EpiSpermHis: A Docker container to perform the analysis of sperm histone ChIP-Seq data in Galaxy

Deepak K. Tanwar, Department of Animal Science
Macdonald Campus, McGill University, Canada

January, 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

"Education is what remains after one has forgotten what one has learned in school."

- Albert Einstein

Abstract

The advent of next-generation sequencing (NGS) technology caused a fundamental change in molecular biology, as sequencing the genome is not time consuming. It is well established in the literature that epigenetics and chromatin modifications are important factors that can govern gene activity and are a requirement for normal embryonic development and cell differentiation. The mechanisms of epigenetic reprogramming, which are believed to be essential for normal growth and development, are still poorly understood. Effects of the paternal environment such as stress, diet and toxicants have been linked to various negative outcomes for offspring including birth defects and increased risks for complex diseases. Although it is well known that a woman's preconception health is essential to the development and health of her children, the reproductive health of men has been comparatively ignored. New studies highlight the role of fathers in disease transmission via epigenetic mechanisms (i.e. non-genetic inheritance). While there is growing consensus among researchers that paternal health is critical to that of offspring, little is known about the effect of paternal nutrition on children, or whether potential effects are passed onto future generations. Despite the linkage of paternal effects to birth defects and complex diseases such as cancer, diabetes and infertility, to date, there are no studies linking environmental exposure to environmental factors (e.g. toxins) to changes in the sperm epigenome of men. These paternal effects may occur via non-genetic inheritance, through epigenetic mechanisms including DNA methylation, post-translational modifications of histones and noncoding RNAs (ncRNAs). These epigenetic factors are passed onto offspring through the highly compacted paternal genome in sperm. Sperm has a unique chromatin conformation with the majority of somatic histones being replaced by protamines with only 1% retained in sperm from mice and 15% from men. There is evidence in literature that unlike a typical chromatin immunoprecipitation sequencing (ChIP-Seq) profile generated from targeting a histone modification, there are fewer histone peaks in sperm, and these tend to be distributed over CpG (5'-C-phosphate-G-3') enriched regions. Specific parameters are required for alignment, peak calling and quantitating sperm ChIP-Seq data. The challenges in analyzing and quantifying ChIP-Seq data from sperm with currently available software are the ability to detect and quantify differences, not just in peak enrichment but also the broad domains. The goal of my research is to develop a stand-alone Docker container with Galaxy data analysis pipelines and with tools for quantitative comparison of histone modification levels in sperm of mice and men. The Galaxy framework will allow biologists with no command-line knowledge to perform optimal pre-processing, peak calling and statistical analysis. To address challenges in data

analysis, I have incorporated efficient bioinformatics pipelines for analyzing sperm epigenome data by using currently available tools (Bowtie2, Trimmomatic, Picard tools, MACS2) with optimized parameters and in-house developed tools for statistical analysis, while considering the unique chromatin configuration in sperm. All pipelines have been tested on datasets from mice and men using those available from the lab (see section 4.1 for details) and from public datasets (GSE15594 and GSE79230). All the in-house built tools or scripts are made available through the Galaxy Test ToolShed, only after the tool passed a test. I have included several other tools and R packages for statistical analysis, which would not restrict the users to the provided pipelines, and they would be able to adapt it according to their research requirements. This pipeline will help biologists and clinicians who want to work with sperm ChIP-Seq data, but who have minimal to no bioinformatics background.

Résumé

L'avènement de la technologie de séquençage de nouvelle génération (NGS) a provoqué un changement fondamental dans la biologie moléculaire, le séquençage du génome ne prenant pas beaucoup de temps à être effectué. Il est bien établi dans la littérature que l'épigénétique et les modifications de la chromatine sont des facteurs importants pour contrôler l'activité des gènes et qui sont nécessaires au développement embryonnaire normal et à la différenciation cellulaire. Par contre, les mécanismes de la reprogrammation épigénétique, considérés comme essentiels pour la croissance et le développement normaux, sont encore peu connus. Les effets de l'environnement paternel tels que le stress, l'alimentation et les substances toxiques ont été associés à divers résultats négatifs pour la progéniture, y compris des malformations congénitales et des risques accrus pour les maladies complexes. Malgré qu'il soit bien connu que la santé d'une femme avant la conception est essentielle au développement et à la santé de ses enfants, la santé reproductive des hommes a été relativement ignorée. De nouvelles études mettent en évidence le rôle des pères dans la transmission de maladies complexes par des mécanismes épigénétiques (hérédité non génétique). Alors que les chercheurs s'entendent de plus en plus sur le fait que la santé paternelle est essentielle à la progéniture, on en sait peu sur les effets de la nutrition paternelle sur les enfants ou sur les effets potentiels sur les générations futures. En dépit du lien entre les effets paternels et les malformations congénitales et les maladies complexes telles que le cancer, le diabète et l'infertilité, aucune étude n'a établi de lien entre les facteurs environnementaux et l'épigénome du sperme qui pourrait être transmis à l'enfant et ainsi affecter ses risques à développer des maladies complexes. Ces effets paternels peuvent se produire par transmission non génétique, c'est-à-dire, par des mécanismes épigénétiques tels que la méthylation de l'ADN, des modifications post-traductionnelles des histones et des ARN non codants (ncRNAs). Ces facteurs épigénétiques sont transmis à la progéniture à travers l'épigénome paternel hautement compacté dans le sperme. Les spermatozoïdes ont une conformation unique de la chromatine. En effet, la majorité des histones somatiques sont remplacées par les protamines. Seulement 1% des histones sont donc présentes dans les spermatozoïdes des souris et 15% chez les hommes. Il existe des preuves dans la littérature que contrairement à un profil typique de séquençage d'immunoprécipitation de la chromatine (ChIP-Seq) généré en ciblant une modification d'histone, ces modifications d'histones sont généralement moins abondantes dans les spermatozoïdes et celles-ci sont distribuées principalement dans les régions riches en CpG (5'-C-phosphate-G -3 '). Des paramètres spécifiques sont requis pour l'alignement des séquences génomiques, pour la détection de régions riches à la modification d'histone d'intérêt, ainsi que la quantification des

données générées par le ChIP-Seq du sperme. L'analyse et la quantification des données ChIP-Seq du sperme avec les logiciels actuellement disponibles représentent un réel défi quant à notre capacité à détecter et à quantifier les différences entre des échantillons comparés provenant de différents groupes expérimentaux. Le but de ma recherche est de développer un contenant Docker permettant l'analyse automatisée de données de façon autonome sur l'interface Galaxy incluant des outils de comparaison quantitative des modifications des histones dans les spermatozoïdes des souris et des hommes. L'interface Galaxy, disponible gratuitement et publiquement, permettra aux biologistes d'effectuer l'analyse bioinformatique et statistique de leurs données de ChIP-Seq de façon optimale et rapide, tout en donnant une flexibilité dans les choix de paramètres d'analyse à sélectionner, sans toutefois nécessiter de connaissances de codage. Pour relever les défis de l'analyse des données, j'ai intégré une chaine de traitements bioinformatiques efficace pour analyser les données épigénétiques des spermatozoïdes en utilisant les outils actuellement disponibles (Bowtie2, Trimmomatic, outils Picard, MACS2) avec des paramètres optimisés et des outils internes d'analyse statistique, en considérant la configuration unique de la chromatine du sperme. Toutes les étapes comprises dans la chaîne de traitement de données ont été testées par une série de données de ChIP-Seq de souris et d'hommes générée par notre laboratoire (voir section 4.1 pour plus de détails) et à partir de données publiques (GSE15594 et GSE79230). Tous les outils ou scripts créés lors du processus sont mis à disposition via le Galaxy Test ToolShed, uniquement après que l'outil ait passé un test. J'ai inclus plusieurs autres outils pour l'analyse statistique, ce qui ne limitera pas les utilisateurs de cette chaîne de traitement de données, et leur permettra d'adapter leurs analyses en fonction de leurs besoins de recherche. Ce pipeline aidera les biologistes et les cliniciens voulant travailler avec des données ChIP-Seq sur les spermatozoïdes, sans nécessiter de connaissances au niveau bioinformatique.

Acknowledgements

At the first instance, I deem it to be my bounden duty and pinnacle of happiness to acknowledge **Prof. Sarah Kimmins** and **Jianguo Xia** for supervising my Master's thesis. The main part of my work was carried out from August 2016 to January 2018 in the Kimmins lab at McGill University, Canada.

First and foremost, I would like to thank **Prof. Sarah Kimmins** for accepting me to undertake my thesis in her supervision. I am greatly indebted to her for giving me an opportunity to work in her lab, and to introduce me to the fantastic and hot topics of epigenetics and chromatin. Without her encouragement and numerous enthusiastic discussions, I would not have been able to finish my MSc. I am convinced that my research will be related to epigenetics for a very long time. I also take this opportunity to thank **Dr. Jianguo Xia** to be my scientific co-supervisor, for his great advices, support and constructive comments.

I am extremely endowed to have **Prof. Martina Strömvik** on my M.Sc. thesis committee and, for all valuable feedbacks provided by her during various phases of research and thesis writing.

I sincerely thank to the Graduate studies Director of Animal Science Department, **Prof. Roger Cue**, who was always available for a discussion towards my progress to obtain a Master's degree.

No words deep in my mind I can express to my father, Capt. Sheo Ram Tanwar (retired), mother, Mrs. Munni Devi Tanwar, and Miss. Palni Kundra, whose consistent inspiration and endless love helped me to make all efforts fruitful.

I would like to express my deep sincere gratitude to my colleagues whose direct or indirect support, guidance, and wishes made it possible for me to complete the study successfully.

Montréal, 2018

Deepak K. Tanwar

Contents

List of Figures

List of Tables

Abbreviations

A	Adenine
API	Application Programming Interface
ATP	Adenosine triphosphate
BAM	Binary Alignment Map
bp	Base pair
BWT	Burrows-Wheeler Transformation
C	Cytosine
C57BL/6	C57 black 6
cDNA	Complementary DNA
ChIP	Chromatin Immunoprecipitation
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
CLI	Command line interface
CpG	5'-C-phosphate-G-3'
Ctfr	Cystic fibrosis transmembrane conductance regulator
DE	Differentially Enriched
DMRs	Differentially Methylated Regions
DNA	Deoxyribonucleic acid
DNA-Seq	DNA Sequencing
DTT	Dithiothreitol
EB	Elution buffer
ENCODE	Encyclopedia of DNA Elements
FCS	Functional Class Scoring
G	Guanine
GC regions	Genomic regions rich in Guanine and Cytosine nucleotides
gDNA	Genomic DNA
GMT	Gene Matrix Transposed
GO	Gene Ontology
GEO	Gene Expression Omnibus
GQ	Genome Québec
GREAT	Genomic Regions Enrichment of Annotations Tool
GSEA	Gene Set Enrichment Analysis

GUI	Graphical User Interface
H	Histone
H1	Histone H1
H2A	Histone H2A
H2B	Histone H2B
H3	Histone H3
H3K4	Histone H3 Lysine 4
H3K4me3	Histone 3 Lysine 4 trimethylation
H3K9	Histone H3 Lysine 9
H3K27	Histone H3 Lysine 27
H3K27me3	Histone H3 Lysine 27 trimethylation
H3K36	Histone H3 Lysine 36
H3K36me3	Histone H3 Lysine 36 trimethylation
H3K79	Histone H3 Lysine 79
H4	Histone H4
H4K20	Histone H4 Lysine 20
HATs	Histone acetyltransferases
HKMTs	Histone lysine methyltransferases
HMM	Hidden Markov Model
HMT	Histone methyltransferase
Hox	Homeobox gene
HTML	Hypertext Markup Language
IRB	Institutional Review Boards
K	Lysine
kbp	Kilobase pair
LOWESS	Locally weighted scatterplot smoother
MACS	Model-based analysis of ChIP-Seq
MDS	Multidimensional scaling
me	methylation
me1	mono-methylation
me2	di-methylation
me3	tri-methylation
MNase	Micrococcal nuclease

mRNA	Messenger ribonucleic acid
MSigDB	Molecular Signature Databases
MTHFR	Methylenetetrahydrofolate reductase
N-ChIP	Native Chromatin Immunoprecipitation
ncRNA	non-coding RNA
NGS	Next Generation Sequencing
OS	Operating System
PCR	Polymerase Chain Reaction
PE	paired-end
PRM1	Protamine 1
PRM2	Protamine 2
PTMs	Post Translational Modifications
QC	Quality Check
qPCR	quantitative PCR
Q-scores	Quality scores
R	Arginine
RNA	Ribonucleic acid
RNA-Seq	RNA-Sequencing
SE	single-end
SGE	Sun Grid Engine
SNP	Single Nucleotide Polymorphism
SOLiD	Sequencing by Oligo Ligation Detection
T	Thymidine
TMM	Trimmed Mean of M-values
TNP2	Transition Protein 2
TSS	Transcription Start Site
UQ	Upper Quantile
VM	Virtual Machine
WHO	World Health Organization
X-ChIP	Cross-linking Chromatin Immunoprecipitation
XML	Extensible Markup Language

1 Introduction

Deoxyribonucleic acid (DNA) is compacted into chromosomes in the form of chromatin. The unit of chromatin is the nucleosome that consists of eight histone proteins: H2A, H2B, H3 and H4 (two copies of each), with about 147 base pair (bp) of DNA wrapped around each nucleosome [1]. Gene expression can be affected by various changes in chromatin structure, for example: DNA methylation, replacement of histone variants by canonical histones, post-translational changes in histones, and chromatin remodeling [2].

Emergence of next generation sequencing (NGS) technologies allowed researchers to study histone modifications by performing Chromatin Immunoprecipitation (ChIP), followed by deep sequencing (ChIP-Seq) and finally mapping the modified regions to the reference genome. There could be various histone modifications present in a genome, including methylation, acetylation and ubiquitination. One of which, histone methylation is the most complex, since it occurs in mono, di or tri form (me1, me2 or me3).

Previously, sperm were thought to only deliver the genome to the oocyte at fertilization. Advancements in science helped researchers to understand that apart from the transfer of DNA to the oocyte, the epigenome is also transferred [3]. Various studies in mice, men and zebrafish, have shown that genes related to development, are marked by histone H3 lysine 4 trimethylation (H3K4me3) and histone H3 lysine 27 trimethylation (H3K27me3) [4–7]. It has also been hypothesized that these histone marks are inherited during fertilization [8] through a mechanism called epigenetic inheritance. Infertility is one of the major problems faced by couples around the world, including Canada. It is defined as the problem of not conceiving after unprotected intercourse for one year. It is estimated that 11.5% to 15.7% of couples in Canada are having problems with conception [9], and more than 60% of infertility is due to male factors [10]. In the Kimmins lab at McGill University, the focus is primarily infertility, which can be associated with epigenetic modifications (histone modifications) in the sperm of both men and mice. We study several histone modifications in sperm, and a primary focus is on histone modifications occurring at histone H3 lysine 4 (H3K4). The peaks (represents enrichment of histone modifications in genome) due to H3K4me3 (one of the modifications occurring at H3K4) are narrow in somatic cells but very broad in the sperm [11]. The changes in histone modification levels are studied by combining native ChIP (N-ChIP) and deep sequencing, and mapping the sequenced reads to the reference genome. Computational analysis of ChIP-Seq is intensive and generates huge amount of data, which is further analyzed to find the location of changes in enrichment of a specific histone modification of interest. Several pipelines

have been published in this regard [12–14], but none focuses on sperm histone ChIP-Seq data analysis, which required specialized parameters for alignment, mismatches [15] and peak calling. This thesis focuses on histone modification H3K4me3 occurring in sperm due to the following reasons:

- Histone methylation in sperm serves in the transmission of epigenetic information that influences embryonic development and phenotypes transgenerationally in mice [6].
- Nucleosomes are retained at promoters. Histones in sperm are retained at genes implicated in cell processes, metabolism and embryo development [4].
- Different profiles were observed in fertile vs idiopathic infertile men.

2 Review of the Literature

2.1 Epigenetics

The word "epigenetics" was coined by a British developmental biologist Conrad Hal Waddington in 1942. He also laid the foundation for theories that underlie epigenetics, systems biology and evolutionary developmental biology. All the cells in a multicellular organism have the same genome, and C. H. Waddington defined a model, called "Waddington's model", which describes how genes interact with the environment to produce a phenotype [16]. The word "epigenetics" is made up of two words, "epi" and "genetics", which means "on the top of genetics".

Epigenetic mechanisms can be divided into four major categories: DNA methylation, histone modifications (e.g. methylation, acetylation), histone variants and non-coding RNA (ncRNA). DNA methylation is associated with gene regulation by transcriptional silencing [17], occuring at cytosine residues in CpG (5'-C-phosphate-G-3') dinucleotides. Histone modifications on histone N-terminal tails play a role in either activation or repression of transcription [18, 19], and histone variants, such as non-histones, can be associated with alteration of epigenetic states, such as, the exchange between histones and protamines in chromatin to alter transcription [20]. ncRNA is known to regulate gene expression at the transcriptional and post-transcriptional level [21].

2.2 Sperm cell structure and function

Sperm are the smallest cells in mammals, and are the carriers of the haploid paternal genome to the oocyte. Each sperm is 50 to 60μm in length and made up of three major parts, Figure 2.1 [22]:

The head: The sperm head contains the tightly compacted father's haploid genome and has less cytoplasm, as compared to somatic cells.

The midpiece: The midpiece contains mitochondria, which generates adenosine triphosphate (ATP) to drive tail movement.

The tail: Sperm motile tail consists of microtubules. Flagellar movement in sperm cells is caused by the sliding of microtubules by ATP driven dynein proteins [23].

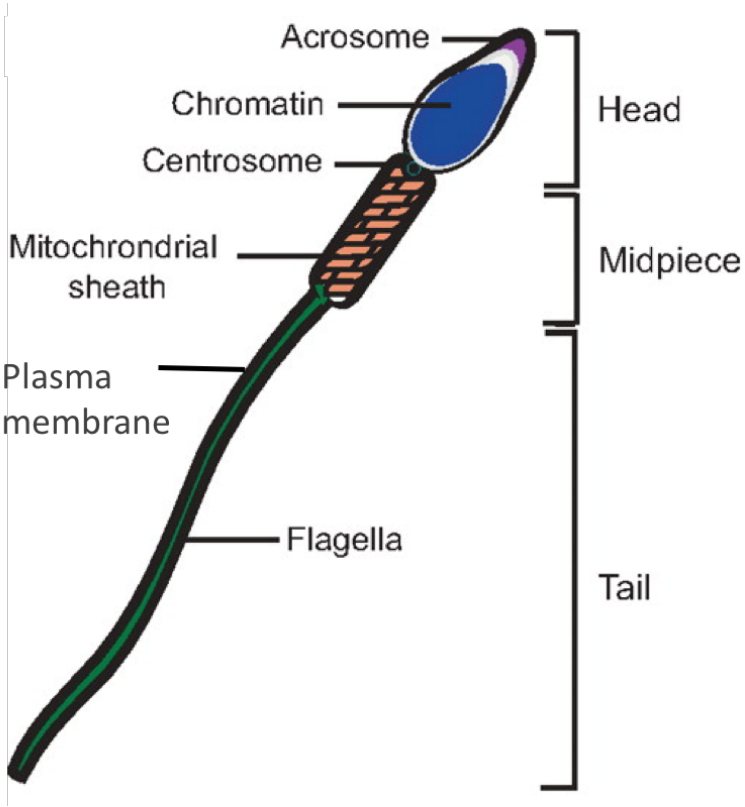

Figure 2.1: Representative schematic of sperm from men. Sperm is composed of three regions. "The head region contains the highly compacted sperm DNA and associated proteins (blue) surrounded by the cellular membrane (black) that in the head region has receptors for recognizing oocyte factors. The acrosome (purple) houses digestive enzymes used for penetrating the oocyte outer layers. The mid-piece region consists of the paternally contributed centrosome (green circle) and mitochondria (tan) used for energy generation. The flagellar tail (green) provides motility" [24]. Reproduced with minimal modifications from [24].

