seq (Riboo)	F1	8	7	NA	Not published	mansuylab/SC_postnatal_adult mansuylab/SC_longRNA
Testis	Spermatogonial Cells	PND8	RNA-Seq (Riboo)	F1	6	NA	NA	Not published	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND15	RNA-Seq (Riboo)	F1	9	8	NA	Not published	mansuylab/SC_postnatal_adult mansuylab/SC_longRNA
Testis	Spermatogonial Cells	PND15	RNA-Seq (Riboo)	F1	6	NA	NA	Not published	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PNW21	RNA-Seq (polyA)	F1	6	6	NA	Not published	mansuylab/SC_postnatal_adult mansuylab/SC_adult
Testis	Spermatogonial Cells	PNW21	RNA-Seq (Riboo)	F1	6	NA	NA	Not published	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND8	short RNA-Seq	F1	8	7	NA	Not published	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND15	short RNA-Seq	F1	6	6	NA	Not published	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND8	RRBS	F1	6	5	NA	Not published	mansuylab/SC_postnatal_adult mansuylab/ssc_pnd8_RRBS
Testis	Spermatogonial Cells	PND15	ATAC-Seq	F1	6	6	NA	Not published	mansuylab/SC_postnatal_adult mansuylab/SC_atac
Testis	Spermatogonial Cells	PNW21	ATAC-Seq	F1	5	5	NA	Not published	mansuylab/SC_postnatal_adult mansuylab/SC_atac
Testis	Sertoli	NA	RNA-Seq (Riboo)	F1	6	7	NA	Not published	NA
Testis	Testis	NA	RNA-Seq (polyA)	F1	5	5	NA	Not published	mansuylab/MSUS23_tissue_RNASeq
Testis	Testis	NA	RNA-Seq (polyA)	F3	5	6	NA	Not published	mansuylab/MSUS23_tissue_RNASeq
Liver	Liver	NA	RNA-Seq (polyA)	F1	5	5	NA	Not published	mansuylab/MSUS23_tissue_RNASeq
Liver	Liver	NA	RNA-Seq (polyA)	F3	5	6	NA	Not published	mansuylab/MSUS23_tissue_RNASeq
Brain	Hippocampus	NA	RNA-Seq (polyA)	F1	5	5	NA	Not published	mansuylab/MSUS23_tissue_RNASeq
Brain	Hippocampus	NA	RNA-Seq (polyA)	F3	5	6	NA	Not published	mansuylab/MSUS23_tissue_RNASeq
Brain	Hypothalamus	NA	RNA-Seq (polyA)	F1	17	7	NA	Not published	NA
Zygote	Zygote	NA	RNA-Seq (Riboo)	F1	4	4	E-MTAB-6589	10.1038/s41380-018-0271-6	mansuylab/zygote_kathi
Embryo	Embryo	4.6.8 cell	RNA-Seq (polyA)	F1	41	36	NA	Not published	mansuylab/embryos_martin
Ovum	Ovum	NA	short RNA-Seq	F1	7	8	NA	Not published	mansuylab/anastasia_ocyto_shortRNA
Cauda	Epididymosmes	NA	short RNA-Seq	F1	5	5	GSE175976	10.1093/biolre/ioab156	mansuylab/alsharabayeva_et_al_2021
HSC	HSC	NA	RNA-Seq (Riboo)	F1	4	5	NA	Not published	mansuylab/HSC_RNA-Seq
Uterosomes	Uterosomes	NA	short RNA-Seq	F1	6	6	NA	Not published	mansuylab/uterosomes

Human dataset from the Schratt lab

Tissue	Cells	HTS techniques	Data availability	Source	GitHub
Blood	PBMC	short RNA-Seq	Not published	Schratt Lab (INS, ETHZ)	ETHZ-INS/schratt_PBMC

Mouse datasets analyzed from literature

Tissue	Cells	Age/Time	HTS techniques	Generations	Group	nsamples	Data availability	DOI	Source	GitHub
Brain	mPFC	NA	RNA-Seq (Ribod0)	F1	Control	3	From author	10.1038/s41386-020-0727-7	Literature	dktanwar/mPFC_MSUS
Brain	mPFC	NA	RNA-Seq (Ribod0)	F1	MSUS	3	From author	10.1038/s41386-020-0727-7	Literature	dktanwar/mPFC_MSUS
Sperm	Sperm	NA	short RNA-Seq	F1	Control	8	GSE162112	10.1016/j.isci.2021.102870	Literature	dktanwar/sperm_GSE162112
Testis	Spermatogonial Cells	PND7	RNA-Seq (Ribod0)	F1	Control	2	GSE62355	10.1101/gad.261925.115	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND14	RNA-Seq (Ribod0)	F1	Control	1	GSM1525703	10.1101/gad.261925.115	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PNW8	RNA-Seq (Ribod0)	F1	Control	1	GSM1415671	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND7	WGBS	F1	Control	1	GSE49623	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND14	WGBS	F1	Control	1	GSE49623	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PNW8	WGBS	F1	Control	7	GSE49623	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PNW8	ChIP-Seq (H3K4me3)	F1	Control	1	GSM1202705	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PNW8	ChIP-Seq (H3K27me3)	F1	Control	1	GSM1202708	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PNW8	ChIP-Seq (H3K27ac)	F1	Control	1	GSM1202713	10.1016/j.stem.2014.04.006	Literature	mansuylab/SC_postnatal_adult
Sperm	Sperm	PNW 10-13	WGBS	F1	Control	6	GSE75323	10.1016/j.devcel.2015.11.024	Literature	dktanwar/GSE75323_reanalysis
Sperm	Sperm	PNW 10-13	WGBS	F1	Low protein	5	GSE75323	10.1016/j.devcel.2015.11.024	Literature	dktanwar/GSE75323_reanalysis
Sperm	Sperm	PNW 10-13	WGBS	F1	High-fat	5	GSE75323	10.1016/j.devcel.2015.11.024	Literature	dktanwar/GSE75323_reanalysis
Testis	Spermatogonial Cells	PND6	ATAC-seq	F1	ID4-eGFP-Bright	3	GSE131657	10.1016/j.isci.2020.101596	Literature	mansuylab/SC_postnatal_adult
Testis	Spermatogonial Cells	PND6	ATAC-seq	F1	ID4-eGFP-Dim	3	GSE131657	10.1016/j.isci.2020.101596	Literature	mansuylab/SC_postnatal_adult

Appendix B

Other publications and manuscripts during doctoral thesis work

Involvement of circulating factors in the transmission of paternal experiences through the germline

Gretchen van Steenwyk[†], Katharina Gapp[†], Ali Jawaid, Pierre-Luc Germain, Francesca Manuella, **Deepak K Tanwar**, Nicola Zamboni, Niharika Gaur, Anastasiia Efimova, Kristina M Thumfart, Eric A Miska, Isabelle Mansuy[#]

[†]Equal contributions

[#]Corresponding author

Journal: *The EMBO Journal*

DOI: 10.15252/embj.2020104579

Contributions: *Analysis of RNA-seq data.*

Symposium summary: Epigenetic inheritance-impact for biology and society 26–28 August 2019, Zurich, Switzerland

Irina Lazar-Contes, Martin Roszkowski, **Deepak K Tanwar** and Isabelle M Mansuy[#]

[#]Corresponding author

Journal: *Environmental Epigenetics*

DOI: 10.15252/embj.2020104579

Contributions: *Writing manuscript.*

OmniSperm: Multiomic analyses of sperm and offspring production from a single male

Martin Roszkowski, Irina Lazar-Contes, Pierre-Luc Germain, **Deepak K Tanwar**, Anara Alshanbayeva, Niharika Obrist, Ali Jawaid, Gretchen van Steenwyk, Eloise Kremer, Dalila Korkmaz, Mark Ormiston, Francesca Manuella, Johannes vom Berg, Jorg Tost, Johannes Bohacek, Isabelle M Mansuy[#]

#Corresponding author

DOI: 10.3929/ethz-b-000489446 (Doctoral thesis, Chapter 2)

Contributions: *Analysis of RNA-seq from sperm and embryos, and WGBS data from sperm.*

Early life stress alters chromatin accessibility landscape and transcript usage in spermatogonial cells during postnatal testis maturation

Irina Lazar-Contes, Gretchen van Steenwyk, **Deepak K Tanwar**, Pierre-Luc Germain, Francesca Manuella, Martin Roszkowski, Niharika Gaur, Isabelle M Mansuy[#]