2.3 Spermatogenesis

Spermatogenesis begins with the onset of puberty in males. It is the process by which a developing sperm cell (primary spermatocytes) creates two cells (secondary spermatocytes), and each of those produce two cells, called spermatids, which are further differentiated into spermatozoa. Spermatogenesis takes place in the seminiferous tubules and is divided into three different phases [25]:

Proliferative phase: spermatogonia undergo a series of mitotic divisions and differentiates into primary spermatocytes.

Meiotic phase: spermatocytes divide to give rise to spermatids, and exchange of genetic information by recombination takes place when the haploid spermatids are produced.

Spermiogenesis: also known as post-meiotic phase or differentiation phase, involving a number of morphological events, required for the production of mobile mature sperm.

In the course of spermatogenesis, sperm chromatin undergoes dramatic changes, resulting in the replacement of the majority of histones by special proteins called protamines. This process allows for an extreme compaction of chromatin in sperm, as compared to the somatic cell nuclei [26]. However, the replacement of histone proteins with protamines is incomplete, and it has been shown that, in men, proteins associated with sperm DNA contains approximately 85% protamine, 15% histone and other proteins, in men [27, 28].

2.4 Chromatin Structure

One of the most interesting challenges in the history of chromatin biology has been to understand how DNA is organized in a single cell nucleus. The cell nucleus measures around two meters in length and is contained within 5 to 10 μm in diameter. The cell adjustment is accomplished by organizing DNA into chromatin. DNA and histones are the components of chromatin in eukaryotic organisms. DNA is wrapped around an octamer of histone proteins to form a nucleosome (fundamental repeating subunit of chromatin), and all nucleosomes are connected by a linker DNA (associated with linker histone H1) forming a "beads-on-a-string" structure (11 nm thickness) [29, 30]. Two copies of each of the four different types of histones -H2A, H2B, H3 and H4- form a nucleosome. Chromatin also contains non-histone proteins that facilitates the packaging of DNA, DNA replication, gene transcription and DNA repair [31]. Chromatin is not uniform, with regards to gene distribution and organization. Its organization can be divided into euchromatin and heterochromatin. Heterochromatin is densely stained and is highly condensed, whereas Euchromatin is weakly stained and contains actively transcribed genes [30].

2.5 Sperm chromatin

Sperms are remarkably complex cells having the important task of delivering the paternal genome and associated factors to the oocyte during fertilization. Sperm DNA is important for reproductive success including fertilization and embryonic development. Chromatin in sperm is different as

compared to somatic cells. Most of the genome in sperm is packed with protamines not nucleosomes, though there are a few histones retained in sperm, Figure 2.2. It has been shown in literature that the nucleosomes are retained at GC rich sequences, and are present at the transcriptional start site (TSS) of genes that regulate development [4]. DNA packaging in sperm with protamines occurs towards the end of sperm maturation and, resulting in a very highly compact genome into the sperm head [26].

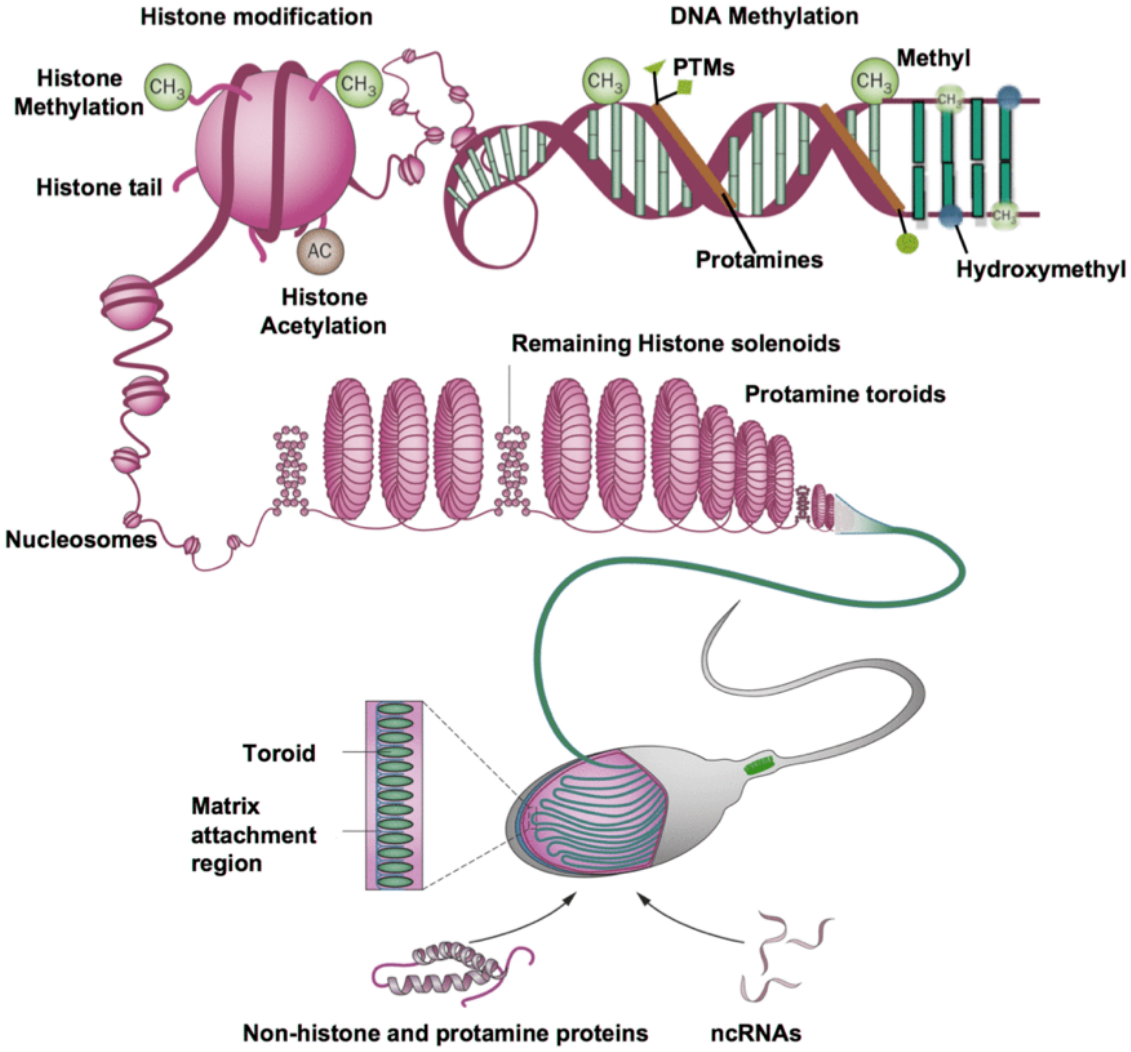

Figure 2.2: Epigenetic modifications in sperm. Sperm has a unique chromatin structure that is unlike any somatic cells as only 10% to 15% of histones are retained in men and only 1% in mice, and the rest are converted to protamines. Both DNA methylation and Histone modifications are found in sperm. Histone tail modifications (methylation and acetylation) are known to play a regulatory role in gene expression. Reproduced with minimal modifications from [32].

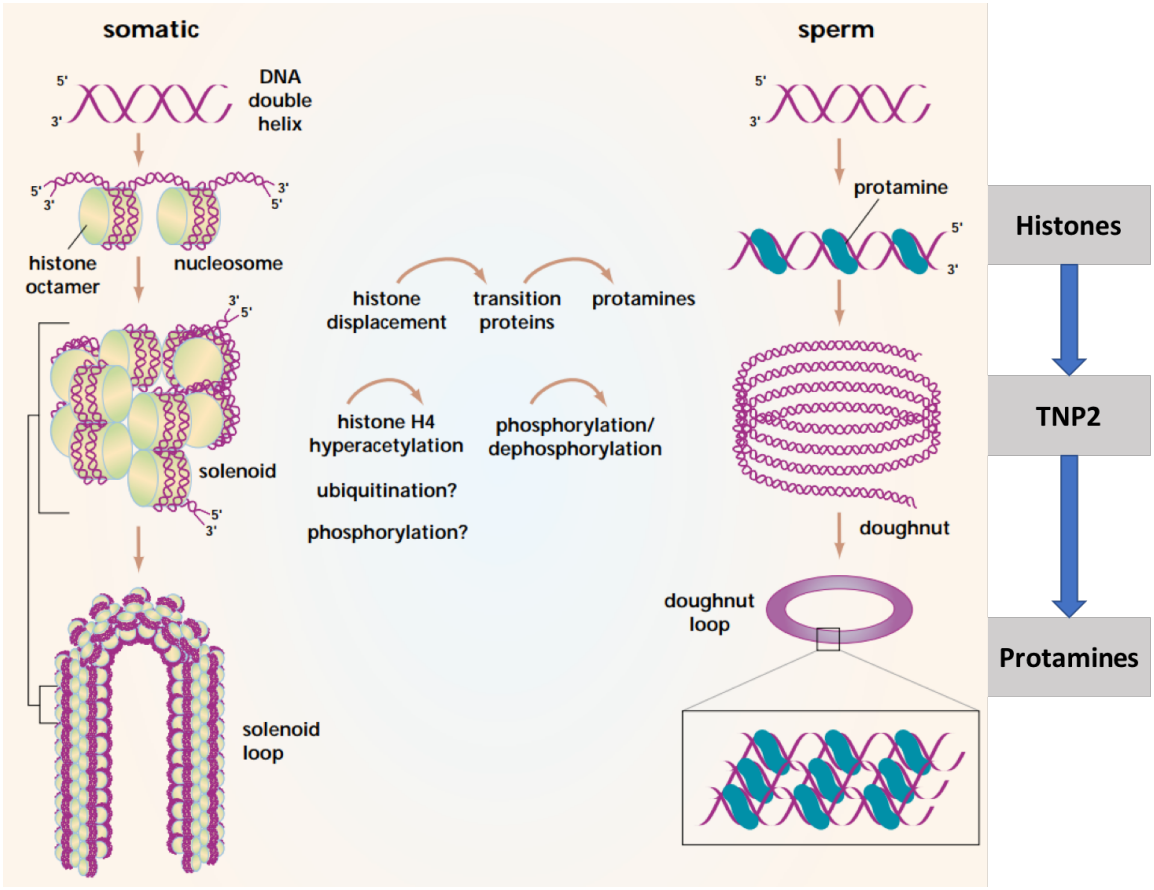

Figure 2.3: Packaging chromatin. "A model of chromatin packaging in somatic cells (left) and mammalian sperm (right). In somatic cells, the DNA is wound twice around histone octamers to form nucleosomes, which are then coiled into solenoids. The solenoids are attached at intervals to the nuclear matrix at their bases and form DNA loop domains. In the sperm nucleus, protamines replace the histones and the protamine-DNA complex is coiled into a doughnut shape. Inset shows the tight compacting of protamine-DNA strands. Displacement of the histones is facilitated by post-translational modifications (PTMs) of the proteins, in the form of histone H4 acetylation, ubiquitination and phosphorylation. Phosphorylation and dephosphorylation of the transition proteins facilitate their displacement before protamines bind" [34]. Reproduced with minimal modifications from [34].

Protamines are small, arginine-rich, nuclear proteins that replace histones late in the haploid phase of spermatogenesis. The protamine gene cluster resides on chromosome 16 in men and mice. It contains codes for protamine 1 (*PRM1*), protamine 2 (*PRM2*) and transition protein 2 (*TNP2*) [33]. These genes are responsible for the correct packaging of DNA into the sperm head, during spermiogenesis [34]. They are believed to be essential for sperm head condensation and DNA stabilization by binding to one turn of DNA helix contrast to ~147 bp nucleosome [36]. In

mature sperm, protamines compact the genome into doughnut-shaped toroids, and each toroid has ~50 kilo base pair (kbp) of the haploid genome [38]. This allows for denser packaging of DNA in the spermatozoon than histones. Differences in chromatin packaging in sperm compared to somatic cells is shown in Figure 2.3.

2.6 Epigenetic Inheritance

Epigenetic information is passed onto children and grandchildren by a mechanism called epigenetic inheritance [3]. Epigenetic modifications are established in mammals by enzymes such as histone acetyltransferases (HATs) and histone lysine methyltransferases (HKMTs) [37], and are altered by environmental factors. Paternally transmitted epigenetic information is passed to oocytes via the sperm and are propagated through their ability to escape epigenome reprogramming that occurs in the zygote and germline. This leads to phenotypic trait inheritance in an offspring that has not been exposed to the environmental factor in question [35].

2.7 Histone modifications

Histones are subjected to PTMs on their N-terminal tails and globular domains. Researches have shown that histone modifications are dynamic during development, vary among different tissues, regulated by specific enzymes, play major role in gene expression, and interact with other epigenetic control systems such as DNA methylation [39]. There are over 60 different amino acid residues on the histone tails, where modifications have been detected [40]. Histones can undergo various modifications, including acetylation, methylation, phosphorylation, ubiquitination and sumoylation among others [41]. Histone methylation occurs on Lysine (K) and Arginine (R) amino acids by histone methyltransferase (HMT) enzymes. Histone lysine residues can be mono-methylated, di-methylated or tri-methylated, and modifications can act as either active or repressive marks. Methylation of H3K4, H3K36 and H3K79 are linked with activation of gene transcriptions, and methylation of H3K9, H3K27 and H4K20 are correlated with repression [42, 43].

2.8 CpG islands

CpG sites are regions of DNA where a cytosine (C) nucleotide occurs next to a guanine (G) nucleotide in the linear sequence, separated by a phosphate which links the two nucleotides together in DNA. The "CpG" notation is used to distinguish a cytosine followed by guanine from a cytosine base paired to a guanine. CpG islands are found near the TSS of genes and are not distributed uniformly in the genome. They are also found at the promoter site of housekeeping genes [23]. The human genome contains approximately 20,000 CpG islands in the promoter of first exons, and marks the 5' ends of the genes [23]. Over 60% of human genes have CpG islands in the promoter site; over 80% CpG islands have no genes and therefore do not regulate expression [44].

2.9 Chromatin Immunoprecipitation

The study of the interaction between a DNA sequence and a histones can be assessed using a technique, called ChIP, followed by sequencing. ChIP can be performed in two ways in sperm: cross-linked ChIP (X-ChIP) and native ChIP (N-ChIP). In X-ChIP, formaldehyde is used to cross-link histones to the sperm DNA and ultrasound sonication is performed to fragment the DNA, whereas in N-ChIP, the native covalent bond between histone and DNA is used. The most important feature in N- ChIP is the fractional generation of two phases: soluble phase with histones and insoluble phase with protamines [45], Figure 2.4. The genomic DNA (gDNA) is extracted and further digested by Micrococcal nuclease (MNase) to obtain DNA fragments, which is followed by ChIP. Samples are treated with Dithiothreitol (DTT, C4H10O2S2) to obtain chromatin. The chromatin is then treated with MNase enzyme to digest free DNA, after which immunoprecipitation with antibodies targeting the chromatin modification of interest such as, H3K4me3 is performed.

2.10 Chromatin Immunoprecipitation Sequencing

ChIP-Seq is a powerful technique to study the physical interaction between proteins and DNA on a genome-wide scale by coupling ChIP with high-throughput sequencing [46, 47]. After performing ChIP on samples, the next step is library preparation, followed by polymerase chain reaction (PCR) amplification, and finally deep sequencing. Sequencing can be either single end (reads generated by sequencing DNA fragment from one end) or paired end (single DNA fragment sequenced from both ends). Both single-end (SE) and paired-end (PE) type of sequencing are used for ChIP-Seq.

Although PE sequencing has been widely used in DNA sequencing (DNA-Seq) and RNA sequencing (RNA-Seq) experiments to uncover genomic rearrangements and genomic structural variation, but it is not clear how a ChIP-Seq experiment can be benefited from PE libraries [48]. When there is low cell input, PE can improve the number of mappable reads and the quality of the data. According to the Encyclopedia of DNA Elements (ENCODE) ChIP-Seq guidelines, a ChIP-Seq experiment should have 20 million DNA reads mapped uniquely (without duplicates) [49]. As of today, the combination of ChIP with NGS technology has already generated plenty of ChIP-Seq data. These experiments produce high volume of data that must be pre processed before any further analysis can be carried out, such as, peak calling and downstream analysis for quantitative analysis of ChIP-Seq data.

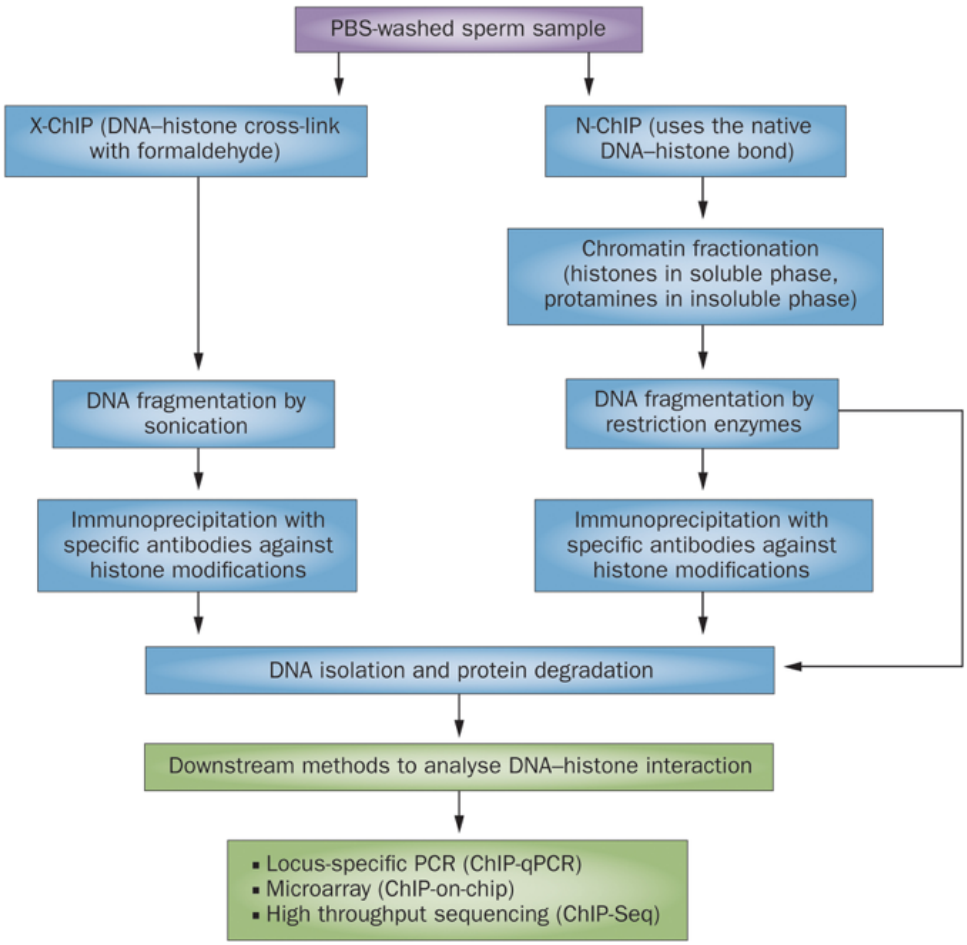

Figure 2.4: Analysis of DNA–histone interaction in sperm. Taken from [45].

2.11 Next-generation sequencing

DNA sequencing is performed to determine the order of nucleotides in a DNA fragment ("A", "T", "G", "C"). This technique is helping biologists around the world in a broad range of applications such as molecular cloning, breeding, pathogenetics, gene regulations, comparative evolution and epigenomics. In the last three decade, DNA Sequencing technologies and applications have undergone tremendous development. In the last few years, the application of semi-automated Sanger sequencing for genome analysis has been replaced by NGS methods, as they allow for parallel sequencing, producing thousands or millions of sequences at once.

The three most common massively parallel sequencing systems launched and that have high accuracy and low costs are: 1) 454 (now Roche), which came out after completion of human genome project, 2) Solexa (now Illumina) which released a genome analyzer the very next year, and 3) SOLiD (Sequencing by Oligo Ligation Detection) which was provided by Agencourt (now Life Technologies). All the three platforms work on different chemistries for sequencing and give different outputs, and they are competitive and exhibit better performance and accuracy [51]. The ability to sequence millions of DNA fragments in less than one day is the major advancement in NGS. In summary, the huge amount of low-cost sequenced reads makes NGS technologies useful for several applications.