#Corresponding author

DOI: 10.1101/2020.08.20.259374 (Doctoral thesis, Chapter 3)

Contributions: *Analysis of RNA-seq and ATAC-seq data, and figure generation.*

Appendix C

shortRNA analysis steps in R (a short vignette)

Package installation

```
BiocManager::install("mansuylab/shortRNA")
library(shortRNA)
```

Quality check and trimming

```
qc_SE(file = fq_files, outdir = "output/", ad1 = adapter_sequence)
```

FastQ files to sequence by counts matrix

```
m <- fastq2SeqCountMatrix(files = trimmed_fastq_files)
```

Unique sequences for alignment

```
fa <- DNAStringSet(row.names(m))
names(fa) <- paste0("S", 1:length(fa))
writeXStringSet(fa, fasta_file)
```

Obtaining databases for analysis

```
db <- getDB()
```

Annotation preparation and genome index generation for alignment

```
a <- prepareAnnotation(
  ensdb = db$ensdb,
  output_dir = genomeDir,
  extra.gr = list(piRNA = db$piRNA_GR, miRNA = db$miRNA_GR),
  extra.seqs = list(rRNA = db$rRNA_fa, tRNA = db$tRNA_fa),
  resolveSplicing = NULL,
  rules = defaultAssignRules(),
  tRNAEnsembleRemove = FALSE,
  clusterMiRNA = TRUE
)
```

Alignment

```
alignShortRNA(
  fastq = "unique.fasta",
  index = "genomeDir/customGenome",
  outDir = "align", GTF = exonsBy(db$ensdb),
  GTF.featureType = "exon", GTF.attrType = "gene_id"
)
```

Overlapping aligned data with annotations

```
o <- overlapWithTx2(
  bamFile = align_file, annotation = a,
  ignoreStrand = TRUE, nbthreads = 16
)
```

Assignment of overlapping reads to features with assignment rules

```
ar <- assignReads(sources = o, rules = defaultAssignRules())
```

Converting features to FactorList

```
fl <- featuresAnnoToFL(a)
names(fl) <- lapply(fl, function(x) as.character(x[length(x)]))
```

Assignment of reads to the features tree

```
mappedFeaturesDF <- ar
mappedFeaturesDF$seq <- rownames(mappedFeaturesDF)
ar_tree <- addReadsToTree(
  fL = fl,
  mappedFeaturesDF = mappedFeaturesDF,
  unassigned = FALSE,
  extraTreeBranch = NULL
)
```

Creation of TreeSummarizedExperiment object

```
library(TreeSummarizedExperiment)

rt <- ar_tree
as <- list(counts = m)
cd <- DataFrame(samples = c(s1, s2), group = c("group1", "group2"))

tse <- TreeSummarizedExperiment(
  assays = as,
  rowTree = rt,
  colData = cd,
  rowData = ar[row.names(m), ],
  metadata = list()
```

```
    assignedReads = ar,
    counts = m,
    notAligned = getReadsFromBam(
      bam = align_file
    )
  )
)
```

Differential analysis

```
library(treeClimbR)
dea <- runDA(TSE = tse, filter_min_count = 20)
out <- nodeResult(object = dea, n = Inf)

cand <- getCand(
  tree = dea$tree,
  score_data = out, node_column = "node",
  p_column = "PValue", sign_column = "logFC",
  message = FALSE, threshold = 0.05
)

candB <- evalCand(
  tree = dea$tree,
  levels = cand$candidate_list,
  score_data = out, node_column = "node",
  p_column = "PValue", sign_column = "logFC",
  method = "BH", limit_rej = 0.05,
  use_pseudo_leaf = FALSE,
  message = FALSE
)

result <- topNodes(object = candB, n = Inf, p_value = 0.05)
```

Appendix D

Simulated reads for testing shortRNA

Libraries

```
library(Biostrings)
library(rtracklayer)
```

miRNAs

```
fa <- readRNASTringSet("https://www.mirbase.org/ftp/CURRENT/mature.fa.gz")
fa <- DNAStringSet(fa)
names(fa) <- gsub(" .+", "", names(fa))
fa <- fa[grep(pattern = "mmu", x = names(fa))]
```

cluster chr17:17830188-17830879

```
clusterMirs <- c(
  "mmu-miR-125a-5p", "mmu-miR-99b-5p", "mmu-let-7e-5p",
  "mmu-let-7e-3p", "mmu-miR-99b-3p", "mmu-miR-125a-3p"
)
clusterMirs <- fa[clusterMirs]
```

Imputing differences for differential analysis

```
sizefacts <- c(
  SRR13129036 = 1.2, SRR13129037 = 0.99, SRR13129038 = 1.01,
  SRR13129039 = 1.01, SRR13129040 = 1.07, SRR13129041 = 0.56,
  SRR13129042 = 0.72, SRR13129043 = 0.51
)
```

Read counts

```
set.seed(123)
clusterMirsCounts <- t(sapply(rgamma(length(clusterMirs),
  shape = 3
), function(x) {
  sizefacts *
  exp(1 + x) *
  rnorm(8, mean = rep(c(1, rnorm(1, 2, sd = 0.1)), each = 4), sd = 0.4)
}))
row.names(clusterMirsCounts) <- names(clusterMirs)
clusterMirsCounts <- matrix(rpois(length(clusterMirsCounts),
  lambda = clusterMirsCounts
),
nrow = nrow(clusterMirsCounts), dimnames = dimnames(clusterMirsCounts)
)
clusterMirsCounts
row.names(clusterMirsCounts) <- as.character(clusterMirs)
```

miRNA precursor

```
mir144prec <- DNAStringSet(c(mir144precursor = substr(
  "GGCTGGGATATCATCATATACTGTAAGTTGTGATGAGACACTACAGTATAGATGATGTACTAGTC",
  1, 50
)))
mir144 <- c(
  mir144prec, fa["mmu-miR-144-3p"], fa["mmu-miR-144-5p"],
```

```
DNAStrongSet(c(
  mir144_invalid1 = "AAGTTTGTGATGAGACACTAC",
  mir144_invalid2 = "CTGTAAGTTGTGATGAGAC",
  mir144_invalid3 = "ATGTAAGTCTGGGTA"
))
)

set.seed(123)
mu <- rbind(
  abs(rnorm(8, 1)),
  abs(rnorm(8, rep(2:1, each = 4), 0.3)),
  abs(rnorm(8, rep(2:1, each = 4), 0.3)),
  abs(rnorm(8, 1)), abs(rnorm(8, 1)), abs(rnorm(8, 1))
)
mu <- 300 * t(t(mu) * sizefacts)
mir144Counts <- matrix(rpois(length(mu), lambda = mu),
  nrow = nrow(mu),
  dimnames = list(names(mir144), names(sizefacts)))
)
row.names(mir144Counts) <- as.character(mir144)
```

Data to be simulated

```
e <- rbind(clusterMirsCounts, mir144Counts)
e
```


Appendix E

Curriculum Vitae of the doctoral student.

DEEPAK KUMAR TANWAR

Bioinformatics, multivariate analysis, methods development



EDUCATION

- Mar. 2018 - ● **ETH Zurich**
PhD Candidate
Supervisor: Prof. Dr. Isabelle Mansuy
Thesis committee: Prof. Mark Robinson and Prof. Tuncay Baubec
📍 Zurich, Switzerland
- Aug. 2016 | Jan. 2018 ● **McGill University**
Masters of Science (M.Sc.)
Supervisors: Prof. Sarah Kimmins and Prof. Jianguo Xia
CGPA: 3.57 / 4
📍 Montréal, Canada
- Aug. 2009 | Feb. 2014 ● **Amity University Rajasthan**
Bachelor of Technology (B.Tech.) in Bioinformatics
Supervisor: Prof. Dr. Rainer König
Advisor: Prof. Dr. A. N. Pathak
CGPA: 7.11 / 10 (*First Class*)
📍 Jaipur, India

PhD student specializing in multivariate data analysis, methods development, reproducible analyses, and epigenetics.

RESEARCH EXPERIENCE

- Mar. 2018 | Dec. 2021 ● **Scientific Assistant**
ETH Zurich
• **Supervisor:** Prof. Dr. Isabelle Mansuy
• Epigenetic inheritance research
📍 Zurich, Switzerland
 - Methods and software development
 - Multi-omics data analysis
- Aug. 2016 | Jan. 2018 ● **Graduate Research Assistant**
McGill University
• **Supervisor:** Prof. Sarah Kimmins
• **Co-supervisor:** Prof. Jianguo Xia
📍 Montréal, Canada
 - Epigenetic inheritance research
 - Tools and pipelines development
- Oct. 2014 | Aug. 2016 ● **Visiting Research Scientist**
University of Alabama at Birmingham
• **Supervisor:** Prof. Malay Basu
• Cancer research
📍 Birmingham, USA
 - Multi-omics data analysis
 - Language of protein domain architecture
- Jun. 2013 | Jan. 2014 ● **Research Associate (Intern)**
University of Jena
• **Supervisors:** Prof. Dr. Rainer König
• Sepsis data analysis
📍 Jena, Germany
 - Network modeling
 - Mathematical modeling
- Jun.-Jul. 2012 ● **Internship**
Rajiv Gandhi Centre for Biotechnology (RGCB)
Supervisors: Dr. Sathish Mundayoor (Scientist G) and Mr. Siva Kumar
Project Title: Molecular Docking of NCI Drug Ligands into HIV-1, using Schrödinger.
📍 Thiruvananthapuram, India
- Jun.-Jul. 2011 ● **Industrial internship**
IBI Biosolutions
Supervisor: Dr. Rajnikant Singh
Project Title: Designing of PERL Biological Module.
📍 Chandigarh, India

CONTACT INFO

- ✉ tanward@ethz.ch
📞 +41 44 635 33 49

LINKS

- [ID 0000-0001-8036-1989](#)
[dktanwar](#)
[d_k_tanwar](#)
[in dktanwar](#)
[g Deepak Tanwar](#)
[Deepak_Tanwar2](#)
[Deepak Tanwar](#)

TECHNICAL SKILLS

- Programming:** R, Perl, Python, UNIX, JavaScript, HTML and CSS
Documentation: LaTeX, Markdown
Literate Programming: Sweave, Knitr
Version control: git
Cluster computing

TEACHING

2021	● Bioinformatics and genomic data analysis ETH Zurich Planned the computational part of the block course and delivered lectures in bioinformatics and literate programming	📍 Zurich, Switzerland	Study of Epigenetic Mechanisms in Mental Health (376-1346-00L)
2020	● Introduction to Bioinformatics ETH Zurich Introduction to bioinformatics and data analysis	📍 Zurich, Switzerland	Study of Epigenetic Mechanisms in Mental Health (376-1346-00L)
2019	● Introduction to Biostatistics ETH Zurich Introduction to variables, distributions and tests in statistics, applied towards biology	📍 Zurich, Switzerland	Study of Epigenetic Mechanisms in Mental Health (376-1346-00L)
2017	● BTEC501: Bioinformatics McGill University Taught R and, statistics and data visualization in R; organized weekly tutorials	📍 Montréal, Canada	Semester course
2016	● GBSC 703-01E: Computational Biology and Bioinformatics University of Alabama at Birmingham Taught R and data visualization in R and, assisted participants	📍 Birmingham, USA	Two weeks intensive course
2015	● GBSC 703: Introduction to Scientific Computing University of Alabama at Birmingham Assisted in teaching literate programming and helped participants	📍 Birmingham, USA	Two weeks intensive course

MENTORING

2021	● Julien Chabrey M.Sc. in Biology, ETH Zurich Title: Investigating differential exon usage and differential 3' untranslated regions usage in spermatogonial cells across development	📍 Zurich, Switzerland	Block course Study of Epigenetic Mechanisms in Mental Health (376-1346-00L)
2021	● David Bugliani M.Sc. in Biology, ETH Zurich Title: Investigating differential exon usage and differential 3' untranslated regions usage in spermatogonial cells across development	📍 Zurich, Switzerland	Block course Study of Epigenetic Mechanisms in Mental Health (376-1346-00L)
2019	● Andrew Acciardo M.Sc. in Computational Biology and Bioinformatics, ETH Zurich Title: Computational study of the effects of early life trauma on gene expression and exon usage in various tissues and cells in <i>Mus musculus</i>	📍 Zurich, Switzerland	M.Sc. Thesis
2019	● Hana Parizkova M.Sc. in Computational Biology and Bioinformatics, ETH Zurich Title: Detecting and simulating inheritance of differential methylation	📍 Zurich, Switzerland	Semester project
2019	● Daniela Schildknecht M.Sc. in Computational Biology and Bioinformatics, ETH Zurich Title: An extension of IsoformSwitchAnalyzeR	📍 Zurich, Switzerland	Lab Rotation

SCHOLARSHIP & AWARDS

- Sep. 2018 | Aug. 2021 ● **PhD Scholarship**
Swiss Government Excellence Scholarship: Three years graduate scholarship for a PhD at ETH Zurich
- 2019 ● **Summer School**
Bioinformatics Summer School: Travel award by UCLouvain
- 2017 ● **Travel and Workshop Awards**
Graduate Research Enhancement and Travel Award (GREAT): Travel award by Animal Science Department, McGill University for the 4th Canadian Conference on Epigenetics
Galaxy Community Conference 2017: Travel and registration award by GCC
Epigenomic Data Analysis Workshop: Workshop Registration award by CRRD, McGill University

VOLUNTEERING AND LEADERSHIP

- 2018 | 2021 ● **PhD student representative for SIB**
Swiss Institute of Bioinformatics PhD Training Network co-representative for Zurich area
- 2021 ● **Abstract reviewing for BC2**
Invited to review abstracts for workshop and tutorial session of the Basel Computational Biology Conference [BC]2 2021
- 2019 ● **Symposium organization**
Organizing member of the ISCB Student Council Symposium (SCS) 2019
- 2019 ● **Abstract reviewing**
Reviewed abstracts for GIW/ABACBS/COMBINE conference

PROFESSIONAL SERVICE

- **Reviewer for PLOS Genetics**
- **Reviewed book proposals for CRC press**

PRESENTATIONS AND POSTERS

- Sep. 2019 ● **shortRNA: A flexible framework for the analysis of short RNA sequencing data**
Basel Life Conference

 Basel, Switzerland

Short presentation &
Poster

- Aug. 2019
- **shortRNA: A flexible framework for the analysis of short RNA sequencing data applicable to studies on epigenetic inheritance**
Epigenetics Inheritance Conference, ETH Zurich
📍 Zurich, Switzerland
- Jun. 2019
- **Understanding the molecular mechanisms of germline- dependent epigenetic inheritance: Computational analysis of multi-omics data**
HiFo-INI symposium, ETH Zurich
📍 Zurich, Switzerland
- Jun. 2019
- **Computational analysis of multi-omics data from germ cells across development**
PhD Training Network Retreat, Swiss Institute of Bioinformatics
📍 Zurich, Switzerland
- May 2019
- **Computational analysis of multi-omics data across biological systems**
ZNZ PhD Retreat, ETH Zurich
HiFo PhD Day, ETH Zurich
📍 Zurich, Switzerland
- Nov. 2018
- **Computational analysis of the genetic and epigenetic impact of environmental insults across generations**
D-HEST 5th Research Day, ETH Zurich
📍 Zurich, Switzerland
- May & Jun. 2018
- **The genome and epigenome from a bioinformatician's perspective**
Inaugural Symposium of the Institute for Neuroscience (INS), ETH Zurich
HiFo PhD Day, ETH Zurich
📍 Zurich, Switzerland
- 2017
- **Pipeline for H3K4me3 data analysis from sperm**
Presentation

Science share at McGill University: A bioinformatics pipeline for sperm epigenome analysis (Montréal, Canada)
📍 Canada and France
- YouTube
- Posters**
- The 4th Canadian Conference on Epigenetics* (Whistler, Canada)
Animal Science Research Day (Montréal, Canada)
Galaxy Community Conference (Montpellier, France)
CRRD Research Day (Montréal, Canada)

💻 HACKATHONS PARTICIPATION AND PROJECT

- 2017
- **Hackathons**
DeLEG: Deep Learning for EpiGenomics data to predict phenotype; Montréal, Canada
Hack the Galaxy: ChIP-Seq flavored Galaxy image; Montpellier, France
📍 Canada and France
- GitHub
GitHub
- 2017
- **Project**
CB2 McGill: co-founded a usergroup for *Computational Biology and Bioinformatics* at McGill University
📍 Canada