Sequencing can be either SE or PE. In SE sequencing, DNA is sequenced from one end only (5' to 3'), whereas in PE, both ends are sequenced (5' to 3' and 3' to 5'). There are some advantages of sequencing longer and PE DNA reads, as opposed to SE and short reads. PE sequencing is done in order to achieve optimal sequencing depth when less input DNA is present for library preparation (for example, in the case of embryos). It is further helpful in identifying duplicates, but, still, it is unclear whether the duplicate has arisen from amplification bias or not. One way to avoid PCR duplicates is to reduce their generation by using a good amount of input material and keeping amplification to a minimum.

2.12 Library generation

Before ChIP-Seq is performed, libraries are created by amplifying templates, in order to allow for massively parallel sequencing (by illumina technology), which will be used in sequencing-by-synthesis [52, 53]. Then a short PCR cycle with primers for the adapter sequence is performed to

generate a library of adapter-ChIP fragments [53, 54]. The adapter sequences added to each ChIP fragment contain universal priming sites and allow different and complex genomes to be amplified using the same common primers [53]. Sample DNA fragments are then denatured to produce single strands and hybridized to oligonucleotide sequences, complementary to their adapter sequences. These are further immobilized on the surface of a flow cell [55]. Multiple cycles of this process result in the generation of clusters of clonally amplified templates [53, 55].

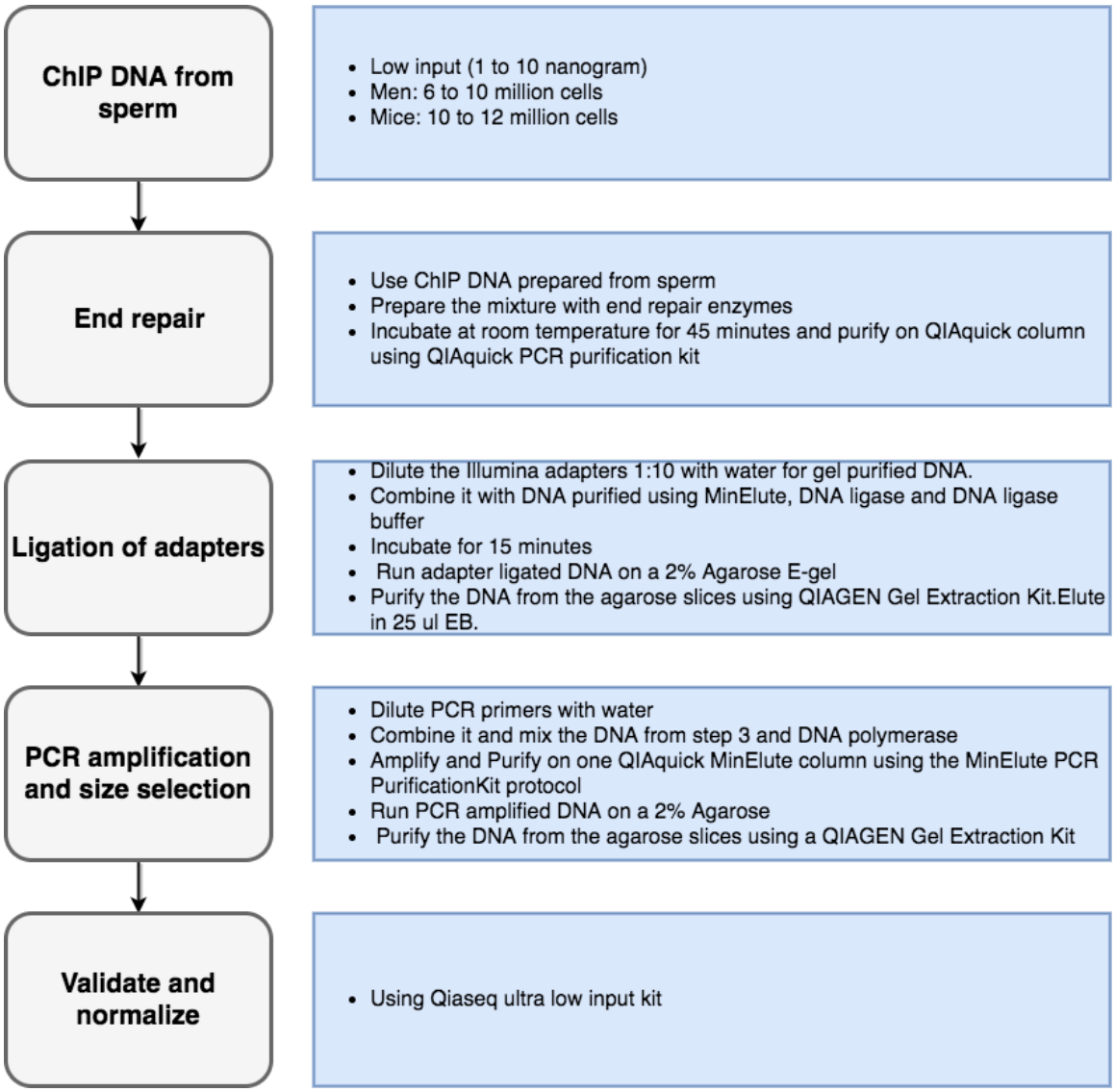

Figure 2.5: Library preparation steps using low input kit from Qiagen (Qiaseq). This protocol is most frequently used in the Kimmins lab [6].

In the Kimmins lab, a Qiagen library preparation kit is used for library generation, Figure 2.5, and then they are sent to Genome Quebec (GQ) for sequencing. GQ provides sequencing results to us in either FastQ format or Binary Alignment/Map (BAM) format. A FastQ format is a format of NGS file, which stores information of raw reads (nucleotides) and their quality scores (Phred).

2.13 Multiplexing samples with barcodes/ indices

High-throughput sequencers sequence millions of reads in a single reaction. Considering time and cost, it is often desirable to pool ("multiplex") libraries from multiple experiments into a single sequencing reaction. For example, a single flow cell lane from an Illumina HiSeq 2000 instrument, for example, routinely yields 150 million sequences. Most ChIP-seq libraries do not require that level of coverage, so it makes sense to combine samples to save costs. To identify the sequence (by the library it comes from), each library is prepared using adapters containing different index or barcode (tags), which are typically short (5-7 bp) sequences and, are read during sequencing. Two kinds of sequence tags are commonly used, barcodes and indices. Various available software packages can be used to sort the sequence reads by their barcode. The Illumina sequencing platform also supports a second type of tag (indices). The index is located within the 3'-adapter and is read using a separate sequencing reaction that follows the primary sequencing reaction. After sequencing, each read may be traced back to its original sample using the index sequence and is binned accordingly, called demultiplexing.

3 Rationale and Objectives

Sperm has a unique chromatin structure in that most of the histones are replaced by protamines. Protamines are required for sperm head condensation and DNA stabilization. Histones carry multiple PTMs, such as lysine acetylation, lysine and arginine methylation, and serine phosphorylation. Lysine methylation can have different effects on gene expression depending on which residue is modified. For example, genome-wide chromatin analysis in mice and humans demonstrates that H3K4 is associated with transcribed chromatin, whereas histone H3K9 is symbolic of gene repression. Furthermore, histone H3 methylation can take the form of either mono-, di- or tri-methylation. Histone modifications are dynamic, and they can be established and removed by various enzymes. More than 50% of histones are localized in intergenic regions, while approximately 25% are localized at promoter regions. In mice, the sperm epigenome, specifically H3K4me3, can influence offspring development and health [6].

Using currently available software to analyze and quantitate ChIP-Seq data, to detect differences in broad domains (not just at peaks) of sperm poses a number of challenges. Although several pipeline exists for ChIP-Seq data analysis, none of them are specific to sperm. Further, there is a lack of graphical user interface (GUI) based pipelines for data from genomic experiments. The objective of this research project is to develop and to validate a bioinformatics pipeline, with optimized parameters for the efficient analysis of sperm ChIP-Seq data. The development of this pipeline will allow for the detection of differences in sperm epigenome. I aim to:

- build an efficient bioinformatics pipeline for analyzing sperm ChIP- Seq data
- improve or develop a (new) peak calling method that is customized to the unique epigenome profile of sperm
- make the pipeline available as a complete independent tool

4 Methods, Tools selection and Results

4.1 Data

For the development of pipelines, the following data were used from Kimmins lab in this thesis:

1. **Reference population sample, fertile and infertile samples from men:** These samples were obtained from the CReATe fertility clinic in Toronto to study the epigenetic factors linked with idiopathic infertility. Thirty men were pooled (24 fertile and 6 infertile) to create a reference population sample. These men were between 27 years and 61 years in age: 6 were smokers; all had Methylenetetrahydrofolate reductase (*MTHFR*) 677 polymorphism (risk factor for male infertility) CC: 13, CT: 2, TT: 15; and 6 men had a high dose folic acid supplementation (5 mg). We had 5 samples from each group (fertile and infertile): 4 were pooled and one individual. All, fertile and infertile men had normal semen parameters, as determined using the criteria of semen analysis established by the World Health Organization (WHO). Men were categorized as idiopathic infertile as they were not able to produce babies, provided no female factor was involved.
2. **High-fat and control samples from mice:** These samples were from high-fat diet fet and control "C57 black 6" (C57BL/6) mice that were obtained from the Sloboda lab at McMaster University to study the epigenetic factors linked with obesity. Control mice were fed regular chow diet and high-fat mice were fed 60% kcal fat for 10 weeks (starting at 6 weeks of age). For the preliminary study, we obtained three biological replicates for each group. This project have animal care (McMaster University) and Institutional Review Boards (IRB) approval from the University of Toronto and McGill University. Note: no samples, only data were used for this thesis.

4.2 Methods

4.2.1 Preprocessing of ChIP-Seq data

4.2.1.1 Quality Control (QC)

After receiving sequenced data from GQ, the first step was to assess the quality of the data to determine the reliability of the sequenced reads. To examine the sequencing errors and trimming

low quality reads, the quality assessment of the raw data was performed using FastQC [56], a widely-accepted software. FastQC gives a capital overview of the data, Figure 4.1 including but not limited to the number of sequenced reads, base sequence quality, 'N' content (whenever a sequencer is unable to call a base, it will input 'N'), sequence duplication levels and adaptor information. A Hypertext Markup Language (HTML) report generated by the software helps to understand the pre-processing steps required.

FastQC

Measure	Value
Filename	Reference_sample
File type	Conventional base calls
Encoding	Sanger/ Illumina 1.9
Total sequences	143,077,301
Sequences flagged as poor quality	0
Sequence length	100
%GC	47
Duplicates	37,625,112
Sequencing type	Single end

Figure 4.1: Basic statistics of reference population sample.

One of the important steps in quality control is to assess the depth of sequencing. The datasets used in this thesis had at least 20 million uniquely mapped reads, as recommended by ENCODE guidelines [49]. ChIP-seq enrichment profiles are expected to saturate after 20 million sequenced reads, in terms of enrichment regions [57].

The quality assessment is done using the quality scores (Q-scores) in the raw files, which are translated to quality statistics in FastQC reports and plotted as boxplots, as shown in Figure 4.2. Q score is defined as the base calling error probabilities and is calculated by the formula: **Q = -log10 (*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 NGS [58].

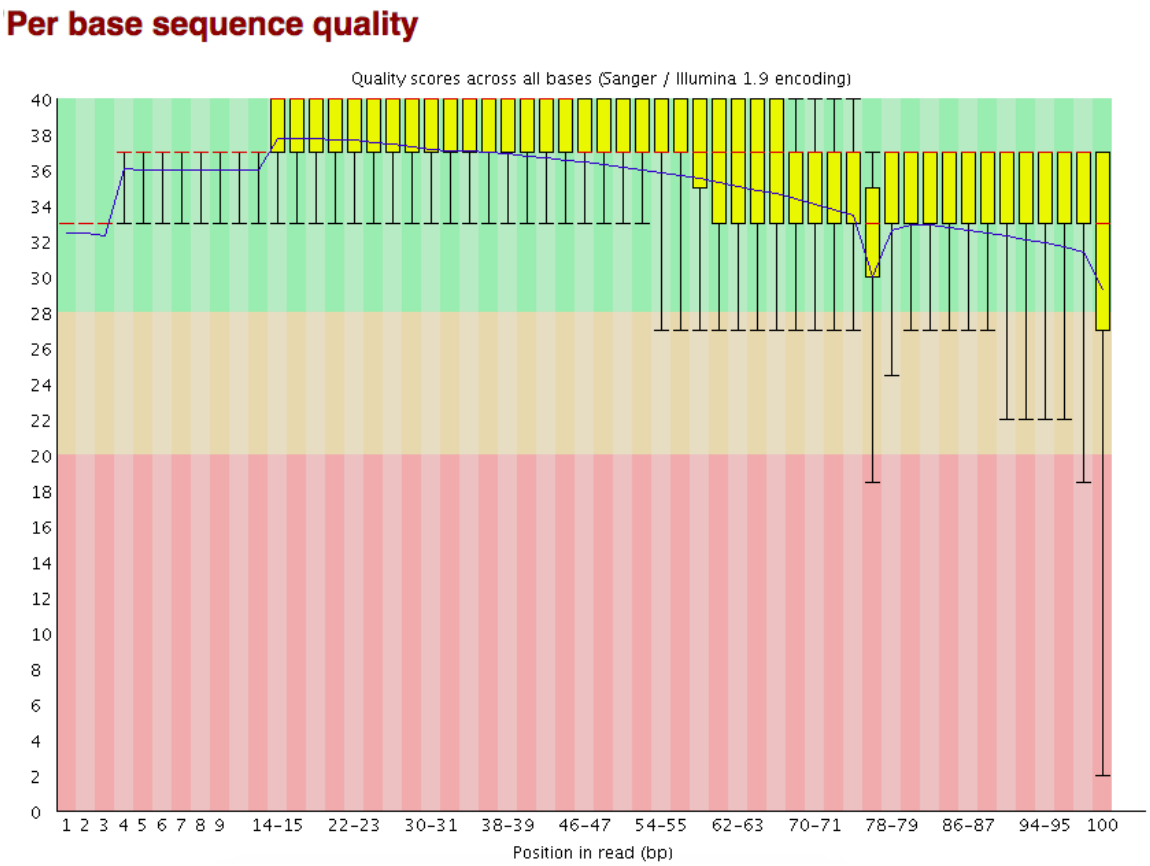

Figure 4.2: Per base sequence quality from FastQC: Boxplots represents the bases. If the boxplots are in the green area, the reads are very reliable; if in orange, they are not so reliable and if in red, they are not at all reliable.

4.2.1.2 Adaptor and read trimming

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 and include only longer reads in the analysis, as shorter reads may have low quality due to

technical errors and will lead to misinterpretation. In our lab, we select reads with minimum length of 80% of the sequenced reads, by that we get rid of short reads, and increases the quality of the data. Although the QC steps are multiple processes, they can be handled easily by one software, Trimmomatic [59].

4.2.1.3 Alignment of sequenced DNA reads (tags alignment)

The alignment of the reads to the genome is a fundamental step that must accurately assign the sequenced reads to the right position in the genome where they were generated. The sequenced DNA reads are just big chunks, unless we know where they belong (mapping and annotation). To provide some meaning to the sequenced DNA reads, we align them (mapping) to a reference genome. The two major issues in the alignment of ChIP-Seq data are: a tag can be mapped to multiple locations in the genome and multiple tags can be mapped to the same locations in the genome. The simplest solution to the first problem is to ignore the tags mapped to multiple locations in the genome as it is not possible to determine their true origin, and for the second problem to take one instance of multiple tags aligned to the same location. An exception to this would be if researchers are interested in repeat regions, for example, H3K9. The most common software (highly cited) used for alignment are BWA [60], Bowtie [61] and Bowtie2 [62] and these have implemented Burrows-Wheeler Transformation (BWT) algorithm [63] for alignment.

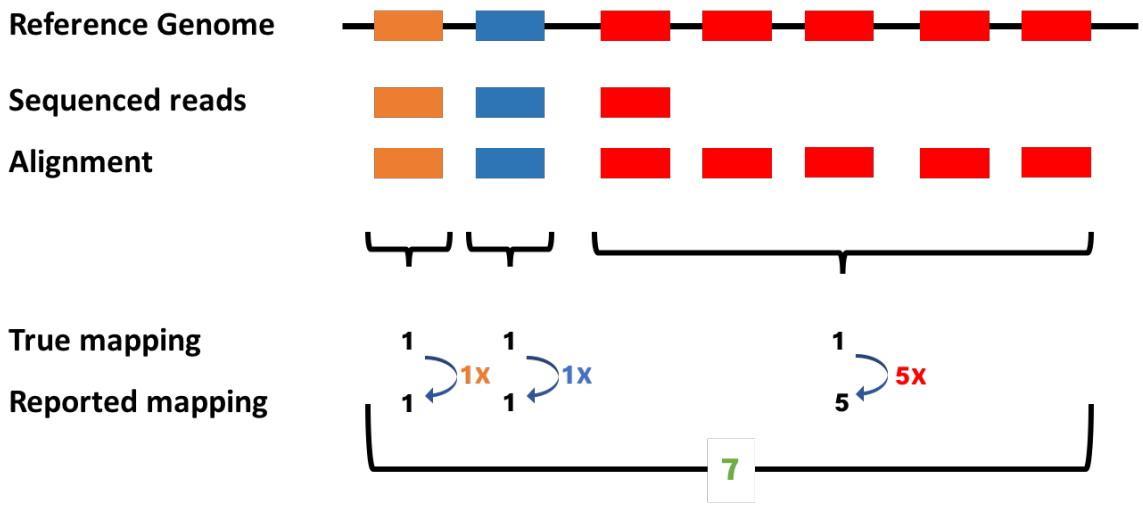

Figure 4.3: Optimal alignment parameters required for aligning sperm ChIP-Seq data. Analysis Pipeline from [64] leads to an artifactual signal in repetitive DNA elements [15]. Reproduced with minimal modifications from [15].

As sperm chromatin structure is unique, we need to align sperm ChIP-Seq reads with optimized parameters. Royo et al. [15] analyzed the parameters used in [64] and found that the pipeline leads to an artifact signal in repetitive DNA elements, as shown in Figure 4.3. The main disagreement discussed in [15] is using the bowtie parameter "-a". Saman et al [64] is counting each multimapping read multiple times. Basically, if a read maps to 10 positions in the genome, it will be assigned to all 10 positions. Royo et al. [15], when re-analysing the data, found that this affects reads mapping to repeat regions disproportionately with a large increase in mappings to those regions. This is not biologically relevant in the case of sperm, as very few histones are retained in sperm. The authors of [15] have compared various tools and parameters for aligning raw sperm reads. They suggested the use of bowtie for aligning sperm deep sequenced data by allowing a mismatch/ Single Nucleotide Polymorphism (SNP) of 3 and alignment limit of 100 per read.

Advancement of NGS has helped researchers to sequence longer reads and Bowtie2 is mainly designed to handle the reads which are longer than 50 bps, while using less memory and being more sensitive and faster. I have duplicated the alignment parameters from [15] for sperm ChIP-Seq data alignment for bowtie2, originally stated for bowtie, which is an advanced version of bowtie and has greater compression as compared to bowtie. By default, Bowtie2 reports a single alignment per read for reads that map multiple times i.e., duplicated. But, Bowtie2 does not allow to filter out the reads with a higher number of SNPs; therefore, in-house script was written to filter reads (with more than 3 SNPs) after alignment (see 4.2.1.4).

4.2.1.4 Mismatch removal

The genome of human beings differs from one another by a minute percentage. If we randomly select two people and sequence their genomes, we will find that they differ by one nucleotide base per 1,200 to 1,500 DNA bases. If we consider differences in the whole genome, one person would be different from another for around three million DNA bases and although this number seems bigger, it accounts for only 0.1% (any two individuals have 99.9% same DNA bases) [65]. The question that arises is; how many dis-similarities should we allow in sperm for ChIP-Seq data analysis? A published study [15] suggests that for sperm, we should allow a maximum of three mismatches in the sequenced DNA reads of 100bp. By allowing up to three mismatches or, by using an alternative "ENCODE-unified analysis pipeline" (tolerating up to five mismatches), the percentage of mapped reads increases. Allowing for some mismatches is acceptable as we already know that humans differ from each other with approximately 0.1% of DNA sequence. Therefore, in

the designed pipeline 3 SNPs are permitted per 100 bp.