PUBLICATIONS

● Research papers and theses

Published papers

Anar Alshanbayeva, **Deepak K. Tanwar**, Martin Roszkowski, Francesca Manuella, Isabelle M. Mansuy. Early life stress affects the miRNA cargo of epididymal extracellular vesicles in mouse. *Biology of Reproduction*. doi: [10.1093/biolre/ioab156](https://doi.org/10.1093/biolre/ioab156)

Gretchen van Steenwyk, Katharina Gapp, Ali Jawaid, Pierre-Luc Germain, Francesca Manuella, **Deepak K. Tanwar**, Nicola Zamboni, Niharika Gaur, Anastasiia Efimova, Kristina M. Thumfart, Eric A. Miska, Isabelle M Mansuy. *The EMBO Journal*. doi: [10.1525/embj.2020104579](https://doi.org/10.1525/embj.2020104579)

Ejimedo Madogwe, **Deepak K. Tanwar**, Milena Taibi, Yasmin Schuermann, Audrey St-Yves and Raj Duggavathi. Global analysis of FSH-regulated gene expression and histone modification in mouse granulosa cells. *Molecular Reproduction and Development*. doi: [10.1002/mrd.23419](https://doi.org/10.1002/mrd.23419)

Irina Lazar-Contes, Martin Roszkowski, **Deepak K. Tanwar**, Isabelle M. Mansuy. Symposium summary: Epigenetic inheritance-impact for biology and society 26-28 August 2019, Zurich, Switzerland. *Environmental Epigenetics*. doi: [10.1093/EEP/DVAA004](https://doi.org/10.1093/EEP/DVAA004)

Yu, L., **Tanwar, D.**, Penha, E., Wolf, Y., Koonin, E., & Basu, M. (2019). Grammar of protein domain architectures. *Proceedings Of The National Academy Of Sciences*. doi: [10.1073/pnas.1814684116](https://doi.org/10.1073/pnas.1814684116)

Tanwar, D., Parker, D., Gupta, P., Spurlock, B., Alvarez, R., Basu, M., & Mitra, K. (2016). Crosstalk between the mitochondrial fission protein, Drp1, and the cell cycle is identified across various cancer types and can impact survival of epithelial ovarian cancer patients. *Oncotarget*. doi: [10.18632/oncotarget.11047](https://doi.org/10.18632/oncotarget.11047)

Preprints

Deepak K. Tanwar[†], Irina Lazar-Contes[†], Pierre-Luc Germain, Niharika Gaur, Isabelle M. Mansuy. Transcriptome and epigenome characterization of mouse spermatogonial cells reveals distinct chromatin regulatory landscapes in postnatal and adult testis. doi: [10.1101/2020.08.20.259374](https://doi.org/10.1101/2020.08.20.259374)

In preparation

Kassandra Ma, **Deepak K. Tanwar**, Nicholas Petronella, Swapan Banerjee, Jennifer Ronholm. The Bivalve Microbiome Loses Diversity in the Retail Environment and is Distinct in Bivalves Colonized with *Vibrio vulnificus*

Deepak K. Tanwar, Jianguo Xia, Sarah Kimmins. EpiSpermHis: A Docker container to study H3K4me3 modifications in sperm using Galaxy

Theses

Deepak Tanwar, 2018. M.Sc. Thesis. EpiSpermHis: A Docker Container to Perform the Analysis of Sperm Histone ChIP-Seq Data in Galaxy [McGill University Libraries](#)

Deepak Tanwar, 2014. B.Tech. Thesis. Comprehensive Reanalysis of Genomic Storm (Transcriptomic) Data, Integrating Clinical Variables and Utilizing New and Old Approaches Munich, [GRIN Verlag](#), Available from [ResearchGate](#)

[†]: Equal contributions

References

- Abels, E. R., & Breakefield, X. O. (2016). Introduction to extracellular vesicles: Biogenesis, RNA cargo selection, content, release, and uptake. *Cellular and Molecular Neurobiology*, 36(3), 301–312. <http://doi.org/10.1007/s10571-016-0366-z>
- Alshanbayeva, A., Tanwar, D. K., Roszkowski, M., Manuella, F., & Mansuy, I. M. (2021). Early life stress affects the miRNA cargo of epididymal extracellular vesicles in mouse†. *Biology of Reproduction*, 105(3), 593602. <http://doi.org/10.1093/biolre/ioab156>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [http://doi.org/10.1016/s0022-2836\(05\)80360-2](http://doi.org/10.1016/s0022-2836(05)80360-2)
- Amezquita, R. A. (2018). Marge: An API for analysis of motifs using HOMER in r. <http://doi.org/10.1101/249268>
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., ... Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493), 455–461. <http://doi.org/10.1038/nature12787>
- Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012). FastQC.
- Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1), 111. <http://doi.org/10.1186/s13059-020-02015-1>
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., ... Stegle, O. (2018). Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124. <http://doi.org/10.15252/msb.20178124>
- Barraud, P., & Tisné, C. (2019). To be or not to be modified: Miscellaneous aspects influencing nucleotide modifications in tRNAs. *IUBMB Life*, 71(8), 1126–1140. <http://doi.org/10.1002/iub.2041>
- Barth, T. K., & Imhof, A. (2010). Fast signals and slow marks: The dynamics of histone modifications. *Trends in Biochemical Sciences*, 35(11), 618–626.

- http://doi.org/10.1016/j.tibs.2010.05.006
- Baumgart, M., Barth, E., Savino, A., Groth, M., Koch, P., Petzold, A., ... Cellerino, A. (2017). A miRNA catalogue and ncRNA annotation of the short-living fish Nothobranchius furzeri. *BMC Genomics*, 18(1). http://doi.org/10.1186/s12864-017-3951-8
- Beacon, T. H., Delcuve, G. P., López, C., Nardocci, G., Kovalchuk, I., van Wijnen, A. J., & Davie, J. R. (2021). The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. *Clinical Epigenetics*, 13(1), 138. http://doi.org/10.1186/s13148-021-01126-1
- Benz, F., Roderburg, C., Vargas Cardenas, D., Vucur, M., Gautheron, J., Koch, A., ... Luedde, T. (2013). U6 is unsuitable for normalization of serum miRNA levels in patients with sepsis or liver fibrosis. *Experimental & Molecular Medicine*, 45, e42. http://doi.org/10.1038/emm.2013.81
- Berest, I., Arnold, C., Reyes-Palomares, A., Palla, G., Rasmussen, K. D., Giles, H., ... Zaugg, J. B. (2019). Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF. *Cell Reports*, 29(10), 3147–3159.e12. http://doi.org/10.1016/j.celrep.2019.10.106
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1), 6–21. http://doi.org/10.1101/gad.947102
- Bogenhagen, D. F. (2012). Mitochondrial DNA nucleoid structure. *Biochimica Et Biophysica Acta*, 1819(9-10), 914–920. http://doi.org/10.1016/j.bbagr.2011.11.005
- Bogliotti, Y. S., & Ross, P. J. (2012). Mechanisms of histone H3 lysine 27 trimethylation remodeling during early mammalian development. *Epigenetics*, 7(9), 976–981. http://doi.org/10.4161/epi.21615
- Bohacek, J., & Mansuy, I. M. (2015). Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. *Nature Reviews. Genetics*, 16(11), 641–652. http://doi.org/10.1038/nrg3964
- Bonsergent, E., Grisard, E., Buchrieser, J., Schwartz, O., Théry, C., & Lavieu, G. (2021). Quantitative characterization of extracellular vesicle uptake and content delivery within mammalian cells. *Nature Communications*, 12(1), 1864. http://doi.org/10.1038/s41467-021-22126-y
- Bouckenheimer, J., Fauque, P., Lecellier, C.-H., Bruno, C., Commes, T., Lemaître, J.-M., ... Assou, S. (2018). Differential long non-coding RNA expression profiles in human oocytes and cumulus cells. *Scientific Reports*, 8(1). http://doi.org/10.1038/s41598-018-20727-0
- Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525–527. http://doi.org/10.1038/nbt.3519

- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. *Cell*, *128*(6), 1089–1103. <http://doi.org/10.1016/j.cell.2007.01.043>
- Brown, J. (2018). Git-fat. Retrieved from <https://github.com/jedbrown/git-fat>
- Bruning, O., Rodenburg, W., Wackers, P. F. K., van Oostrom, C., Jonker, M. J., Dekker, R. J., ... Breit, T. M. (2016). Confounding factors in the transcriptome analysis of an in-vivo exposure experiment. *Plos One*, *11*(1), e0145252. <http://doi.org/10.1371/journal.pone.0145252>
- Brunner, A. M., Nanni, P., & Mansuy, I. M. (2014). Epigenetic marking of sperm by post-translational modification of histones and protamines. *Epigenetics & Chromatin*, *7*(1), 2. <http://doi.org/10.1186/1756-8935-7-2>
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. <http://doi.org/10.1038/nmeth.2688>
- Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, *109*, 21.29.1–21.29.9. <http://doi.org/10.1002/0471142727.mb2129s109>
- Buffalo, V. (2020). Qrqc: Quick read quality control. Retrieved from <http://github.com/vsbuffalo/qrqc>
- Canovas, S., & Ross, P. J. (2016). Epigenetics in preimplantation mammalian development. *Theriogenology*, *86*(1), 69–79. <http://doi.org/10.1016/j.theriogenology.2016.04.020>
- Casas, E., & Vavouri, T. (2014). Sperm epigenomics: Challenges and opportunities. *Frontiers in Genetics*, *5*, 330. <http://doi.org/10.3389/fgene.2014.00330>
- Casas, E., & Vavouri, T. (2020). Mechanisms of epigenetic inheritance of variable traits through the germline. *Reproduction*, *159*(6), R251–R263. <http://doi.org/10.1530/REP-19-0340>
- Chacon, S., & Straub, B. (2021). Pro git. Retrieved from <https://git-scm.com/book/en/v2>
- Chan, F., Oatley, M. J., Kaucher, A. V., Yang, Q. E., Bieberich, C. J., Shashikant, C. S., & Oatley, J. M. (2014). Functional and molecular features of the Id4+ germline stem cell population in mouse testes. *Genes and Development*, *28*(12), 1351–1362. <http://doi.org/10.1101/gad.240465.114>
- Chan, P. P., & Lowe, T. M. (2016). GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research*, *44*(D1), D184–9. <http://doi.org/10.1093/nar/gkv1309>

- Chan, P., Lin, B., Mak, A., & Lowe, T. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096. <http://doi.org/10.1093/nar/gkab688>
- Chen, C.-J., Servant, N., Toedling, J., Sarazin, A., Marchais, A., Duvernois-Berthet, E., ... Barillot, E. (2012). ncPRO-seq: A tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics*, *28*(23), 3147–3149. <http://doi.org/10.1093/bioinformatics/bts587>
- Chen, Z., & Riggs, A. D. (2011). DNA methylation and demethylation in mammals. *The Journal of Biological Chemistry*, *286*(21), 18347–18353. <http://doi.org/10.1074/jbc.R110.205286>
- Consortium, G. O. (2021). The gene ontology resource: Enriching a GOld mine. *Nucleic Acids Research*, *49*(D1), D325–D334. <http://doi.org/10.1093/nar/gkaa1113>
- Cooley, L., Appel, B., & Soll, D. (1982). Post-transcriptional nucleotide addition is responsible for the formation of the 5' terminus of histidine tRNA. *Proceedings of the National Academy of Sciences*, *79*(21), 6475–6479. <http://doi.org/10.1073/pnas.79.21.6475>
- Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., ... Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, *14*(10), 959–962. <http://doi.org/10.1038/nmeth.4396>
- Cozen, A. E., Quartley, E., Holmes, A. D., Hrabetá-Robinson, E., Phizicky, E. M., & Lowe, T. M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nature Methods*, *12*(9), 879–884. <http://doi.org/10.1038/nmeth.3508>
- Darwin, C. (1871). Pangenesis. *Nature*, *3*(78), 502–503. <http://doi.org/10.1038/003502a0>
- Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, *25*(10), 1010–1022. <http://doi.org/10.1101/gad.2037511>
- Desvignes, T., Batzel, P., Sydes, J., Eames, B. F., & Postlethwait, J. H. (2019). miRNA analysis with prost! Reveals evolutionary conservation of organ-enriched expression and post-transcriptional modifications in three-spined stickleback and zebrafish. *Scientific Reports*, *9*(1), 3913. <http://doi.org/10.1038/s41598-019-40361-8>
- Donati, S., Ciuffi, S., & Brandi, M. L. (2019). Human circulating miRNAs real-time qRT-PCR-based analysis: An overview of endogenous reference genes used for data normalization. *International Journal of Molecular Sciences*, *20*(18). <http://doi.org/10.3390/ijms20184353>
- Erkek, S., Hisano, M., Liang, C.-Y., Gill, M., Murr, R., Dieker, J., ... Peters, A. H. F. M. (2013). Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nature Structural & Molecular Biology*, *20*(7), 868–875. <http://doi.org/10.1038/nsmb.2599>

- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <http://doi.org/10.1093/bioinformatics/btw354>
- Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3), 186–194. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9521922>
- Franklin, T. B., Russig, H., Weiss, I. C., Gräff, J., Linder, N., Michalon, A., ... Mansuy, I. M. (2010). Epigenetic transmission of the impact of early stress across generations. *Biological Psychiatry*, 68(5), 408415. <http://doi.org/10.1016/j.biopsych.2010.05.036>
- Freist, W., Verhey, J. F., Rühlmann, A., Gauss, D. H., & Arnez, J. G. (1999). Histidyl-tRNA synthetase. *Biological Chemistry*, 380(6), 623–646. <http://doi.org/10.1515/BC.1999.079>
- Fromant, M., Plateau, P., & Blanquet, S. (2000). Function of the Extra 5'-Phosphate Carried by Histidine tRNA. *Biochemistry*, 39(14), 4062–4067. <http://doi.org/10.1021/bi9923297>
- Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., ... Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5), 1827–1831. <http://doi.org/10.1073/pnas.89.5.1827>
- Fu, L., & Zhang, L. (2019). Serotonylation: A novel histone H3 marker. *Signal Transduction and Targeted Therapy*, 4, 15. <http://doi.org/10.1038/s41392-019-0048-7>
- Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics*, 19(1), 531. <http://doi.org/10.1186/s12864-018-4933-1>
- Gapp, K., & Bohacek, J. (2018). Epigenetic germline inheritance in mammals: Looking to the past to understand the future. *Genes, Brain, and Behavior*, 17(3), e12407. <http://doi.org/10.1111/gbb.12407>
- Gapp, Katharina, Parada, G. E., Gross, F., Corcoba, A., Kaur, J., Grau, E., ... Miska, E. A. (2021). Single paternal dexamethasone challenge programs offspring metabolism and reveals multiple candidates in RNA-mediated inheritance. *iScience*, 24(8), 102870. <http://doi.org/10.1016/j.isci.2021.102870>
- Gapp, K., van Steenwyk, G., Germain, P. L., Matsushima, W., Rudolph, K. L. M., Manuella, F., ... Miska, E. A. (2020). Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma. *Molecular Psychiatry*, 25(9), 21622174. <http://doi.org/10.1038/s41380-018-0271-6>
- Gerber, S., Schratt, G., & Germain, P.-L. (2021). Streamlining differential exon and 3' UTR usage with diffUTR. *BioRxiv*. <http://doi.org/10.1101/2021.02.12.430963>

- Germain, P.-L., Sonrel, A., & Robinson, M. D. (2020). pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. *Genome Biology*, 21(1), 227. <http://doi.org/10.1186/s13059-020-02136-7>
- GM, F. (2018). tRNADBImport. <http://doi.org/10.18129/B9.BIOC.TRNADBIMPORT>
- Gong, H., Liu, C.-M., Liu, D.-P., & Liang, C.-C. (2005). The role of small RNAs in human diseases: Potential troublemaker and therapeutic tools. *Medicinal Research Reviews*, 25(3), 361–381. <http://doi.org/10.1002/med.20023>
- Gontarz, P., Fu, S., Xing, X., Liu, S., Miao, B., Bazylianska, V., ... Zhang, B. (2020). Comparison of differential accessibility analysis strategies for ATAC-seq data. *Scientific Reports*, 10(1), 10150. <http://doi.org/10.1038/s41598-020-66998-4>
- Goodnight, A. V., Kremsky, I., Khampang, S., Jung, Y. H., Billingsley, J. M., Bosinger, S. E., ... Chan, A. W. S. (2019). Chromatin accessibility and transcription dynamics during in vitro astrocyte differentiation of huntington's disease monkey pluripotent stem cells. *Epigenetics & Chromatin*, 12(1), 67. <http://doi.org/10.1186/s13072-019-0313-6>
- Green, R., & Noller, H. F. (1997). Ribosomes and translation. *Annual Review of Biochemistry*, 66, 679–716. <http://doi.org/10.1146/annurev.biochem.66.1.679>
- Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: Tools for microRNA genomics. *Nucleic Acids Research*, 36(Database issue), D154–8. <http://doi.org/10.1093/nar/gkm952>
- Hammoud, S. S., Low, D. H. P., Yi, C., Carrell, D. T., Guccione, E., & Cairns, B. R. (2014). Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell*, 15(2), 239–253. <http://doi.org/10.1016/j.stem.2014.04.006>
- Hammoud, S. S., Low, D. H. P., Yi, C., Lee, C. L., Oatley, J. M., Payne, C. J., ... Cairns, B. R. (2015). Transcription and imprinting dynamics in developing postnatal male germline stem cells. <http://doi.org/10.1101/gad.261925>
- Handzlik, J. E., Tastsoglou, S., Vlachos, I. S., & Hatzigeorgiou, A. G. (2020). Manatee: Detection and quantification of small non-coding RNAs from next-generation sequencing data. *Scientific Reports*, 10(1), 705. <http://doi.org/10.1038/s41598-020-57495-9>
- Hanson, M. A., & Skinner, M. K. (2016). Developmental origins of epigenetic transgenerational inheritance. *Environmental Epigenetics*, 2(1). <http://doi.org/10.1093/eep/dvw002>
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., ... Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9), 1760–1774. <http://doi.org/10.1101/gr.135350.111>

- Heard, E., & Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: Myths and mechanisms. *Cell*, 157(1), 95–109. <http://doi.org/10.1016/j.cell.2014.02.045>
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., ... Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4), 576–589. <http://doi.org/10.1016/j.molcel.2010.05.004>
- Hou, Y.-M. (2010). CCA addition to tRNA: Implications for tRNA quality control. *IUBMB Life*, 62(4), 251–260. <http://doi.org/10.1002/iub.301>
- Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., ... Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, 49(D1), D884–D891. <http://doi.org/10.1093/nar/gkaa942>
- Huang, C., Xu, M., & Zhu, B. (2013). Epigenetic inheritance mediated by histone lysine methylation: Maintaining transcriptional states without the precise restoration of marks? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1609), 20110332. <http://doi.org/10.1098/rstb.2011.0332>
- Huang, R., Soneson, C., Ernst, F. G. M., Rue-Albrecht, K. C., Yu, G., Hicks, S. C., & Robinson, M. D. (2020). TreeSummarizedExperiment: A S4 class for data with hierarchical structure. *F1000Research*, 9, 1246. <http://doi.org/10.12688/f1000research.26669.1>
- Huang, R., Soneson, C., Germain, P.-L., Schmidt, T. S. B., Mering, C. V., & Robinson, M. D. (2021). treeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses. *Genome Biology*, 22(1). <http://doi.org/10.1186/s13059-021-02368-1>
- Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8, 84. <http://doi.org/10.3389/fgene.2017.00084>
- Ibba, M., & Soll, D. (2000). Aminoacyl-tRNA synthesis. *Annual Review of Biochemistry*, 69, 617–650. <http://doi.org/10.1146/annurev.biochem.69.1.617>
- Iwasaki, W., Miya, Y., Horikoshi, N., Osakabe, A., Taguchi, H., Tachiwana, H., ... Kurumizaka, H. (2013). Contribution of histone n-terminal tails to the structure and stability of nucleosomes. *FEBS Open Bio*, 3, 363–369. <http://doi.org/10.1016/j.fob.2013.08.007>
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., ... D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1), D498–D503. <http://doi.org/10.1093/nar/gkz1031>
- Jawaid, A., Jehle, K.-L., & Mansuy, I. M. (2021). Impact of parental exposure on offspring health in humans. *Trends in Genetics*, 37(4), 373–388. <http://doi.org/10.1016/j.tig.2020.10.006>
- Jawaid, A., & Mansuy, I. M. (2019). Inter- and transgenerational inheritance of behavioral phenotypes. *Current Opinion in Behavioral Sciences*, 25, 96–101.

- http://doi.org/10.1016/j.cobeha.2018.12.004
- Johnson, N. R., Yeoh, J. M., Coruh, C., & Axtell, M. J. (2016). Improved placement of multi-mapping small RNAs. *G3 (Bethesda, Md.)*, 6(7), 2103–2111. http://doi.org/10.1534/g3.116.030452
- Jones, P. A., & Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science*, 293(5532), 1068–1070. http://doi.org/10.1126/science.1063852
- Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., & Pütz, J. (2009). tRNADB 2009: Compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(Database issue), D159–62. http://doi.org/10.1093/nar/gkn772
- Jung, Y. H., Bixler, B. J., Ruiz, D., Wang, H.-L. V., Linsenbaum, H., Xiang, J.-F., ... Corces, V. G. (2020). Transgenerational inheritance of BPA-induced obesity correlates with transmission of new CTCF sites in the fto gene. *BioRxiv*. http://doi.org/10.1101/2020.11.20.391672
- Jung, Y. H., Kremsky, I., Gold, H. B., Rowley, M. J., Punyawai, K., Buonanotte, A., ... Corces, V. G. (2019). Maintenance of CTCF- and transcription factor-mediated interactions from the gametes to the early mouse embryo. *Molecular Cell*, 75(1), 154–171.e5. http://doi.org/10.1016/j.molcel.2019.04.014
- Jung, Y. H., Sauria, M. E. G., Lyu, X., Cheema, M. S., Ausio, J., Taylor, J., & Corces, V. G. (2017). Chromatin states in mouse sperm correlate with embryonic and adult regulatory landscapes. *Cell Reports*, 18(6), 1366–1382. http://doi.org/10.1016/j.celrep.2017.01.034
- Kaisers, W. (2020). seqTools: Analysis of nucleotide, sequence and quality content on fastqfiles. Retrieved from <https://bioconductor.org/packages/release/bioc/html/seqTools.html>
- Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1), D545–D551. http://doi.org/10.1093/nar/gkaa970
- Kia, A., Gloeckner, C., Osothprarop, T., Gormley, N., Bomati, E., Stephenson, M., ... He, M. M. (2017). Improved genome sequencing using an engineered transposase. *BMC Biotechnology*, 17(1), 6. http://doi.org/10.1186/s12896-016-0326-1
- Kolasa, A., Misiakiewicz, K., Marchlewicz, M., & Wiszniewska, B. (2012). The generation of spermatogonial stem cells and spermatogonia in mammals. *Reproductive Biology*, 12(1), 5–23. http://doi.org/10.1016/S1642-431X(12)60074-6
- König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., ... Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7), 909–915. http://doi.org/10.1038/nsmb.1838

- Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2016). Fast gene set enrichment analysis. *BioRxiv*. <http://doi.org/10.1101/060012>
- Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, *128*(4), 693–705. <http://doi.org/10.1016/j.cell.2007.02.005>
- Kremsky, I., & Corces, V. G. (2020). Protection from DNA re-methylation by transcription factors in primordial germ cells and pre-implantation embryos can explain trans-generational epigenetic inheritance. *Genome Biology*, *21*(1), 118. <http://doi.org/10.1186/s13059-020-02036-w>
- Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, *27*(11), 1571–1572. <http://doi.org/10.1093/bioinformatics/btr167>
- Kuksa, P. P., Amlie-Wolf, A., Katanic, Ž., Valladares, O., Wang, L.-S., & Leung, Y. Y. (2018). SPAR: Small RNA-seq portal for analysis of sequencing experiments. *Nucleic Acids Research*, *46*(W1), W36–W42. <http://doi.org/10.1093/nar/gky330>
- Lai, X., & Vera, J. (2013). MicroRNA clusters (pp. 1310–1314). Springer New York. http://doi.org/10.1007/978-1-4419-9863-7_1121
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*(4), 357–359. <http://doi.org/10.1038/nmeth.1923>
- Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. <http://doi.org/10.1186/gb-2014-15-2-r29>
- Lawrence, M., Daujat, S., & Schneider, R. (2016). Lateral thinking: How histone modifications regulate gene expression. *Trends in Genetics*, *32*(1), 42–56. <http://doi.org/10.1016/j.tig.2015.10.007>
- Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, *12*(1). <http://doi.org/10.1186/1471-2105-12-253>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Subgroup, 1000. G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. <http://doi.org/10.1093/bioinformatics/btp352>
- Li, J., Han, X., Wan, Y., Zhang, S., Zhao, Y., Fan, R., ... Zhou, Y. (2018). TAM 2.0: Tool for MicroRNA set analysis. *Nucleic Acids Research*, *46*(W1), W180–W185. <http://doi.org/10.1093/nar/gky509>
- Li, X. Z., Roy, C. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W., ... Zamore, P. D. (2013). An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Molecular Cell*, *50*(1), 67–81.