4.2.1.5 Duplicate sequenced DNA reads

A duplicated read in ChIP-Seq data can arise as a PCR artifact, or via reading of the same fragment of DNA twice. It is debatable if duplicated reads are to be removed or not from ChIP-Seq data [66]. Most of the NGS pipelines recommend removing duplicates, or at least marking them for peak calling, as duplicates may lead to a false enrichment profile. It is suggested to retain marked duplicates for downstream analysis, as it decreases enrichment overall. PCR amplification is carried out to intentionally multiply the number of reads to ensure enough coverage. It has been suggested that PCR cycles should not exceed 6 so that the library retains a high degree of complexity. This will result in only 4% of duplicates [67]. The percentage of duplicate reads in the sequenced samples is highly dependent on the depth of the library sequenced. The main aim behind removal of duplicated reads is to mitigate the effects of PCR amplification bias which is introduced during library preparation. Proper duplicate removal is fully dependent on a low error rate in the library. On the other hand, sequenced reads which are mapped to the genome may indeed have the same genomic coordinates, reads sequenced can be of same length and have identical mapping coordinates, but can still belong to different genomic coordinates. The two identical mapping reads with a SNP in the complementary DNA (cDNA) fragments may have been derived from both the parents, and these two cDNA fragments, which differ by just one SNP, could contain biological information that would be erased while removing duplicates on the basis of mapping coordinates [67].

4.2.2 Peak calling

The aim of peak calling is to select real peaks over false positive peaks. Real peaks correspond to ends of the binding regions for both strands [68]. In some cases false positive peaks correspond to a contamination process that may happen during the experiment or, due to a faulty peak caller. Usually, a ranking procedure takes place to identify best peaks among all other peaks. Also, visual examination using a genome browser is required to see if peak calling is accurate. In some software, peaks are ranked based on p-values, as in Model-based analysis of ChIP-Seq (MACS) [69] and MACS2. Peak calling process varies between systems in terms of implementation. In a two samples experiment, the control data is generated by doing the experiment without including any antibody targets.

The data from ChIP-Seq are represented as aligned stacks of reads which represent the enrichment of the targeted epigenetic factor in the genome. Some of these regions represent signals of interest where histone proteins occur, and others could be experimental bias. Based on performance, various papers discuss benchmarking of peak calling methods [70–74]. Previous research has also considered width and number of peaks [75]. As well, a recent publication [76] tested 30 different methods for peak calling for transcription factors and histone modifications in somatic cells and identified the features that define the best peak caller. According to the publication [76], peak calling can be subdivided into two problems: "identifying peaks" and "testing peaks for statistical significance". The authors have also described various features that best describe peak calling methods and checked for the presence of these features in various peak callers. **Importantly, there is no previous study which describes a best peak caller for sperm, though there are various software available for broad peak-calling: MACS2 [69], based on cross-correlation; RSEG [77], based on Hidden Markov Model (HMM); and BroadPeak [78], based on segmentation algorithm.**

4.2.2.1 MACS2

MACS is a tool for peak calling within a single ChIP-Seq experiment while modeling the ChIP generated read distribution along the genome with a Poisson distribution [69]. MACS2 scans the genome with a sliding window of fixed width centered on each nucleotide. Candidate peak regions are identified using a Poisson test of regions where the numbers of reads in the ChIP sample are greater than what one would expect given a fixed background rate (gDNA/ input). These regions are then ranked using a Poisson test that assesses if reads in the ChIP sample are greater than what one would expect, given a local background rate estimated from the input sample. In case of no gDNA/input sample, the background threshold is calculated from around the peak.

4.2.2.2 BroadPeak

BroadPeak is based on a Maximal Segmentation algorithm, in which the data to be binned into non-overlapping 200 bp windows and then evaluated with the high-tag bins compared to the genomic background density (low-tag bins) [78]. Adjacent genomic bins with high tag counts are put together to form the broad peaks. The method can be used as unsupervised as well as supervised (machine learning). One of the major advantages of this software is that it does not use input/ gDNA to filter out the background.

4.2.2.3 RSEG

RSEG is a HMM based algorithm which aims to identify the genomic regions marked by broad histone marks such as H3K36me3 and H3K27me3 [77]. The whole genome is segmented into non-overlapping bins of a defined size and reads within each bin are counted to obtain signal tracks. Resulting signal tracks are further filtered for unmappable regions and remaining signals are adjusted.

Figure 4.4: Peak calling in sperm ChIP-Seq data from men. Peak calling was performed using MACS2, RSEG and BroadPeak methods. MACS2 identified the true peaks, though the boundaries of peaks were not so appropriate. The correct boundaries which could have been identified, are marked with arrows after visual inspection (orange: peak-start, blue: peak-end) for reference sample.

Earlier, the peak calling parameters were assessed in the Kimmins lab by Keith Siklenka (a graduate student) using Homer, MACS1.4 and MACS2, and MACS2 with "–broad" and "–broad-cutoff 10^{-6}" were found to be most reliable for peak calling in sperm ChIP-Seq data. Further, I have tested RSEG and BroadPeak methods to optimize the peak calling in sperm ChIP-Seq data from men and mice (from the lab), Figure 4.4, and it is concluded that peak-calling using MACS2 with "–broad" and "–broad-cutoff 10^{-6}" parameters is most optimal, consistent to what Dr. Siklenka found and is implemented in the pipeline.

4.2.3 Statistical analysis

4.2.3.1 Calculating read counts in the samples

To calculate the read counts from pre-processed alignment files (bam files), the pipeline was tested using an approach where we bin the samples into 200 bp windows, keeping an overlap of 50 bp between windows, and then count the number of reads in 200 bp windows in all samples using csaw [79]. A human haploid genome consists of 3.2 billion nucleotide bases and is binned into approximately 21 million 200 bp windows. Although, binning and calculating read counts in 21 million windows is computationally very intensive, this task can be optimized to perform in less than 30 minutes, using "data.table" and "csaw" packages from R. This reduces the original time by many folds (one week to 30 minutes), and can be faster or/ slower depending on the access to computational power (servers/ clusters).

4.2.3.2 Normalization

The most widely used normalization method for ChIP-Seq data is adjustment of total counts (sequencing depth). However, this approach has a drawback, if the signal-to-noise ratio is very different between two libraries, one library is going to contain more background reads than the other. A classical method used to normalize RNA-Seq data is Upper Quantile (UQ) normalization method [80], where "raw count" values are divided by the 75th percentile of the column (after removing zeros) and multiplied by 1000 (a scaling factor). The normalized file therefore does not take any external factors into account, but simply transforms each sample so the values are relative to the 75th percentile with a x1000 adjustment factor. This is based on the hypothesis that most genes/ regions are not Differentially Enriched (DE) [80].

A new method, called Trimmed Mean of M-values (TMM), also hypothesizes that most genes/ regions between samples are not DE [81]. TMM performs UQ normalization and estimates the scaling factor (based on sequencing depth). The next step is to calculate the weighted trimmed mean of the log enrichment ratios, which accounts for the variability among samples.

Another method, voom [82] was developed in 2013 and can be used to correct over variance across samples. Voom estimates the mean variance of the logged read counts (non-parametrically) and uses it to predict the variance of each log-cpm (Counts Per Million, also used in TMM method) values. Variance is then encapsulated as an inverse weight for the log-cpm value and the weights are incorporated into a linear modeling procedure. We found that this method corrects for variance

in the genes/ regions which have very low enrichment.

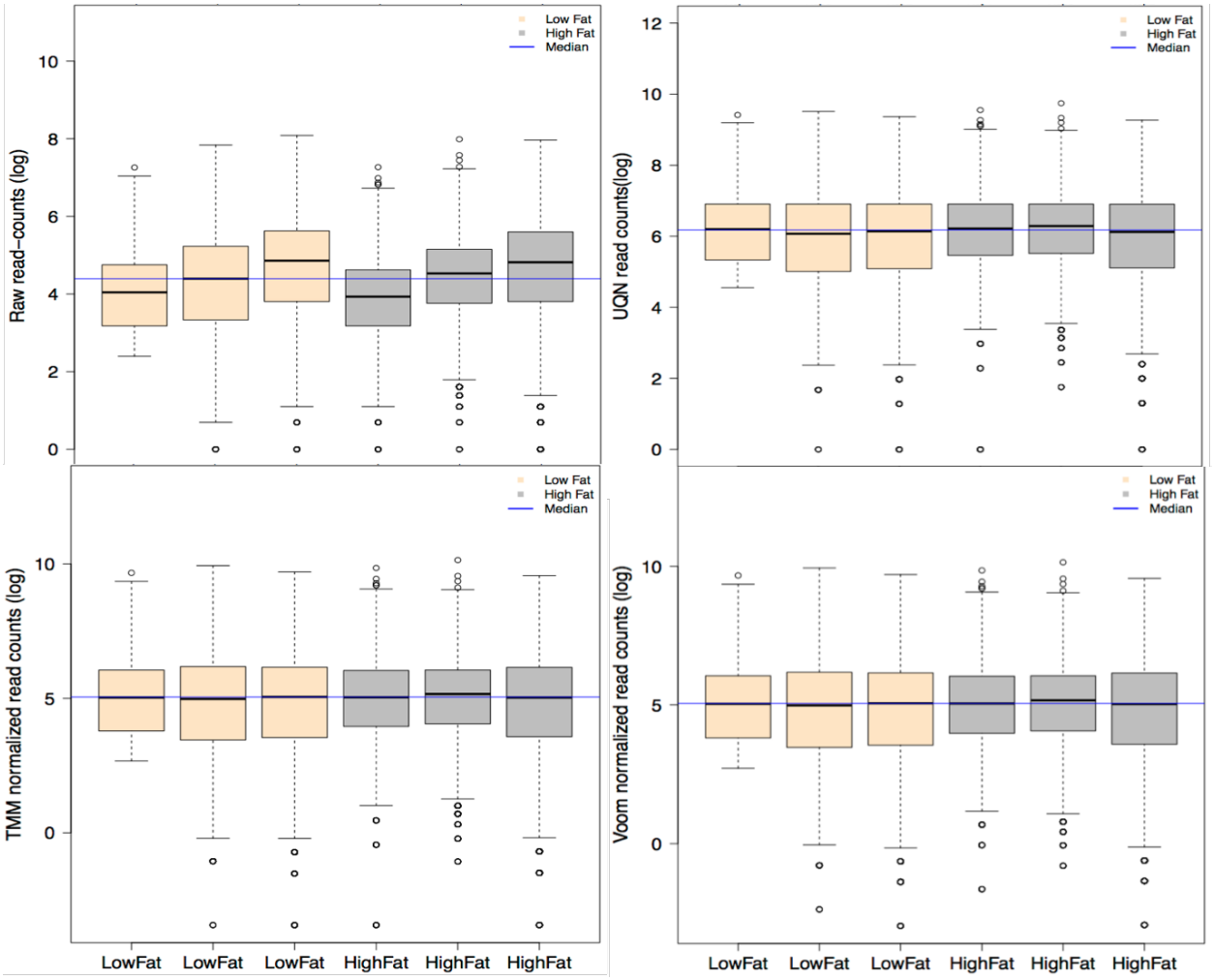

Figure 4.5: Normalization methods. Different normalization methods were tested in all the datasets, and TMM normalization works best in normalizing read-counts. These plots shows the testing of UQ, TMM and voom normalization methods on high-fat and control mice data from the lab.

Correct normalization methods are essential for comparing the two independent groups ChIP-Seq data. I have compared UQ, TMM and voom normalization methods, as shown in Figure 4.5, and I propose that TMM method is the best method for normalization of ChIP-Seq data for the following reasons:

- As many genes/ regions have low enrichment, it did not consider them and this is helpful while performing differential enrichment analysis.

- It performs library normalization (correcting for sequencing depth).
- It corrects for the variance of very high and very low enriched genes/ regions (outliers).
- Although voom can correct for further variance between samples, we observed that it only corrects for the genes/ regions which have very low enrichment values (and those are not of that much importance).

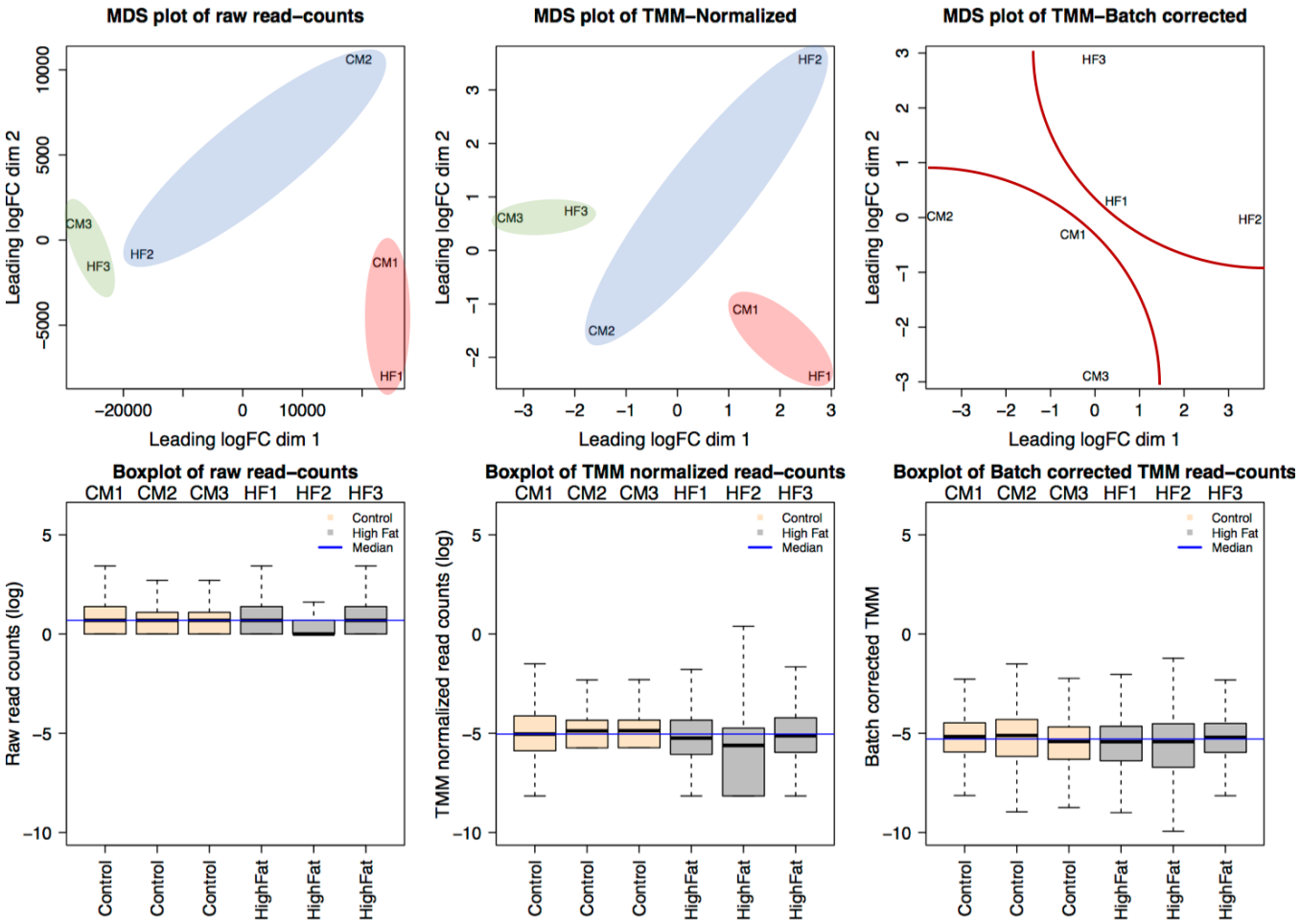

Figure 4.6: Batch effect correction using limma. Batch effect (day bias) was observed in the high-fat and control mice data from lab, and limma (a R package) was used to correct for the batch effect. On the top, multidimensional scaling (MDS) plots represent the grouping of controls and high-fat samples. Batch effect (days bias) can be seen in the raw and TMM normalized data, which was not observed after performing batch correction. At bottom, boxplots of read-counts from high-fat and control mice sperm ChIP-Seq data are shown.

Batch effects are technical experimental data variations that are introduced by processing samples in multiple batches, and can create confounding biological meaningful variations. To address the non-biological variance in data, I used limma [83], which models the batch correction via linear model. As limma method works on linear model, various effects can be modelled together, such as days and batches, in which the experiment was performed. Moreover, the batch correction is performed while retaining the group differences. Limma was tested to correct for the batch effect (day bias) in high-fat and control mice data from the lab, as shown in Figure 4.6

4.2.4 Differential analysis

Detection of differentially enriched regions or genes in ChIP-Seq data is a very important step to find the changes in protein-DNA binding in different samples. A number of tools have been reported to either detect differential gene enrichment or delimit differentially methylated regions (DMRs) between different conditions [69, 84]. A recent review [85] has identified the major problems faced by these tools in identifying the DMRs:

- DMRs are not limited to a certain genomic coordinate (for example genes) and, they could be widespread in intergenic regions. Hence, we cannot restrict our search. The changes in the enrichment (signal) is not constant in data and transformation is required (normalization).
- It is challenging to detect the changes when variation in the data is large and differences are subtle.
- Histone marks have different bindings with the DNA (width), and the boundaries are not easy to detect.

It is hard to rely on the output from differential peak calling because of the problems stated above. A recent tool, "csaw" [79] published in 2016, gives the power of analyzing the full genome by using a window approach. csaw also provides the power of binning the samples into the same windows (with or without overlaps) and analyzing them for differential binding, as shown in Figure 4.7. Similar to edgeR [86], csaw uses TMM normalization and also implements lowess-based (LOcally WEighted Scatterplot Smoother) normalization to remove trended differences in window counts with respect to abundance and analogous to cyclic loess normalization for microarrays. As mentioned previously, csaw package uses TMM normalization to account for normalization factors and uses edgeR to find the differentially enriched windows. csaw suggests the right pipeline for

analyzing differentially enriched regions by dividing them into small bins, as the differential regions are not necessarily present at certain genomic intervals. The proposed method in the pipeline for differential analysis is to bin the samples into 200 bp windows with a 50 bp overlap and then perform differential analysis of windows. Although the analysis will be a little slow, but overlapping will provide more resolution and better detection of differentially enriched regions. The rationale behind using a 200 bp window is: the DNA which wraps around histone octamer is ~150 bp long (together makes a nucleosome) and two nucleosomes are connected by a ~50 bp long DNA.