- http://doi.org/10.1016/j.molcel.2013.02.016
- Li, Z., Ender, C., Meister, G., Moore, P. S., Chang, Y., & John, B. (2012). Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Research*, 40(14), 6787–6799. http://doi.org/10.1093/nar/gks307
- Liao, Y., Smyth, G. K., & Shi, W. (2019). The r package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8), e47e47. http://doi.org/10.1093/nar/gkz114
- Lo Cicero, A., Stahl, P. D., & Raposo, G. (2015). Extracellular vesicles shuffling inter-cellular messages: For good or for bad. *Current Opinion in Cell Biology*, 35, 69–77. http://doi.org/10.1016/j.ceb.2015.04.013
- Loher, P., Telonis, A. G., & Rigoutsos, I. (2017). MINTmap: Fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Scientific Reports*, 7, 41184. http://doi.org/10.1038/srep41184
- Lorena Pantano [Aut, Cre], Georgia Escaramis [Aut], ChristosArgyropoulos [Aut]. (2017). isomiRs. http://doi.org/10.18129/B9.BIOC.ISOMIRS
- Lun, A. T. L., & Smyth, G. K. (2015). CsaW: A bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*, 44(5), e45. http://doi.org/10.1093/nar/gkv1191
- Luo, C., Hajkova, P., & Ecker, J. R. (2018). Dynamic DNA methylation: In the right place at the right time. *Science*, 361(6409), 1336–1340. http://doi.org/10.1126/science.aat6806
- Magee, R., Loher, P., Londin, E., & Rigoutsos, I. (2017). Threshold-seq: A tool for determining the threshold in short RNA-seq datasets. *Bioinformatics*, 33(13), 2034–2036. http://doi.org/10.1093/bioinformatics/btx073
- Mao, S., Li, Y., Liu, B., & Chi, T. (2017). Mouse models of epigenetic inheritance: Classification, mechanisms, and experimental strategies (pp. 231–243). Elsevier. http://doi.org/10.1016/B978-0-12-805388-1.00015-8
- Marmorstein, R., & Trievel, R. C. (2009). Histone modifying enzymes: Structures, mechanisms, and specificities. *Biochimica Et Biophysica Acta*, 1789(1), 58–68. http://doi.org/10.1016/j.bbapm.2008.07.009
- Mayr, C. (2019). What are 3' utrs doing? *Cold Spring Harbor Perspectives in Biology*, 11(10). http://doi.org/10.1101/cshperspect.a034728
- McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6), 765–771. http://doi.org/10.1093/bioinformatics/btp053
- McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., ... Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5), 495–501. http://doi.org/10.1038/nbt.1630

- Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18), 5868–5877. <http://doi.org/10.1093/nar/gki901>
- Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H., & Gentleman, R. (2009). ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25(19), 2607–2608. <http://doi.org/10.1093/bioinformatics/btp450>
- Morgan, M., Obenchain, V., Hester, J., & Pagès, H. (2020). SummarizedExperiment: SummarizedExperiment container. Retrieved from <https://bioconductor.org/packages/SummarizedExperiment>
- Mutoji, K., Singh, A., Nguyen, T., Gildersleeve, H., Kaucher, A. V., Oatley, M. J., ... Hermann, B. P. (2016). TSPAN8 expression distinguishes spermatogonial stem cells in the prepubertal mouse testis. *Biology of Reproduction*, 95(6), 117. <http://doi.org/10.1095/biolreprod.116.144220>
- Nagy, C., & Turecki, G. (2015). Transgenerational epigenetic inheritance: An open discussion. *Epigenomics*, 7(5), 781–790. <http://doi.org/10.2217/epi.15.46>
- O'Brien, K., Breyne, K., Ughetto, S., Laurent, L. C., & Breakefield, X. O. (2020). RNA delivery by extracellular vesicles in mammalian cells and its applications. *Nature Reviews. Molecular Cell Biology*, 21(10), 585–606. <http://doi.org/10.1038/s41580-020-0251-y>
- Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2020). Biostrings: Efficient manipulation of biological strings. Retrieved from <https://bioconductor.org/packages/Biostrings>
- Park, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews. Genetics*, 10(10), 669–680. <http://doi.org/10.1038/nrg2641>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 417–419. <http://doi.org/10.1038/nmeth.4197>
- Pisco, A. O., Fouquier d'Herouel, A., & Huang, S. (2016). Conceptual confusion: The case of epigenetics. *BioRxiv*. <http://doi.org/10.1101/053009>
- Pratt, A. J., & MacRae, I. J. (2009). The RNA-induced silencing complex: A versatile gene-silencing machine. *The Journal of Biological Chemistry*, 284(27), 17897–17901. <http://doi.org/10.1074/jbc.R900012200>
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue), D590–6. <http://doi.org/10.1093/nar/gks1219>

- Rahman, R.-U., Gautam, A., Bethune, J., Sattar, A., Fiosins, M., Magruder, D. S., ... Bonn, S. (2018). Oasis 2: Improved online analysis of small RNA-seq data. *BMC Bioinformatics*, 19(1), 54. <http://doi.org/10.1186/s12859-018-2047-z>
- Ramirez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., ... Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1), W160W165. <http://doi.org/10.1093/nar/gkw257>
- Rando, O. J. (2016). Intergenerational transfer of epigenetic information in sperm. *Cold Spring Harbor Perspectives in Medicine*, 6(5). <http://doi.org/10.1101/cshtperspect.a022988>
- Reik, W., Dean, W., & Walter, J. (2001). Epigenetic Reprogramming in Mammalian Development. *Science*, 293(5532), 1089–1093. <http://doi.org/10.1126/science.1063443>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47. <http://doi.org/10.1093/nar/gkv007>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139140. <http://doi.org/10.1093/bioinformatics/btp616>
- Robles, V., Valcarce, D. G., & Riesco, M. F. (2019). Non-coding RNA regulation in reproduction: Their potential use as biomarkers. *Non-Coding RNA Research*, 4(2), 54–62. <http://doi.org/10.1016/j.ncrna.2019.04.001>
- Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11), e1005752. <http://doi.org/10.1371/journal.pcbi.1005752>
- Royo, H., Stadler, M. B., & Peters, A. H. F. M. (2016). Alternative computational analysis shows no evidence for nucleosome enrichment at repetitive sequences in mammalian spermatozoa. *Developmental Cell*, 37(1), 98–104. <http://doi.org/10.1016/j.devcel.2016.03.010>
- Rozowsky, J., Kitchen, R. R., Park, J. J., Galeev, T. R., Diao, J., Warrell, J., ... Gerstein, M. (2019). exceRpt: A comprehensive analytic platform for extracellular RNA profiling. *Cell Systems*, 8(4), 352357.e3. <http://doi.org/10.1016/j.cels.2019.03.004>
- Rueda, A., Barturen, G., Lebrón, R., Gómez-Martín, C., Alganza, Á., Oliver, J. L., & Hackenberg, M. (2015). sRNAtoolbox: An integrated collection of small RNA research tools. *Nucleic Acids Research*, 43(W1), W467–73. <http://doi.org/10.1093/nar/gkv555>
- Saxton, D. S., & Rine, J. (2019). Epigenetic memory independent of symmetric histone inheritance. *eLife*, 8. <http://doi.org/10.7554/eLife.51421>
- Sayols, S., Scherzinger, D., & Klein, H. (2016). dupRadar: A bioconductor package for the assessment of PCR artifacts in RNA-seq data. *BMC Bioinformatics*, 17(1), 428. <http://doi.org/10.1186/s12859-016-1276-2>