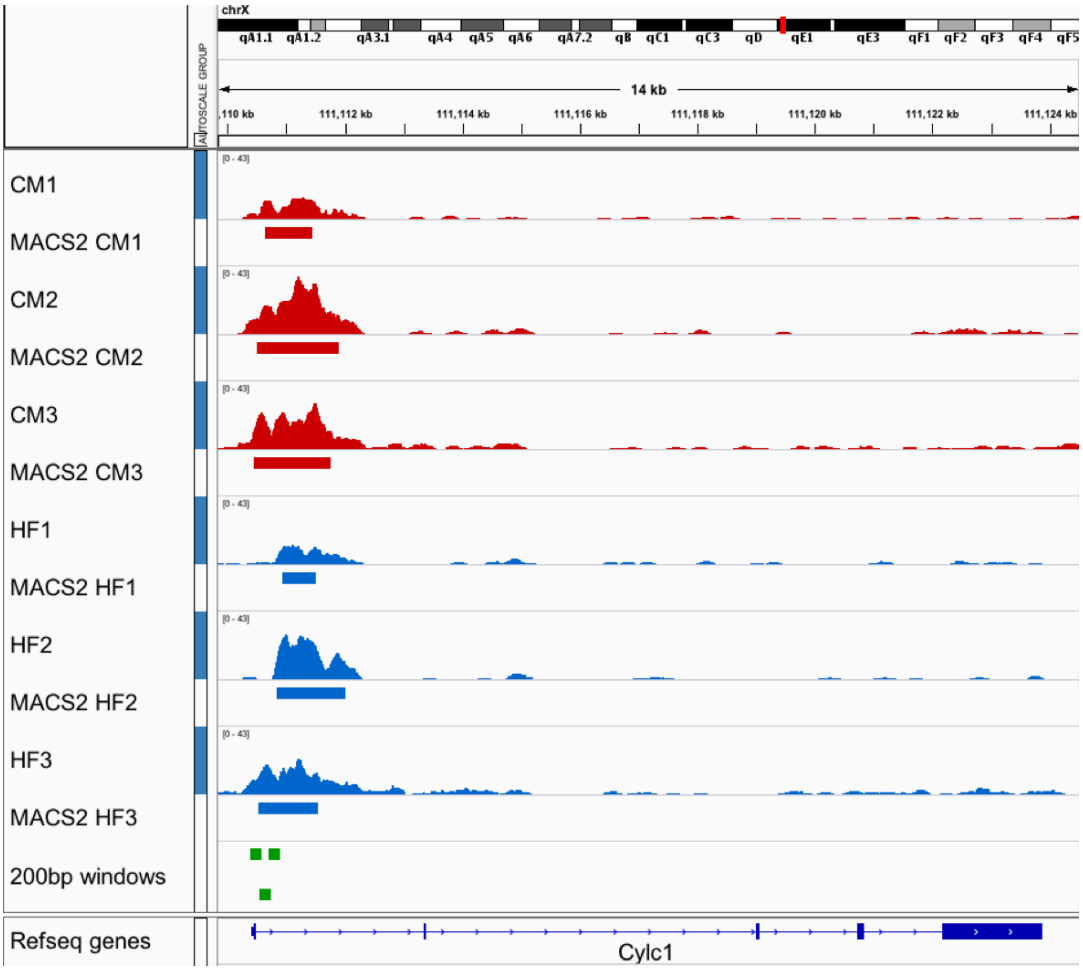

Figure 4.7: Coverage plot with results from differential analysis of high-fat vs control mice data using csaw. With the help of csaw, we were able to detect true differences between high-fat and control samples (even at a small scale). This figure shows the genomic track with peaks identified by MACS2, and at bottom (in green) windows identified as differentially enriched. The plot is for *Ctfr* (cystic fibrosis transmembrane conductance regulator) gene, which was differentially enriched in high-fat (log fold change: 1.133 and p-value: 0.0024).

4.2.4.1 Annotation of genomic regions

Annotation of the genomic regions is an essential step that involves annotating the regions to genes, ncRNAs or enhancers. After differential analysis and setting cutoff for selecting regions, we aim to annotate the DE windows. There are various packages available to perform the annotation, like ChIPseeker [87], Homer [88], Genomic Regions Enrichment of Annotations Tool (GREAT) [89]. Among these, ChIPseeker provides a better overview by incorporating superior graphical representation.

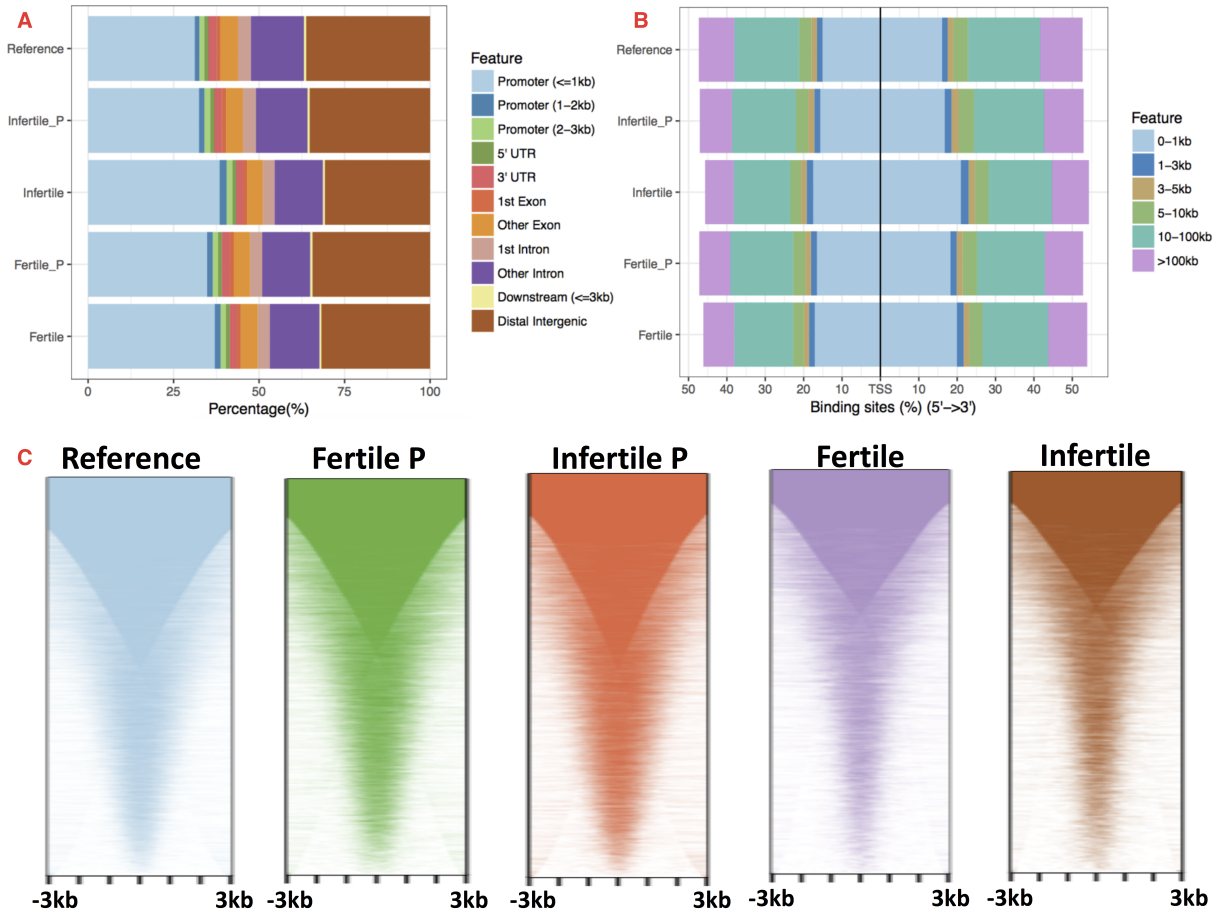

Figure 4.8: Annotation of peaks called by MACS2 on fertile, infertile and reference population samples. "P" represents "pooled" (see section 4.1). A. Annotation of peaks called by MACS2. Different colors represent the type of genomic sites enriched. **B.** Binding site percentage of peaks around TSS. Different colors represent the distance of binding sites for H3K4me3 from TSS. **C.** Tag density heatmap showing the intermediate signal of peaks from all samples +/- 3 kb TSS.

For annotation, ChIPseeker needs a bedgraph file as an input (or a list of bedgraph files). The genomic coordinates are then annotated based on transcript database (for example "TxDb.Hsapiens.UCSC.hg19.knownGene") and boundaries of TSS (for example +/- 3 kb), as specified by the user. In order to make multiple comparisons of regions among samples (for example: peaks called by MACS2), ChIPseeker provides clear visualization plots, as shown in Figure 4.8

4.2.4.2 Pathway Analysis

Genes that are DE should have biological significance, which can be tested against the pathway databases. The most widely used functional analysis method is the **Functional Class Scoring (FCS)** approach. FCS methods have several advantages; for example, they treat the pathway independently as a gene can function in more than one pathway, meaning that pathways can overlap [90]. The first method developed on FCS analysis technique was Gene Set Enrichment Analysis (GSEA) and is still widely used to perform pathway analysis [91, 92]. It requires an input gene list, in which genes are ranked by some values (correlation or Log Fold Change or read counts): walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not. The magnitude of the increment depends on the correlation of the gene with the phenotype (or absolute value of the ranking metric). The gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic. The pathway-level statistics used by current approaches include the Kolmogorov-Smirnov statistic [92]. This method uses Molecular Signature Database (MSigDB) [92] or pathway databases in Gene Matrix Transposed (GMT) file format.

I incorporated "fgsea" [93] for pathway analysis, which is based on GSEA. Similar to GSEA, "fgsea" also requires two inputs: a ranked file and a pathway database in GMT format. Gene Ontology (GO) analysis using "fgsea" between infertile-fertile samples comparison is shown in Figure 4.9.

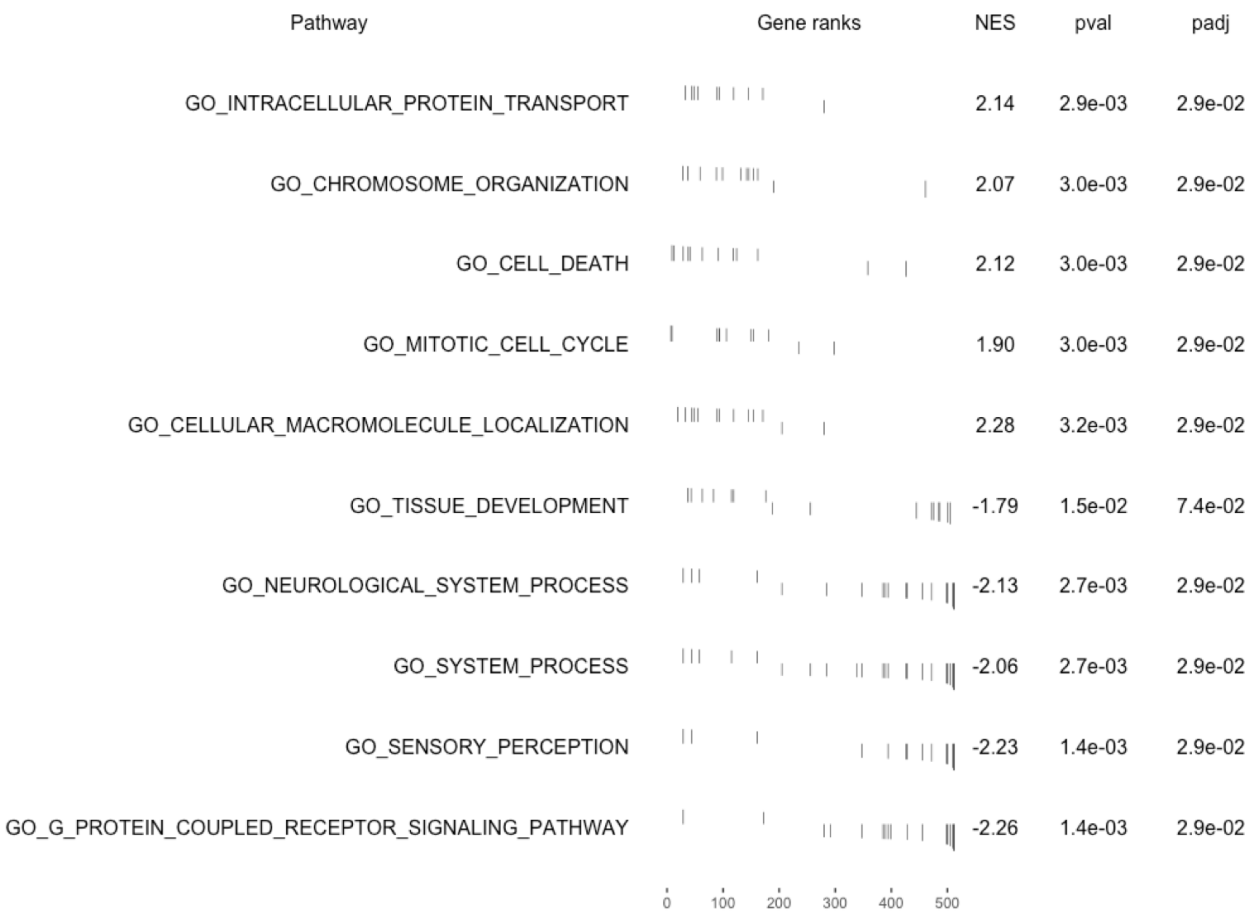

Figure 4.9: GO analysis results of human infertile vs fertile samples using fgsea. GO analysis results on differentially enriched genes from infertile-fertile comparison, using curated GO pathway database from MSigDB [92]. The colum "Gene ranks" represents the distribution of genes in the input ranked list (ranked by log-fold change), that were present in the GO term; "NES" is normalized enrichment score of a GO term; "pval" is p-value; and "padj" is the corrected p-value of the enriched GO term. Result shows that positively-enriched (fertile) genes in infertile samples are represented in cell-cycle and negatively-enriched (infertile) genes are represented in developmental GO terms.

5 Pipeline

5.1 EpiSpermHis

I have developed a Docker container (a container to encapsulate Galaxy, tools and pipelines) for analyzing sperm ChIP-Seq data from broad histone marks in sperm, such as, H3K4me3. There are two main components of the pipeline: Galaxy and Docker.

5.1.1 Galaxy

Analyzing the data from genomics experiments is not an easy task, and requires unique skill-sets or a well-trained bioinformatician. If biologists could analyze their own data with simple clicks on the computer, they would probably be the best to interpret their data. To help biologists, Galaxy framework was developed in 2005 to facilitate the development of GUI for bioinformatics pipelines [94]. It is a web-based framework and integrated tool management system for running command-line utilities from a GUI. It provides a point-and-click web interface alternative to the bioinformatics command line, thus allowing researchers to easily create, run and troubleshoot analytical pipelines. Furthermore, it allows non-computational researchers in the lab to begin learning NGS data analysis without first taking 4-6 months to learn Linux at the command-line. Data analysis in Galaxy can be run through reproducible workflows (pipelines). One can take a Galaxy workflow and run it on another computer (provided Galaxy and required tools are installed) without considering compatibility, which is not the case with a command-line pipeline (one would have to declare various variables on their "command-line user interface" before running a pipeline).

Galaxy maintains a detailed record of what analyses each user has run and in what order. The software also fosters reproducibility, making it possible to repeat and share analyses, and/or revisit them at a later date. Also, depending on the computing power, Galaxy can automatically process samples in parallel mode.

5.1.2 Docker

A pipeline can be run for data analysis if an interface to run it is present and all the essential software (dependencies) are installed. Docker [95], a tool built in 2014, can package an application and its dependencies (external software/ requirements) in a virtual container that can be run on any virtual

machine (VM) and eliminates the term “works on my machine”. Docker machines are containers, that can be used without any dependency on operating systems. Containers are a way to package software in a format that can be run isolated on a shared operating system (OS). They are widely regarded as more effective than VM because of how they allocate resources. Unlike VMs, containers do not bundle a full OS, only libraries and settings are required to make the software work [96].

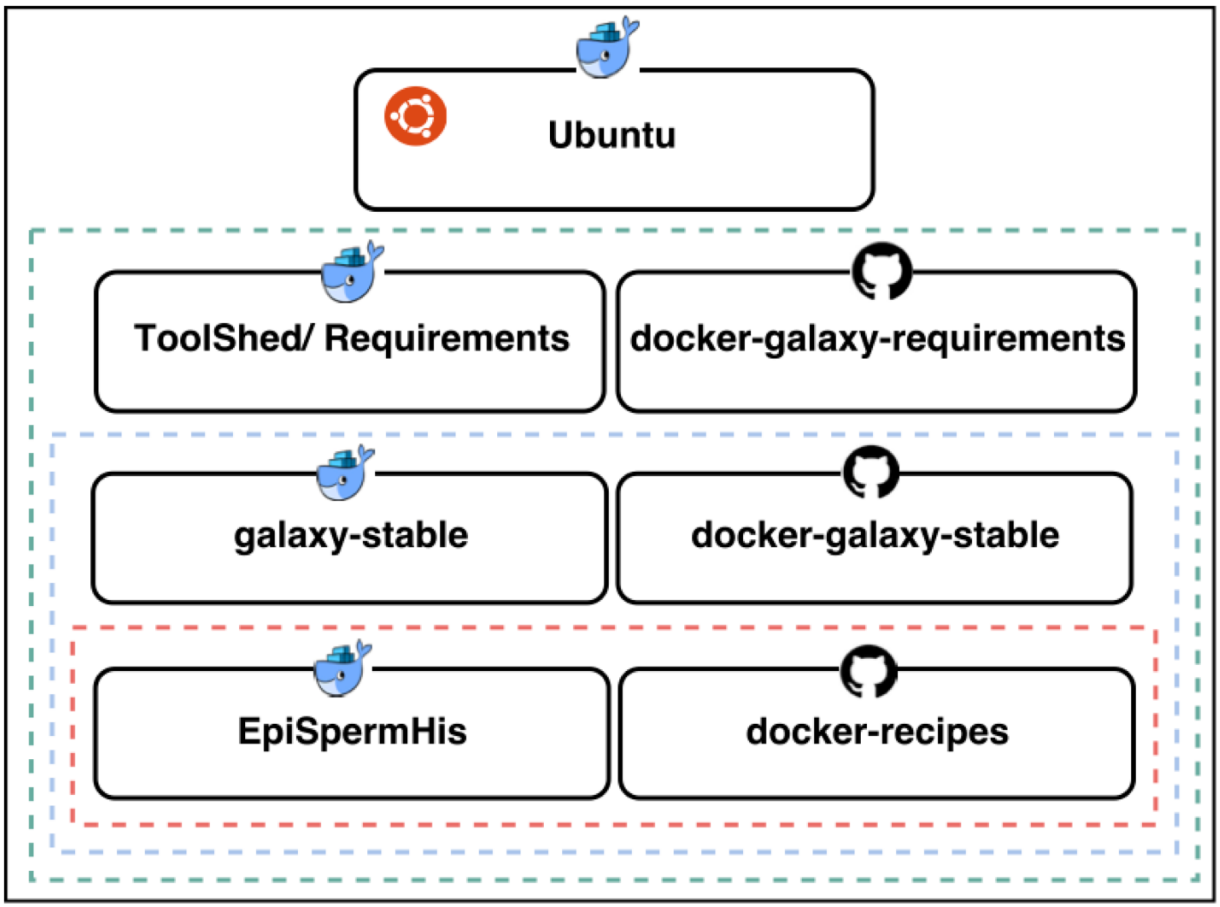

Figure 5.1: EpiSpermHis: Docker containerized Galaxy for sperm ChIP-Seq data analysis.

In 2015, Galaxy was containerized with Ubuntu (a Linux OS) and some base tools into a Docker image [97]. This Docker image has been used to further make it specific to a particular data analysis and the recently published RNA workbench [98] is a perfect example of that. I have used the original Galaxy Docker container and extended it to make it specific for sperm histone ChIP-Seq data analysis, Figure 5.1. The “EpiSpermHis” Docker container encapsulates several data analysis

pipelines, NGS tools and in-house built tools. The container also contains other NGS and statistical analysis tools, which will help researchers to adapt pipelines according to their research concern. This Docker container can be installed easily on any machine and data analysis can be carried out, depending on the computational power.

5.1.3 Graphical User Interface

Galaxy GUI is user friendly and contains various features. Some of the important features are shown in Figure 5.2.