- Schaefer, S., & Nadeau, J. H. (2015). The genetics of epigenetic inheritance: Modes, molecules, and mechanisms. *The Quarterly Review of Biology*, 90(4), 381–415. <http://doi.org/10.1086/683699>
- Shi, J., Ko, E.-A., Sanders, K. M., Chen, Q., & Zhou, T. (2018). SPORTS1.0: A tool for annotating and profiling non-coding RNAs optimized for rRNA- and tRNA-derived small RNAs. *Genomics, Proteomics & Bioinformatics / Beijing Genomics Institute*, 16(2), 144–151. <http://doi.org/10.1016/j.gpb.2018.04.004>
- Siklenka, K., Erkek, S., Godmann, M., Lambrot, R., McGraw, S., Lafleur, C., ... Kimmins, S. (2015). Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science*, 350(6261), aab2006. <http://doi.org/10.1126/science.aab2006>
- Sims, J. K., Houston, S. I., Magazinnik, T., & Rice, J. C. (2006). A trans-tail histone code defined by monomethylated H4 lys-20 and H3 lys-9 demarcates distinct regions of silent chromatin. *The Journal of Biological Chemistry*, 281(18), 12760–12766. <http://doi.org/10.1074/jbc.M513462200>
- Skog, S., Örkenby, L., Kugelberg, U., Tariq, K., Farrants, A.-K. Ö., Öst, A., & Nätt, D. (2021). Seqpac: A new framework for small RNA analysis in r using sequence-based counts. *BioRxiv*. <http://doi.org/10.1101/2021.03.19.436151>
- Smallwood, S. A., & Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. *Trends in Genetics*, 28(1), 33–42. <http://doi.org/10.1016/j.tig.2011.09.004>
- Smith, Z. D., & Meissner, A. (2013). DNA methylation: Roles in mammalian development. *Nature Reviews. Genetics*, 14(3), 204–220. <http://doi.org/10.1038/nrg3354>
- Sprinzl, M., & Cramer, F. (1979). The -c-c-a end of tRNA and its role in protein biosynthesis (pp. 1–69). Elsevier. [http://doi.org/10.1016/s0079-6603\(08\)60798-9](http://doi.org/10.1016/s0079-6603(08)60798-9)
- Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., ... Moulton, V. (2012). The UEA sRNA workbench: A suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, 28(15), 2059–2061. <http://doi.org/10.1093/bioinformatics/bts311>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. <http://doi.org/10.1073/pnas.0506580102>
- Sullivan, R. (2016). Epididymosomes Role of extracellular microvesicles in sperm maturation. *Frontiers in Bioscience*, 8(1), 106–114. <http://doi.org/10.2741/s450>
- Surani, M. A., Hayashi, K., & Hajkova, P. (2007). Genetic and epigenetic regulators of pluripotency. *Cell*, 128(4), 747–762. <http://doi.org/10.1016/j.cell.2007.02.010>

- Susiarjo, M. (2016). Introduction to epigenetic mechanisms (pp. 49–62). Elsevier.
<http://doi.org/10.1016/B978-0-12-801383-0.00004-9>
- Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Computational Biology*, 12(4), e1004873. <http://doi.org/10.1371/journal.pcbi.1004873>
- Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, 31(12), 2032–2034. <http://doi.org/10.1093/bioinformatics/btv098>
- Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., & Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3), 569–583. <http://doi.org/10.1093/biostatistics/kxu001>
- Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., ... Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414), 75–82. <http://doi.org/10.1038/nature11232>
- Tsagiopoulou, M., Maniou, M. C., Pechlivanis, N., Togkousidis, A., Kotrová, M., Hutzenlaub, T., ... Psomopoulos, F. (2021). UMIc: A preprocessing method for UMI deduplication and reads correction. *Frontiers in Genetics*, 12. <http://doi.org/10.3389/fgene.2021.660366>
- Vakoc, C. R., Sachdeva, M. M., Wang, H., & Blobel, G. A. (2006). Profile of histone lysine methylation across transcribed mammalian chromatin. *Molecular and Cellular Biology*, 26(24), 9185–9195. <http://doi.org/10.1128/MCB.01529-06>
- van Steenwyk, G., Gapp, K., Jawaid, A., Germain, P.-L., Manuella, F., Tanwar, D. K., ... Mansuy, I. M. (2020). Involvement of circulating factors in the transmission of paternal experiences through the germline. *The EMBO Journal*, 39(23), e104579. <http://doi.org/10.15252/embj.2020104579>
- van Steenwyk, G., Roszkowski, M., Manuella, F., Franklin, T. B., & Mansuy, I. M. (2018). Transgenerational inheritance of behavioral and metabolic effects of paternal exposure to traumatic stress in early postnatal life: Evidence in the 4th generation. *Environmental Epigenetics*, 4(2), dvy023. <http://doi.org/10.1093/eep/dvy023>
- Veselovska, L., Smallwood, S. A., Saadeh, H., Stewart, K. R., Krueger, F., Maupetit-Méhouas, S., ... Kelsey, G. (2015). Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome Biology*, 16, 209. <http://doi.org/10.1186/s13059-015-0769-z>
- Vitsios, D. M., & Enright, A. J. (2015). Chimira: Analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*, 31(20), 3365–3367. <http://doi.org/10.1093/bioinformatics/btv380>

- Vitting-Seerup, K., & Sandelin, A. (2017). The landscape of isoform switches in human cancers. *Molecular Cancer Research*, 15(9), 1206–1220. <http://doi.org/10.1158/1541-7786.MCR-16-0459>
- Vitting-Seerup, K., & Sandelin, A. (2019). IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, 35(21), 4469–4471. <http://doi.org/10.1093/bioinformatics/btz247>
- Waddington, C. H. (1942). The epigenotype. *Endeavour*, 1(18).
- Wang, W., & Carroll, T. (2020). Rfastp: An ultra-fast and all-in-one fastq preprocessor (QualityControl, adapter, low quality and polyX trimming) and UMISequencing parsing). Retrieved from <http://www.bioconductor.org/packages/release/bioc/html/Rfastp.html>
- Weismann, A. (1893). *The germ-plasm: A theory of heredity. Translated by w. Newton parker and harriet rönnfeldt.* New York: Scribner. <http://doi.org/10.5962/bhl.title.25196>
- Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M., & Wilke, C. O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19(1), 510. <http://doi.org/10.1186/s12864-018-4869-5>
- Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., & Smyth, G. K. (2010). ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17), 2176–2182. <http://doi.org/10.1093/bioinformatics/btq401>
- Wu, D., & Smyth, G. K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17), e133. <http://doi.org/10.1093/nar/gks461>
- Wu, J., Liu, Q., Wang, X., Zheng, J., Wang, T., You, M., ... Shi, Q. (2013). mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biology*, 10(7), 1087–1092. <http://doi.org/10.4161/rna.25193>
- Wu, X., Kim, T.-K., Baxter, D., Scherler, K., Gordon, A., Fong, O., ... Wang, K. (2017). sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Research*, 45(21), 12140–12151. <http://doi.org/10.1093/nar/gkx999>
- Yi, L., Liu, L., Melsted, P., & Pachter, L. (2018). A direct comparison of genome alignment and transcriptome pseudoalignment. *BioRxiv*. <http://doi.org/10.1101/444620>
- Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biology*, 11(2), R14. <http://doi.org/10.1186/gb-2010-11-2-r14>
- Yuan, X., Liu, C., Yang, P., He, S., Liao, Q., Kang, S., & Zhao, Y. (2009). Clustered microRNAs' coordination in regulating protein-protein interaction network. *BMC Systems Biology*, 3, 65. <http://doi.org/10.1186/1752-0509-3-65>

- Zhang, T., Zhang, Z., Dong, Q., Xiong, J., & Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology*, 21(1), 45. <http://doi.org/10.1186/s13059-020-01957-w>
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... Liu, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biology*, 9(9), R137. <http://doi.org/10.1186/gb-2008-9-9-r137>
- Zhu, H., Wang, G., & Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics*, 17(9), 551–565. <http://doi.org/10.1038/nrg.2016.83>
- Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Bin Li, ... Shi, L. (2016). Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using RNA-seq. *Scientific Reports*, 6, 28400. <http://doi.org/10.1038/srep28400>
- Zuo, L., Wang, Z., Tan, Y., Chen, X., & Luo, X. (2016). piRNAs and Their Functions in the Brain. *International Journal of Human Genetics*, 16(1-2), 53–60. <http://doi.org/10.1080/09723757.2016.11886278>