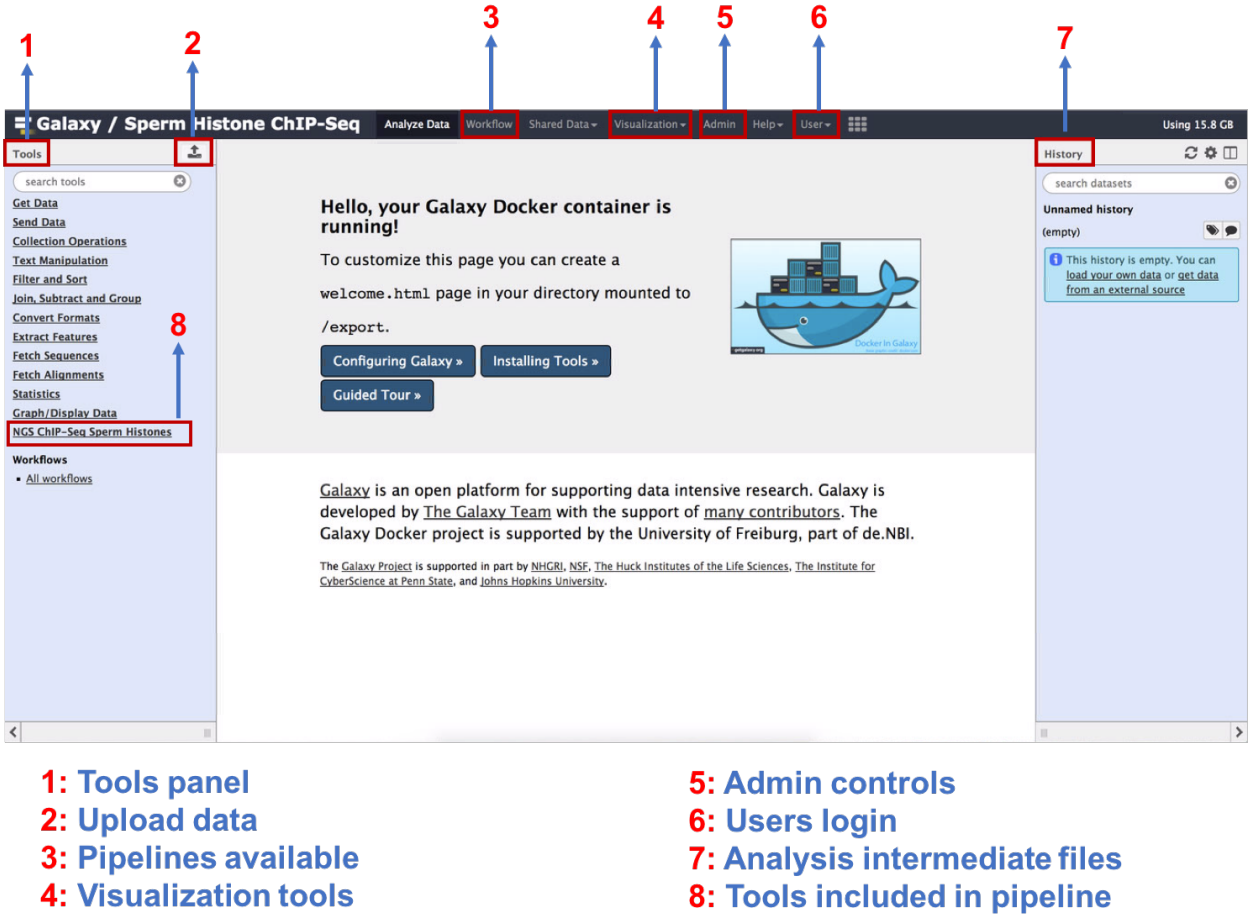

Figure 5.2: GUI of EpiSpermHis. Galaxy interface for sperm ChIP-Seq data analysis.

5.1.4 Galaxy tools

Galaxy tools are written in Extensible Markup Language (XML). A script or a piece of code can also be wrapped into an XML format. An XML file of the tool contains tool ID, tool name, tool version, help for the tool, commands to run the tool, input and output formats and tests. Before putting the tool into the Galaxy Test Toolshed, it needs to be tested and validated using Planemo (a software developed to assist tools development for galaxy) [109].

Various NGS tools are available in the Galaxy Test Toolshed for the data analysis. For "EpiSpermHis", we have included the tools for QC, peak calling and statistical analysis. An example of the in-house tool for removing reads with more than 3 SNPs is shown in Figure 5.3. As per Galaxy guidelines, this tool was wrapped into XML format, as required for publishing on Galaxy toolshed, Figure 5.4. After that, tools were tested with Planemo and they passed the test, Figure 5.5. All the in-house developed tools were tested in the same manner, before publishing them on the Test Toolshed website of Galaxy. Each tool is version controlled in the GitHub.

```
#!/bin/bash

#if [ "$1" == "-h" ]; then
if [[ -z $1 || -z $2 || $1 == "-h" ]]; then
    echo "This program will remove mismatches from sam file."
    echo "Usage: `basename $0` test.sam 3"
    exit 0
fi

cmd="echo perl -ne 'print if((/XM:i:[0-$2][^0-9]/) || (/^@/));' $1"
eval `$cmd`
```

Figure 5.3: RemoveSNPs: a tool coded in bash and Perl to filter reads with more SNPs.

```
<tool id="mismatchRemovalSam" name="Remove Mismatches (SNPs) from SAM" version="0.1.0">
  <requirements>
  </requirements>
  <command detect_errors="exit_code"><![CDATA[
        bash $__tool_directory__/mismatchRemovalSam "$input1" 3 > "$output1"
    ]]></command>
  <inputs>
    <param type="data" name="input1" format="sam" />
    <param type="integer" name="SNPs/ mismatches allowed" value="3" />
  </inputs>
  <outputs>
    <data name="output1" format="sam" />
  </outputs>
  <tests>
    <test>
      <param name="input1" value="input.sam"/>
      <output name="output1" file="output.sam"/>
    </test>
  </tests>
  <help><![CDATA[
This program will remove mismatches from sam file.
Usage: mismatchRemovalSam test.sam 3
    ]]></help>
  <citations>
    <citation type="bibtex">
      @misc{github01_mismatch_removal,
      author = {LastTODO, FirstTODO},
      year = {TODO},
      title = {01_mismatch_removal},
      publisher = {GitHub},
      journal = {GitHub repository},
      url = {https://github.com/dktanwar/Galaxy_Tools/tree/master/01_mismatch_removal},
      }</citation>
  </citations>
</tool>
```

Figure 5.4: XML format of RemoveSNPs tool for Galaxy. Tools for Galaxy are written in XML format. For a tool to run on the Galaxy, all the inputs and outputs for the tool are defined. This is an XML code for RemoveSNPs tool that was written for removing sequenced reads with higher SNPs.

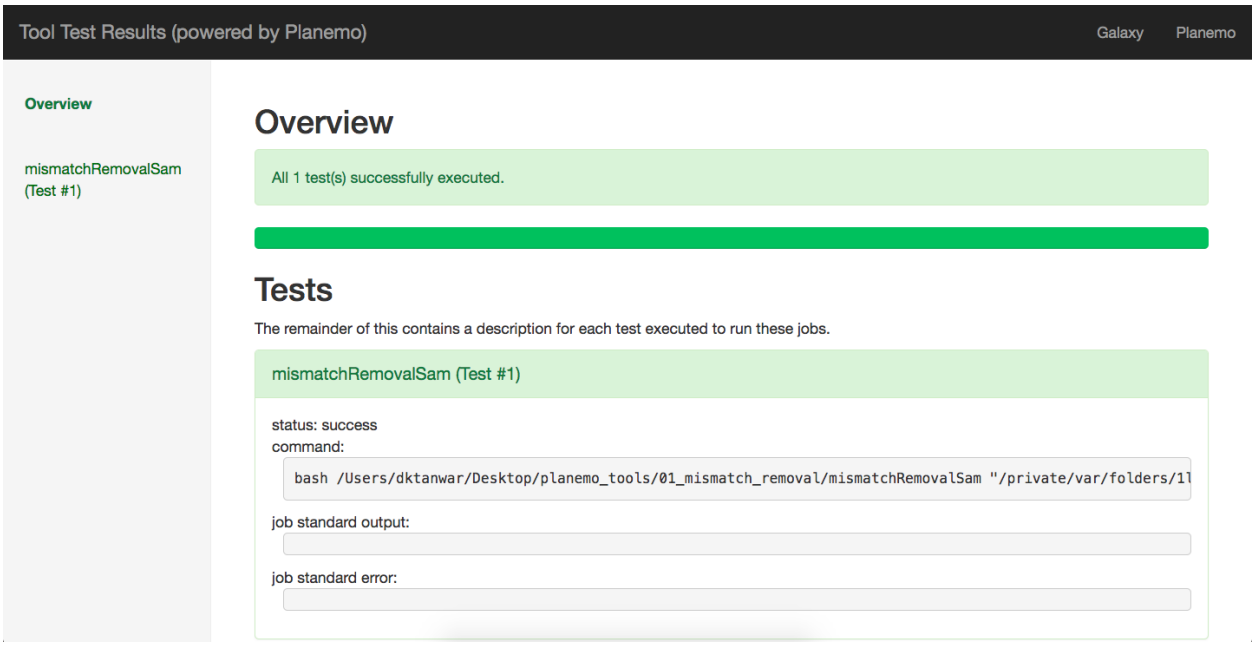

Figure 5.5: Test result of a Galaxy tool: RemoveSNPs tool passes a test.

5.1.5 Pipelines

An overview of a pipeline is shown in Figure 5.6. Pipelines available for sperm ChIP-Seq data analysis are mentioned in table 5.1. All the pipelines can be modified with the help of "workflow editor" in Galaxy to make them work for a specific reference genome (mm10/ hg19).

Table 5.1: Pipelines for sperm ChIP-Seq data analysis

Pipeline	Function
PreprocessingPeakCallingInputSE	Pre-processing and peak calling (including "gDNA/ Input") of SE ChIP-Seq data
PreprocessingPeakCallingInputPE	Pre-processing and peak calling (including "gDNA/ Input") of PE ChIP-Seq data
PreprocessingPeakCallingSE	Pre-processing and peak calling (without "gDNA/ Input") of SE ChIP-Seq data
PreprocessingPeakCallingPE	Pre-processing and peak calling (without "gDNA/ Input") of PE ChIP-Seq data
StatisticalAnalysis	Pipeline to perform statistical analysis (differential analysis, functional analysis and annotation)

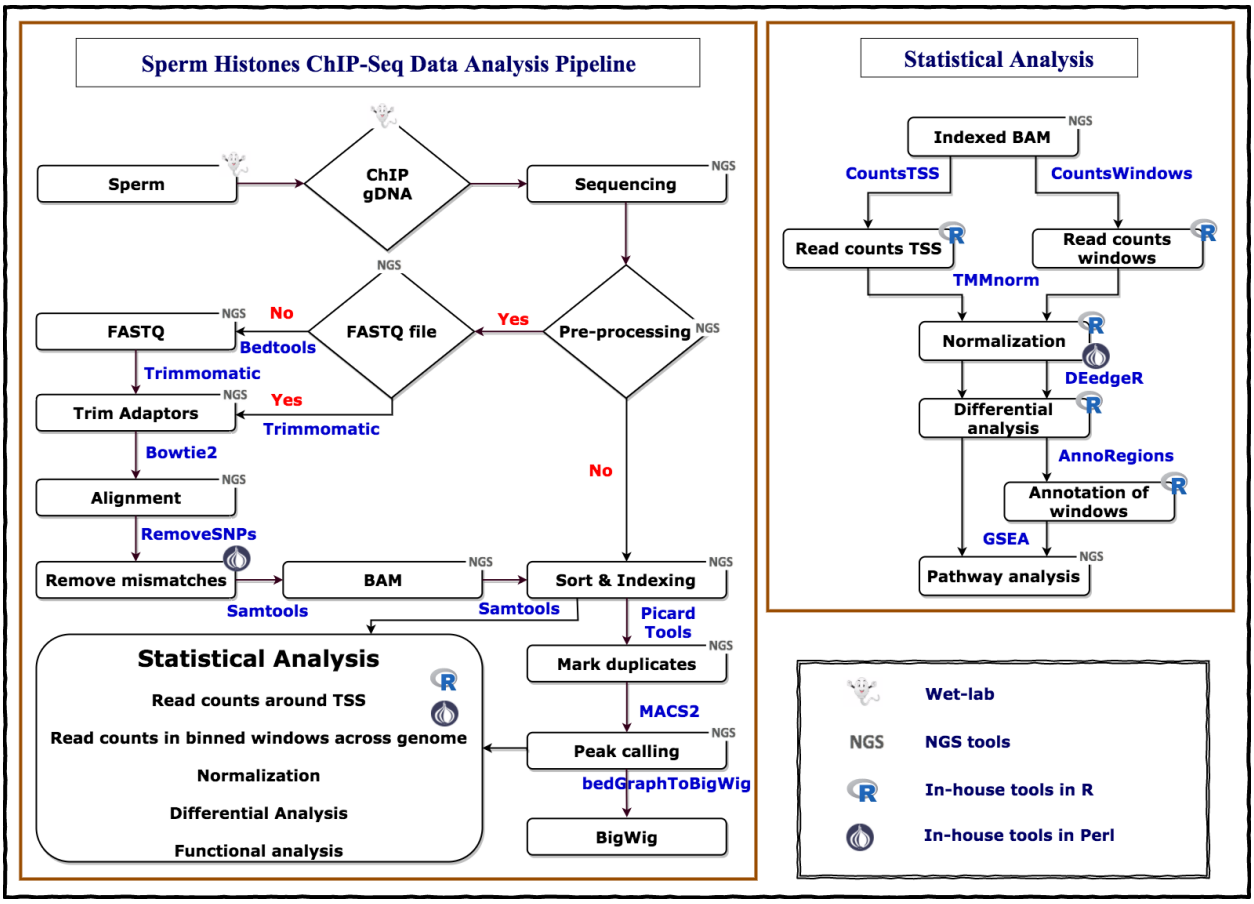

Figure 5.6: Pipeline for analyzing Epigenetic (histone) modification in sperm. Overview of Sperm ChIP-Seq data analysis pipeline with all the steps and tools used.

5.2 Pipeline evaluation

To evaluate the pipeline, I used the publicly available H3K4me3 targeted sperm ChIP-Seq datasets. Human and mice datasets were obtained from Gene Expression Omnibus (GEO). GSE15594 (human sperm ChIP-Seq) was obtained from [4] and GSE79230 (mouse sperm ChIP-Seq) was obtained from [99]. For GSE15594, authors had collected the data from three known fertile donors, who were attending the University of Utah Andrology laboratory, and gave their consent for research.

Data was provided as an input to the pipeline in `fastq` format and was aligned to the reference genomes ("mm10" for mice and "hg19" for human). The pipeline runs well for preprocessing and peak calling, Figure 5.7. As the data from GEO was available from only one biological group, it

was not able to evaluate the pipeline for any statistical analysis.

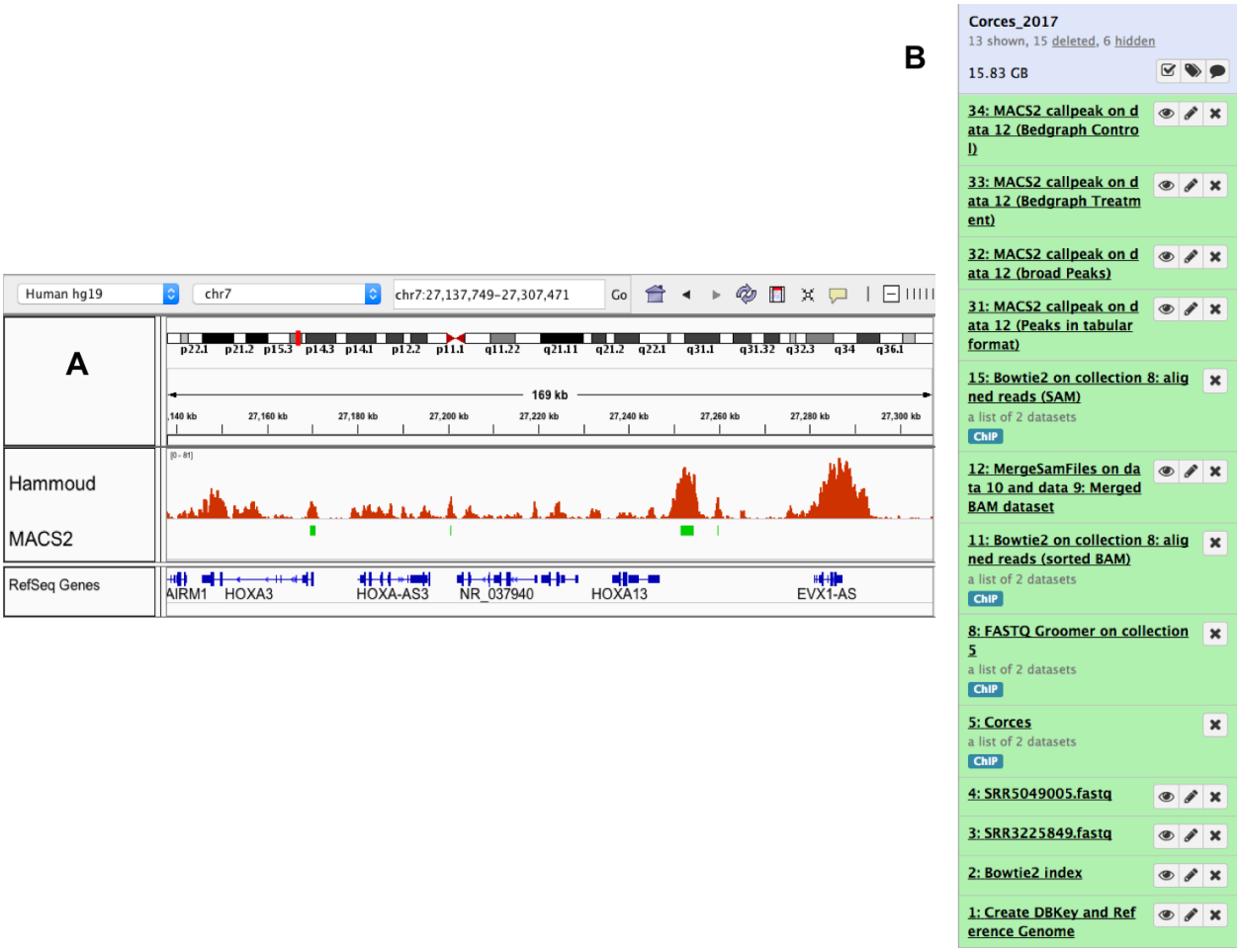

Figure 5.7: Coverage plot of human and mice sperm data. A. Coverage plot of *Hox* (Homeobox) genes from GSE15594 dataset. The figure also depicts peaks identified by MACS2. The peaks were not identified properly as the data is from year 2009 with sequencing depth of only 10 million. Another possible reason could be the compatibility issues of current version of MACS2 software with sperm ChIP-Seq data from year 2009. **B** Screenshot of outputs created while analyzing GSE79230 in Galaxy.

6 Conclusion

In the end, we could say that epigenetics refers to the reversible, heritable changes in gene regulation which occur without changing genetic code (DNA sequence), and sperm ChIP-Seq data analysis requires optimized methods to perform correctly. **From a developer's point of view**, the advantage of scripting method is the flexibility in design, debugging and execution. With scripting, it is possible to perform data analysis, use application programming interface (API) of databases to retrieve data/ information, and to incorporate them into the pipeline. Further, tools can be programmed to implement the missing elements. **From an end user perspective**, once the pipeline is implemented, performing the data analysis is possible by providing the input files and cluster scheduler, such as sun grid engine (SGE)/ slurm, for job submission. However, clusters often do not have a GUI but rather use a command line interface (CLI).

Docker container is lightweight and is the best solution for making a data analysis platform available by wrapping Galaxy, data analysis tools and pipelines. Galaxy has been proven to be very helpful for biologists for data analysis but is less flexible compared to command-line. The pipeline system (Docker container) developed here may be further extended or adapted according to the research questions and type of data available.

Appendix A: Manuscript in preparation

Manuscript in preparation for submission to Bioinformatics (Application note)

EpiSpermHis: A Docker container to study H3K4me3 modifications in sperm using Galaxy

Deepak Tanwar[1], Jianguo Xia[1,3], Sarah Kimmins[1,2]*

[1]Department of Animal Science, McGill University – Sainte-Anne-de-Bellevue, QC, Canada

[2]Department of Pharmacology and Therapeutics, McGill University – Montreal, QC, Canada

[3]Institute of Parasitology, McGill University – Sainte-Anne-de-Bellevue, QC, Canada

*Corresponding author: sarah.kimmins@mcgill.ca

Short title: Docker container for sperm ChIP-Seq data analysis.

Keywords: Docker, Galaxy, ChIP-Seq, Histone, Broad marks, Epigenomics

Grant support: This work was supported by the Canadian Institute of Health Research (CIHR) - to SK

Abstract

Unlike a typical ChIP-seq profile in somatic cells, in sperm, there are fewer histone peaks, and these tend to be distributed over CpG (5'-C- phosphate-G-3') enriched regions. We have developed a standalone Docker container with Galaxy, that has embedded tools for NGS and quantitative comparison of histone modification enrichments in sperm and includes downstream data analysis pipelines. The Galaxy framework will allow biologists without command-line knowledge to perform optimal pre-processing and to address other challenges in data analysis. We have incorporated efficient bioinformatics pipelines for analyzing sperm epigenome data by using sperm specific parameters in currently available tools and tools developed in-house, including tools for differential analysis and functional analysis. We have included several tools, statistical analyses and R packages, allowing users to adapt the analysis to their research requirements. This Docker image is available from our DockerHub repository (https://hub.docker.com/r/dktanwar/ngs_chip-seq_sperm_histones/).

Keywords

ChIP-Seq, Histones, H3K4me3, Bioinformatics, Docker, Galaxy

Findings

Background

The recent support for environmental factors causing epigenetics alterations has generated data suggesting how fertility can be affected by environmental exposures occurring in histones [101–104]. ChIP-Sequencing enables us to understand and study these alterations in sperm histones by mapping them throughout genome.

The challenges in analyzing and quantifying ChIP-seq data from sperm with currently available software is the need to detect and quantify differences not just in peak enrichment, but also to call peaks with parameters specific to sperm. It is even more challenging for biologists to study these modifications by virtue of computers (bioinformatics data analyses). To analyze these datasets, a person needs to have command line knowledge as well as knowledge of other programming languages, including R for computational statistical analysis. Also, the installation of tools and then

fighting dependencies (external software on which installed software depends on) of software is always problematic.

There are several tools available that can assist in making automated pipelines, including makefiles and snakemake [100]. The graphical user interface (GUI) is limited in the field of Bioinformatics. A GUI has a unique feature of automating pipeline development and reproducibility for genomic data analysis. The GUI based platforms available to date for bioinformatics data analysis are Galaxy [105], Taverna [106], Pegasus [107], Geneious (commercial) [108]. The Galaxy community is dedicated to increase the availability of bioinformatics tools for researchers. Further, Galaxy offers a variety of features and is "an open source, web-based platform for performing accessible, reproducible, and transparent genomic science" [105]. Galaxy is one of the best GUI for bioinformatics, that provides web-based platform for large-scale interactive data analysis. It allows scientists to share the whole analysis workflow and datasets. Further, Galaxy is a user-friendly analytical platform, and researchers without computational background can navigate their way through an investigation and use various analytical tools and workflows. It provides various features such as interactive tours, workflow building from history, sharing history with data and workflows, being open source.

We have developed a comprehensive Docker container with NGS tools, statistical analysis (written in R) and workflows that can be used for the processing of raw data, peak calling and further downstream analysis (statistical and functional analysis). These workflows can be customized depending on factors like sequence read length. Software like Galaxy, provides a GUI for biologists to allow them to analyze their data. This toolbox was generated for the analysis of histone modifications, such as H3K4me3, that can have a very broad coverage in sperm cells, as compared to somatic cells [4] .

Implementation

The developed toolbox makes use of several existing programs and connects them in a convenient interface. Further, it uses recommended practices and can assemble extremely high coverage of sperm epigenome data. In addition, we provide a series of statistical analysis tools which will help researchers to quantify the differences in biological groups.

Galaxy enables the addition of novel software into the environment by individual users to fill in the gaps of tools that haven't been created by the Galaxy team or other researchers. Galaxy

tool integration requires a tool definition file, which is written in Extensible Markup Language (XML). This file dictates the available inputs, outputs and arguments to build the command based on user-specified parameters. We obtained assistance from Planemo tool [109] to construct the XML format configuration file for the tool. In-house programs were developed in different computer languages (Bash, Perl, Python and R) and are configured to work on Galaxy. All repositories are available on the public Galaxy Test Toolshed (https://testtoolshed.g2.bx.psu.edu/) which allows users to automatically install any tool in any Galaxy environment [110]. Full access to the test tool shed repository is provided to Intergalactic Utilities Commission (IUC).

Methods

To run the pipeline, data is accepted in its raw state (FastQ) as well as in binary alignment format (BAM), since these are two formats in which data is made available to researchers after sequencing. Most of the packages and softwares in bioinformatics contain several different functions and cannot be used simultaneously for data analysis and representation. To perform some specific analysis, the code is modified and organized according to the requirement. We have molded the code into various scripts to perform specific functions. All tools are made available to Toolshed and are version controlled on GitHub [111].

Pipeline description

The "EpiSpermHis" Docker container is developed for the analysis of ChIP-Seq data from histone marks targeted via ChIP in sperm and contains a series of analytical methods that are connected. As shown in the }, the processing of sperm histones ChIP-Seq data is required to go through three major steps:

- Quality Control
- Peak Calling
- Quantitative/ Semi-quantitative analysis.

The rationale and details of each step is as follows:

1. **Quality Control of ChIP-Sequencing data:** A sequencing center makes data available either in raw format (fastq) or as data aligned to the preferred reference genome. However, researchers do not necessarily use the same reference genome. Moreover, adapters are also attached to the sequenced reads, which affect the quality of reads. Further, some reads (of shorter length) could have poor base quality score and should be removed from further analysis. If a sequence fails to call a base, it will put a "N", instead of "A", "T", "C" or "G" and these sequences should be removed to improve the quality of the data. All the pre-processing steps are performed using Trimmomatic software [59].

Proper alignment of sperm ChIP-Seq data is required to get an accurate alignment and maximum depth. There are various software discussed for the alignment of sperm ChIP-Seq data [15]. We used bowtie2 for alignment due to its advantage over bowtie [62], using the same parameters as described in [15], except for allowing for a maximum of 100 reads aligned to a region. When reads align more than 100 times, they can be classified as technical duplicates. It is recommended to remove while analyzing ChIP-Seq data [66]; hence, they are filtered out further in the pipeline using Picard tools [112] (keeping only one copy of a read) for peak calling. As specified in [15], the pipeline allows a maximum of three mismatches (SNPs) in a sequenced read.

2. **Peak calling:** Peak calling is an essential step in ChIP-Seq data analysis which identifies the enriched regions in the genome. Three different methods for peak calling were tested, Figure A.2} out of which, MACS2 with "–broad" parameter, along with "–broad-cutoff" of 10^{-6} predicts peaks with a well-defined boundary.

3. **Quantitative/ Semi-quantitative analysis:** For a biologist, statistical analysis is a fundamental step to determine the important changes in the epigenome. We have tailored our scripts to Galaxy tools for various statistical tests including edgeR and DSeq2, and "fgsea" functional enrichment analysis. Tools included are mentioned in table A.1. All tools were developed keeping the standard format for Galaxy, and then were uploaded to test tool shed only after they passes a test.

Conclusion

In this paper we have described a standalone Docker container for sperm histones ChIP-Seq data analysis. To construct the fundamental elements of the pipeline we relied on the tools that are

Table A.1: In-house tools for analysis

Tools	Language	Function
RemoveSNPs	Perl	Remove DNA sequenced reads with more than 'n' SNPs
CountsTSS	R	Calculate the read counts around TSS
CountsWindows	R	Calculate the read counts in whole genome after binning into windows
TMMnorm	R	TMM normalization of read counts
DEedgeR	R	Differential Enrichment analysis using edgeR
AnnoRegions	R	Annotation of windows

already developed in the bioinformatics community. We did not want to reinvent the wheel in each and every step but added wrappers of existing tools to perform specifically. The pipelines make use of highly recommended parameters for data analysis softwares, which are specific to sperm. All our tools are wrapped into a Galaxy instance, which can also be used independently. The Galaxy platform and pipelines can be used by researchers without advanced programming skills, or without high-performance computing clusters.

Taken together, our proposed pipeline includes all relevant sperm histone ChIP-Seq sequencing data processing modules, is easily applicable, and needs no time consuming installation processes. Based on what we experienced through this project, the most challenging task is to find the best program that firstly supports the data of your interest and secondly uses reasonable computational resources and time to perform and finally produces biologically meaningful result. We plan to integrate more tools as well as downstream analysis tools, that can assist with functional annotation.

Availability and requirements

Project Name: EpiSpermHis

Project home page: https://github.com/dktanwar/NGS_ChIP-Seq_Sperm_Histones

Operating system(s): UNIX (Galaxy); Platform independent.

Programming Language: Bash, Perl, Python, R

License: MIT

Any restrictions to use by non-academics: None

All tools described in the article are available in the Galaxy Test Toolshed (https://toolshed.g2.bx.psu.edu/) and under MIT license via the project GitHub repository. The Dockerfile to automatically install these tools and a prebuilt Docker image are also provided.

Declarations

List of abbreviations

ChIP-Seq:	Chromatin Immunoprecipitation Sequencing
CpG:	5'-C-phosphate-G-3'
H3K4me3:	Histone 3 Lysine 4 tri-methylation
GUI:	Graphical User Interface
KEGG:	Kyoto Encyclopedia of Genes and Genomes
GO:	Gene Ontology
SNPs:	Single Nucleotide Polymorphisms

Funding

SK is funded by CIHR grant.

Competing interest

The authors declare that they have no competing interests.

Authors' contributions

DT designed the toolbox, analyzed data and wrote manuscript. SK designed the experiment and along with JX supervised the primary author.

Acknowledgements

We thank Keith Siklenka and Romain Lambrot from SK research group for helping in tools selection and providing feedback on this work. We also thank Galaxy community for their ongoing support. We are particularly grateful to Björn Gruening, Mallory Freiberg, Yvan Le Bras and Devon Ryan for their courteous assistance.

Peaks in somatic cells, mouse sperm cells and human sperm cells from H3K4me3 along with density plot and boxplot of peaks distribution

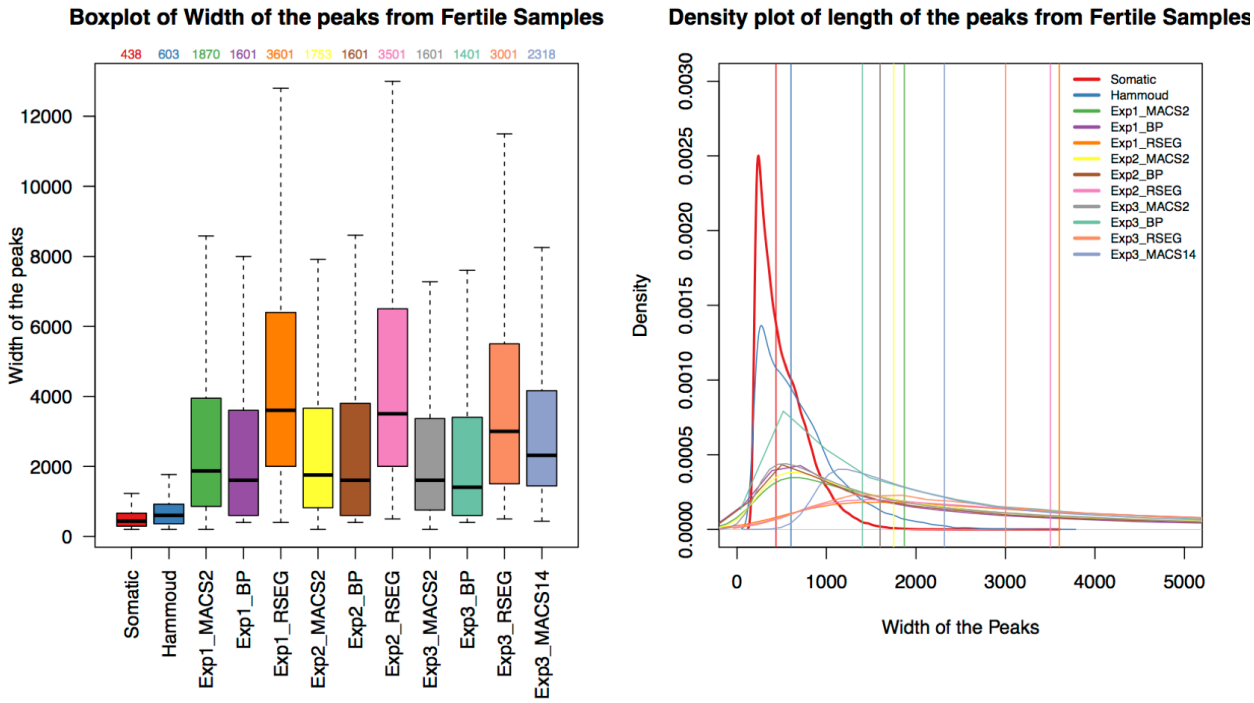

Figure A.1: Peaks in somatic cells, mouse sperm cells and human sperm cells from H3K4me3 along with density plot and boxplot of peaks distribution.

Pipeline overview (along with the programming language used) and also a pipeline of statistical analysis.

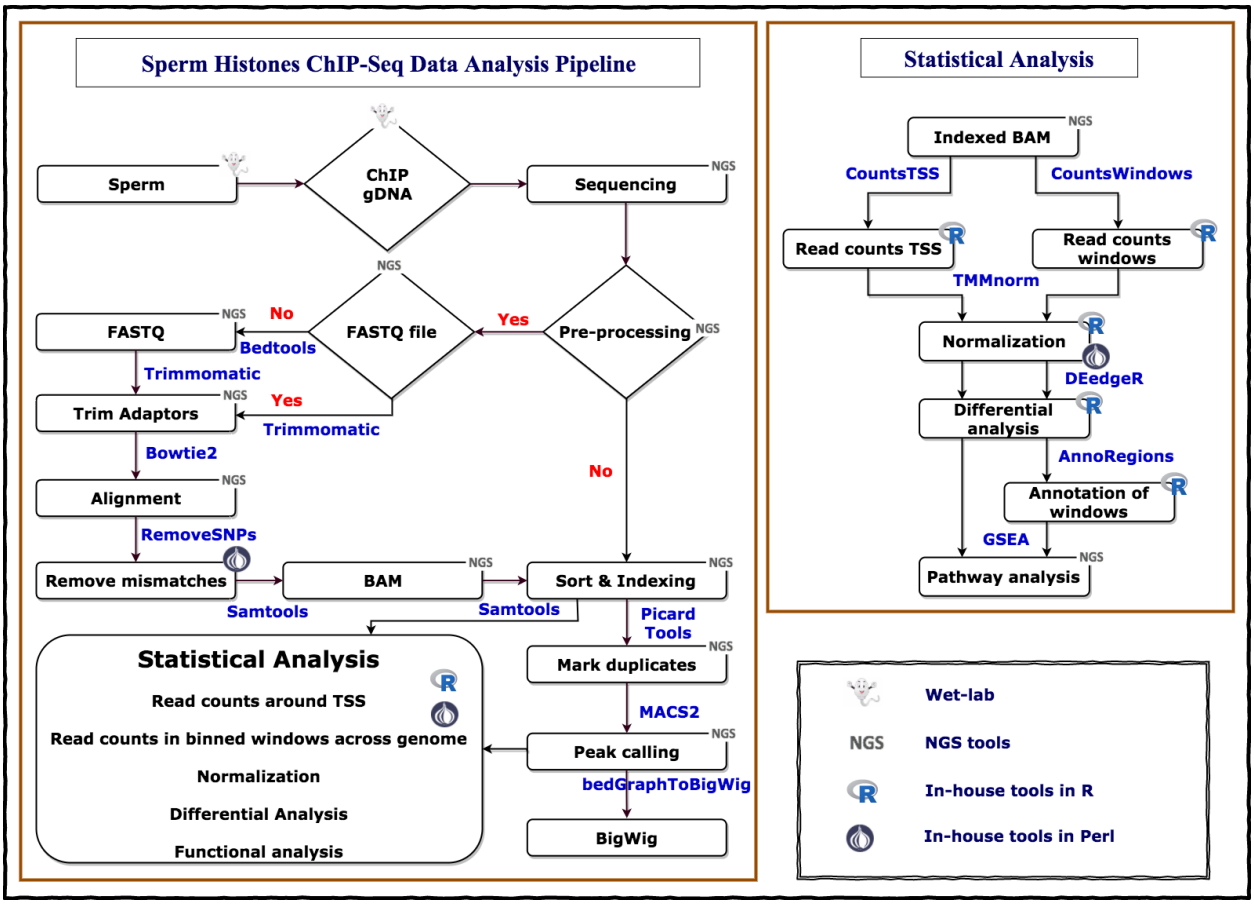

Figure A.2: Pipeline for analyzing Epigenetic(histone) modification in sperm. Overview of Sperm ChIP-Seq data analysis pipeline with all the steps and tools used.

Appendix B: Samples statistics

Following tables show basic statistics of the data mentioned in section 4.1 :

Table B.1: Basic statistics table for human samples

Sample	Sequencing	Sequenced read length	Sequenced read counts	Duplicates	Uniquely mapped reads
Reference (n=30)	SE	100 bp	143,077,301	37,625,112 (26.3%)	105,452,189 (73.7%)
Fertile pooled (n=4)	SE	100 bp	45,952,728	7,187,648 (15.65%)	38,765,080 (84.35%)
Fertile	SE	100bp	38,226,540	40,55,797 (10.6%)	34,170,743 (89.4%)
Infertile pooled (n=4)	SE	100 bp	41,002,736	7,862,122 (19.2%)	33,140,614 (80.8%)
Infertile	SE	100 bp	51,192,266	6,217,476 (12.15%)	44,974,790 (87.85%)

Table B.2: Basic statistics table for mice samples

Sample	Sequencing	Sequenced read length	Sequenced read counts	Duplicates	Uniquely mapped reads
Control replicate 1	SE	100 bp	40,345,634	16,973,812 (42.1%)	23,371,822 (57.9%)
Control replicate 2	SE	100 bp	45,399,661	23,423,502 (51.6%)	21,976,159 (48.4%)
Control replicate 3	SE	100 bp	37,622,335	15,283,698 (40.6%)	22,338,637 (59.4%)
High-fat replicate 1	SE	100 bp	40,701,175	21,179,671 (52%)	19,521,504 (48%)
High-fat replicate 2	SE	100 bp	37,976,375	17,865,227 (47%)	20,111,148 (53%)
High-fat replicate 3	SE	100 bp	42,775,479	15,062,530 (35.2%)	27,712,949 (64.8%)

Appendix C: Data availability

All the computational analysis was carried out using lab facilities and data is available in lab computer in the following locations:

1. **Reference, fertile and infertile samples from humans:**
   `/Volumes/BINF1_Raid/home/dtanwar/projects/Human_Sperm_CReATe_H3K4me3`
2. **High-fat and control samples from mice:**
   `/Volumes/BINF1_Raid/home/dtanwar/projects/H3K4me3_McMasterU`
3. **Data from GEO:**
   `/Volumes/BINF1_Raid/home/dtanwar/projects/testing`

References

[1] D. Koryakov, "Histone modification and regulation of chromatin function", *Russian Journal of Genetics*, vol. 42, no. 9, pp. 970-984, 2006.

[2] N. Vastenhouw and A. Schier, "Bivalent histone modifications in early embryogenesis", *Current Opinion in Cell Biology*, vol. 24, no. 3, pp. 374-386, 2012.

[3] E. Heard and R. Martienssen, "Transgenerational Epigenetic Inheritance: Myths and Mechanisms", *Cell*, vol. 157, no. 1, pp. 95-109, 2014.

[4] S. Hammoud, D. Nix, H. Zhang, J. Purwar, D. Carrell and B. Cairns, "Distinctive chromatin in human sperm packages genes for embryo development", *Nature*, vol. 460, no. 7254, pp. 473-478, 2009.

[5] S. Wu, H. Zhang and B. Cairns, "Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm", *Genome Research*, vol. 21, no. 4, pp. 578-589, 2011.

[6] K. Siklenka, S. Erkek, M. Godmann, R. Lambrot, S. McGraw, C. Lafleur, T. Cohen, J. Xia, M. Suderman, M. Hallett, J. Trasler, A. Peters and S. Kimmins, "Disruption of histone methylation in developing sperm impairs offspring health transgenerationally", *Science*, vol. 350, no. 6261, pp. aab2006-aab2006, 2015.

[7] U. Brykczynska, M. Hisano, S. Erkek, L. Ramos, E. Oakeley, T. Roloff, C. Beisel, D. Schübeler, M. Stadler and A. Peters, "Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa", *Nature Structural & Molecular Biology*, vol. 17, no. 6, pp. 679-687, 2010.

[8] L. Lindeman, I. Andersen, A. Reiner, N. Li, H. Aanes, O. Østrup, C. Winata, S. Mathavan, F. Müller, P. Aleström and P. Collas, "Prepatterning of Developmental Gene Expression by Modified Histones before Zygotic Genome Activation", *Developmental Cell*, vol. 21, no. 6, pp. 993-1004, 2011.

[9] "Canadian Health, Disease, & Medication Information - Canada.com", *Bodyandhealth.canada.com*, 2017. [Online]. Available: http://bodyandhealth.canada.com. [Accessed: 20- Sep- 2017].

[10] S. Esteves, R. Miyaoka and A. Agarwal, "An update on the clinical assessment of the infertile male", *Clinics*, vol. 66, no. 4, pp. 691-700, 2011.

[11] T. Vavouri and B. Lehner, "Chromatin Organization in Sperm May Be the Major Functional Consequence of Base Composition Variation in the Human Genome", *PLoS Genetics*, vol. 7, no. 4, p. e1002036, 2011.

[12] W. Ma, W. Wong, “Chapter Three - The Analysis of ChIP-Seq Data”, Editor(s): Chris Voigt, *Methods in Enzymology*, vol. 497, pp. 51-73, ISSN 0076-6879, ISBN 9780123850751, 2011.

[13] S. Mei, Q. Qin, Q. Wu, H. Sun, R. Zheng, C. Zang, M. Zhu, J. Wu, X. Shi, L. Taing, T. Liu, M. Brown, C. Meyer and X. Liu, "Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse", *Nucleic Acids Research*, vol. 45, no. 1, pp. D658-D662, 2016.

[14] S. Park, J. Kim, B. Yoon and S. Kim, "A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages", *Genomics & Informatics*, vol. 15, no. 1, p. 11, 2017.

[15] H. Royo, M. Stadler and A. Peters, "Alternative Computational Analysis Shows No Evidence for Nucleosome Enrichment at Repetitive Sequences in Mammalian Spermatozoa", *Developmental Cell*, vol. 37, no. 1, pp. 98-104, 2016.

[16] C. Waddington, "The Epigenotype", International *Journal of Epidemiology*, vol. 41, no. 1, pp. 10-13, 2011.

[17] A. Bird, "DNA methylation patterns and epigenetic memory", *Genes & Development*, vol. 16, no. 1, pp. 6-21, 2002.

[18] A. Kimura, K. Matsubara and M. Horikoshi, "A Decade of Histone Acetylation: Marking Eukaryotic Chromosomes with Specific Codes", *The Journal of Biochemistry*, vol. 138, no. 6, pp. 647-662, 2005.

[19] A. Shilatifard, "Chromatin Modifications by Methylation and Ubiquitination: Implications in the Regulation of Gene Expression", *Annual Review of Biochemistry*, vol. 75, no. 1, pp. 243-269, 2006.

[20] S. Henikoff and K. Ahmad, "Assembly of Variant Histones into Chromatin", *Annual Review of Cell and Developmental Biology*, vol. 21, no. 1, pp. 133-153, 2005.

[21] L. Ringrose and R. Paro, "Epigenetic Regulation of Cellular Memory by the Polycomb and Trithorax Group Proteins", *Annual Review of Genetics*, vol. 38, no. 1, pp. 413-443, 2004.

[22] R. Jones and K. Lopez, textitHuman reproductive biology, 4th ed. Amsterdam: Elsevier, 2014.

[23] B. Alberta, D. Bray, J. Lewis, M. Raff, K. Roberta and P. Walter, *Molecular biology of the cell*, 4th ed. New York: Garland Science, 2002.

[24] T. Wu and D. Chu, "Sperm Chromatin: Fig. 1.", *Molecular & Cellular Proteomics*, vol. 7, no. 10, pp. 1876-1886, 2008.

[25] B. Klein, *Cunningham's Textbook of veterinary physiology*, 5th ed. St. Louis: Elsevier Saunders, 2013.

[26] W. Steven Ward and D. Coffey, "DNA Packaging and Organization in Mammalian Spermatozoa: Comparison with Somatic Cell", *Biology of Reproduction*, vol. 44, no. 4, pp. 569-574, 1991.

[27] G. van der Heijden, A. Derijck, L. Ramos, M. Giele, J. van der Vlag and P. de Boer, "Transmission of modified nucleosomes from the mouse male germline to the zygote and subsequent remodeling of paternal chromatin", *Developmental Biology*, vol. 298, no. 2, pp. 458-469, 2006.

[28] J. Gatewood, G. Cook, R. Balhorn, C. Schmid and E. Bradburyeli, "Isolation of four core histones from human sperm chromatin representing a minor subset of somatic histones", *The Journal of Biological Chemistry*, vol. 265, no. 33, pp. 20662-20666, 1990.

[29] K. Maeshima, R. Imai, S. Tamura and T. Nozaki, "Chromatin as dynamic 10-nm fibers", *Chromosoma*, vol. 123, no. 3, pp. 225-237, 2014.

[30] G. Felsenfeld and M. Groudine, "Controlling the double helix", *Nature*, vol. 421, no. 6921, pp. 448-453, 2003.

[31] D. Tremethick, "Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber", *Cell*, vol. 128, no. 4, pp. 651-654, 2007.

[32] D. Carrell, "Epigenetics of the male gamete", *Fertility and Sterility*, vol. 97, no. 2, pp. 267-274, 2012.

[33] R. Martins and S. Krawetz, "Nuclear organization of the protamine locus" *Soc Reprod Fertil Suppl*, vol. 64, pp.1-12, 2007.

[34] R. Braun, "Packaging paternal chromosomes with protamine", *Nature Genetics*, vol. 28, no. 1, pp. 10-12, 2001.

[35] B. Saab and I. Mansuy, "Neurobiological disease etiology and inheritance: an epigenetic perspective", *Journal of Experimental Biology*, vol. 217, no. 1, pp. 94-101, 2013.

[36] L. Yebra, J. Ballesca, J. Vanrell, L. Bassas and R. Oliva, "Complete selective absence of protamine P2 in humans", *The Journal of Biological Chemistry*, vol. 268, no. 14, pp. 10553-10557, 1993.

[37] D. Lu, "Epigenetic modification enzymes: catalytic mechanisms and inhibitors", *Acta Pharmaceutica Sinica B*, vol. 3, no. 3, pp. 141-149, 2013.

[38] W. Ward, "Function of sperm chromatin structural elements in fertilization and development", *Molecular Human Reproduction*, vol. 16, no. 1, pp. 30-36, 2009.

[39] T. Barth and A. Imhof, "Fast signals and slow marks: the dynamics of histone modifications", *Trends in Biochemical Sciences*, vol. 35, no. 11, pp. 618-626, 2010.

[40] W. Iwasaki, Y. Miya, N. Horikoshi, A. Osakabe, H. Taguchi, H. Tachiwana, T. Shibata, W. Kagawa and H. Kurumizaka, "Contribution of histone N-terminal tails to the structure and stability of nucleosomes", *FEBS Open Bio*, vol. 3, no. 1, pp. 363-369, 2013.

[41] T. Kouzarides, "Chromatin Modifications and Their Function", *Cell*, vol. 128, no. 4, pp. 693-705, 2007.

[42] C. Vakoc, M. Sachdeva, H. Wang and G. Blobel, "Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin", *Molecular and Cellular Biology*, vol. 26, no. 24, pp. 9185-9195, 2006.

[43] J. Sims, S. Houston, T. Magazinnik and J. Rice, "A Trans-tail Histone Code Defined by Monomethylated H4 Lys-20 and H3 Lys-9 Demarcates Distinct Regions of Silent Chromatin", *Journal of Biological Chemistry*, vol. 281, no. 18, pp. 12760-12766, 2006.

[44] I. Hatada, M. Fukasawa, M. Kimura, S. Morita, K. Yamada, T. Yoshikawa, S. Yamanaka, C. Endo, A. Sakurada, M. Sato, T. Kondo, A. Horii, T. Ushijima and H. Sasaki, "Genome-wide profiling of promoter methylation in human", *Oncogene*, vol. 25, no. 21, pp. 3059-3064, 2006.

[45] U. Schagdarsurengin, A. Paradowska and K. Steger, "Analysing the sperm epigenome: roles in early embryogenesis and assisted reproduction", *Nature Reviews Urology*, vol. 9, no. 11, pp. 609-619, 2012.

[46] A. Barski, S. Cuddapah, K. Cui, T. Roh, D. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao, "High-Resolution Profiling of Histone Methylations in the Human Genome", *Cell*, vol. 129, no. 4, pp. 823-837, 2007.

[47] D. Johnson, A. Mortazavi, R. Myers and B. Wold, "Genome-Wide Mapping of in Vivo Protein-DNA Interactions", *Science*, vol. 316, no. 5830, pp. 1497-1502, 2007.

[48] Y. Chen, N. Negre, Q. Li, J. Mieczkowska, M. Slattery, T. Liu, Y. Zhang, T. Kim, H. He, J. Zieba, Y. Ruan, P. Bickel, R. Myers, B. Wold, K. White, J. Lieb and X. Liu, "Systematic evaluation of factors influencing ChIP-seq fidelity", *Nature Methods*, vol. 9, no. 6, pp. 609-614, 2012.

[49] S. Landt, G. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. Bernstein, P. Bickel, J. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. Hartemink, M. Hoffman, V. Iyer, Y. Jung, S. Karmakar, M. Kellis, P. Kharchenko, Q. Li, T. Liu, X. Liu, L. Ma, A. Milosavljevic, R. Myers, P. Park, M. Pazin, M. Perry, D. Raha, T. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. Stamatoyannopoulos, M. Tolstorukov, K. White, S. Xi, P. Farnham, J. Lieb, B. Wold and M. Snyder, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia", *Genome Research*, vol. 22, no. 9, pp. 1813-1831, 2012.

[50] B. Wold and R. Myers, "Sequence census methods for functional genomics", *Nature Methods*, vol. 5, no. 1, pp. 19-21, 2007.

[51] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law, "Comparison of Next-Generation Sequencing Systems", *Journal of Biomedicine and Biotechnology*, vol. 2012, pp. 1-11, 2012.

[52] C. Fuller, L. Middendorf, S. Benner, G. Church, T. Harris, X. Huang, S. Jovanovich, J. Nelson, J. Schloss, D. Schwartz and D. Vezenov, "The challenges of sequencing by synthesis", *Nature Biotechnology*, vol. 27, no. 11, pp. 1013-1023, 2009.

[53] M. Metzker, "Sequencing technologies — the next generation", *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31-46, 2009.

[54] M. Fedurco, "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies", *Nucleic Acids Research*, vol. 34, no. 3, pp. e22-e22, 2006.

[55] D. Bentley, S. Balasubramanian, H. Swerdlow, G. Smith, J. Milton, C. Brown, K. Hall, D. Evers, C. Barnes, H. Bignell, J. Boutell, J. Bryant, R. Carter, R. Keira Cheetham, A. Cox, D. Ellis, M. Flatbush, N. Gormley, S. Humphray, L. Irving, M. Karbelashvili, S. Kirk, H. Li, X. Liu, K. Maisinger, L. Murray, B. Obradovic, T. Ost, M. Parkinson, M. Pratt, I. Rasolonjatovo, M. Reed, R. Rigatti, C. Rodighiero, M. Ross, A. Sabot, S. Sankar, A. Scally, G. Schroth, M. Smith, V. Smith, A. Spiridou, P. Torrance, S. Tzonev, E. Vermaas, K. Walter, X. Wu, L. Zhang, M. Alam, C. Anastasi, I. Aniebo, D. Bailey, I. Bancarz, S. Banerjee, S. Barbour, P. Baybayan, V. Benoit, K. Benson, C. Bevis, P. Black, A. Boodhun, J. Brennan, J. Bridgham, R. Brown, A. Brown, D. Buermann, A. Bundu, J. Burrows, N. Carter, N. Castillo, M. Chiara E. Catenazzi, S. Chang, R. Neil Cooley, N. Crake, O. Dada, K. Diakoumakos, B. Dominguez-Fernandez, D. Earnshaw, U. Egbujor, D. Elmore, S. Etchin, M. Ewan, M. Fedurco, L. Fraser, K. Fuentes Fajardo, W. Scott Furey, D. George, K. Gietzen, C. Goddard, G. Golda, P. Granieri, D. Green, D. Gustafson, N. Hansen, K. Harnish, C. Haudenschild, N. Heyer, M. Hims, J. Ho, A. Horgan, K. Hoschler, S. Hurwitz, D. Ivanov, M. Johnson, T. James, T. Huw Jones, G. Kang, T. Kerelska, A. Kersey, I. Khrebtukova, A. Kindwall, Z. Kingsbury, P. Kokko-Gonzales, A. Kumar, M. Laurent, C. Lawley, S. Lee, X. Lee, A. Liao, J. Loch, M. Lok, S. Luo, R. Mammen, J. Martin, P. McCauley, P. McNitt, P. Mehta, K. Moon, J. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. Novo, M. O'Neill, M. Osborne, A. Osnowski, O. Ostadan, L. Paraschos, L. Pickering, A. Pike, A. Pike, D. Chris Pinkard, D. Pliskin, J. Podhasky, V. Quijano, C. Raczy, V. Rae, S. Rawlings, A. Chiva Rodriguez, P. Roe, J. Rogers, M. Rogert Bacigalupo, N. Romanov, A. Romieu, R. Roth, N. Rourke, S. Ruediger, E. Rusman, R. Sanches-Kuiper, M. Schenker, J. Seoane, R. Shaw, M. Shiver, S. Short, N. Sizto, J. Sluis, M. Smith, J. Ernest Sohna Sohna, E. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. Virk, S. Wakelin, G. Walcott, J. Wang, G. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. Mullikin, M. Hurles, N. McCooke, J. West, F. Oaks, P. Lundberg, D. Klenerman, R. Durbin and A. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature*, vol. 456, no. 7218, pp. 53-59, 2008.

[56] S. Andrews, "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data", *Bioinformatics.babraham.ac.uk*. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. [Accessed: 03- Sep- 2017].

[57] Y. Jung, L. Luquette, J. Ho, F. Ferrari, M. Tolstorukov, A. Minoda, R. Issner, C. Epstein, G. Karpen, M. Kuroda and P. Park, "Impact of sequencing depth in ChIP-seq experiments", *Nucleic Acids Research*, vol. 42, no. 9, pp. e74-e74, 2014.

[58] B. Ewing and P. Green, "Base-Calling of Automated Sequencer Traces Using Phred.II. Error Probabilities", *Genome Research*, vol. 8, no. 3, pp. 186-194, 1998.

[59] A. Bolger, M. Lohse and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data", *Bioinformatics*, vol. 30, no. 15, pp. 2114-2120, 2014.

[60] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform", *Bioinformatics*, vol. 25, no. 14, pp. 1754-1760, 2009.

[61] B. Langmead, C. Trapnell, M. Pop and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome Biology*, vol. 10, no. 3, p. R25, 2009.

[62] B. Langmead and S. Salzberg, "Fast gapped-read alignment with Bowtie 2", *Nature Methods*, vol. 9, no. 4, pp. 357-359, 2012.

[63] M. Burrows and D. Wheeler, "A block sorting lossless data compression algorithm", *Digital Equipment Corporation*, 1994.

[64] B. Samans, Y. Yang, S. Krebs, G. Sarode, H. Blum, M. Reichenbach, E. Wolf, K. Steger, T. Dansranjavin and U. Schagdarsurengin, "Uniformity of Nucleosome Preservation Pattern in Mammalian Sperm and Its Connection to Repetitive DNA Elements", *Developmental Cell*, vol. 30, no. 1, pp. 23-35, 2014.

[65] "What's a Genome?", *Genomenewsnetwork.org*, 2017. [Online]. Available: http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp4_1.shtml. [Accessed: 07- Oct- 2017].

[66] M. Dozmorov, I. Adrianto, C. Giles, E. Glass, S. Glenn, C. Montgomery, K. Sivils, L. Olson, T. Iwayama, W. Freeman, C. Lessard and J. Wren, "Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data", *BMC Bioinformatics*, vol. 16, no. 13, 2015.

[67] "How PCR duplicates arise in next-generation sequencing", *Cureffi.org*, 2017. [Online]. Available: http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/. [Accessed: 28- Sep- 2017].

[68] H. Pearson, "What is a gene?", *Nature*, vol. 441, no. 7092, pp. 398-401, 2006.

[69] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nussbaum, R. Myers, M. Brown, W. Li and X. Liu, "Model-based Analysis of ChIP-Seq (MACS)", *Genome Biology*, vol. 9, no. 9, p. R137, 2008.

[70] T. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio and L. Elo, "A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments", *BMC Genomics*, vol. 10, no. 1, p. 618, 2009.

[71] E. Wilbanks and M. Facciotti, "Evaluation of Algorithm Performance in ChIP-Seq Peak Detection", *PLoS ONE*, vol. 5, no. 7, p. e11471, 2010.

[72] M. Rye, P. Sætrom and F. Drabløs, "A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs", *Nucleic Acids Research*, vol. 39, no. 4, pp. e25-e25, 2010.

[73] M. Micsinai, F. Parisi, F. Strino, P. Asp, B. Dynlacht and Y. Kluger, "Picking ChIP-seq peak detectors for analyzing chromatin modification experiments", *Nucleic Acids Research*, vol. 40, no. 9, pp. e70-e70, 2012.

[74] H. Koohy, T. Down, M. Spivakov and T. Hubbard, "A Comparison of Peak Callers Used for DNase-Seq Data", *PLoS ONE*, vol. 9, no. 5, p. e96303, 2014.

[75] A. Szalkowski and C. Schmid, "Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts", *Briefings in Bioinformatics*, vol. 12, no. 6, pp. 626-633, 2010.

[76] R. Thomas, S. Thomas, A. Holloway and K. Pollard, "Features that define the best ChIP-seq peak calling algorithms", *Briefings in Bioinformatics*, p. bbw035, 2016.

[77] Q. Song and A. Smith, "Identifying dispersed epigenomic domains from ChIP-Seq data", *Bioinformatics*, vol. 27, no. 6, pp. 870-871, 2011.

[78] J. Wang, V. Lunyak and I. Jordan, "BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets", *Bioinformatics*, vol. 29, no. 4, pp. 492-493, 2013.

[79] A. Lun and G. Smyth, "csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows", *Nucleic Acids Research*, vol. 44, no. 5, pp. e45-e45, 2015.

[80] M. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj and F. Jaffrezic, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis", *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671-683, 2012.

[81] M. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data", *Genome Biology*, vol. 11, no. 3, p. R25, 2010.

[82] C. Law, Y. Chen, W. Shi and G. Smyth, "voom: precision weights unlock linear model analysis tools for RNA-seq read counts", *Genome Biology*, vol. 15, no. 2, p. R29, 2014.

[83] M. Ritchie, B. Phipson, D. Wu, Y. Hu, C. Law, W. Shi and G. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies", *Nucleic Acids Research*, vol. 43, no. 7, pp. e47-e47, 2015.

[84] L. Chen, C. Wang, Z. Qin and H. Wu, "A novel statistical method for quantitative comparison of multiple ChIP-seq datasets", *Bioinformatics*, vol. 31, no. 12, pp. 1889-1896, 2015.

[85] S. Steinhauser, N. Kurzawa, R. Eils and C. Herrmann, "A comprehensive comparison of tools for differential ChIP-seq analysis", *Briefings in Bioinformatics*, p. bbv110, 2016.

[86] M. Robinson, D. McCarthy and G. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics*, vol. 26, no. 1, pp. 139-140, 2009.

[87] G. Yu, L. Wang and Q. He, "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization", *Bioinformatics*, vol. 31, no. 14, pp. 2382-2383, 2015.

[88] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. Lin, P. Laslo, J. Cheng, C. Murre, H. Singh and C. Glass, "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities", *Molecular Cell*, vol. 38, no. 4, pp. 576-589, 2010.

[89] C. McLean, D. Bristor, M. Hiller, S. Clarke, B. Schaar, C. Lowe, A. Wenger and G. Bejerano, "GREAT improves functional interpretation of cis-regulatory regions", *Nature Biotechnology*, vol. 28, no. 5, pp. 495-501, 2010.

[90] P. Khatri, M. Sirota and A. Butte, "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges", *PLoS Computational Biology*, vol. 8, no. 2, p. e1002375, 2012.

[91] V. Mootha, C. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. Daly, N. Patterson, J. Mesirov, T. Golub, P. Tamayo, B. Spiegelman, E. Lander, J. Hirschhorn, D. Altshuler and L. Groop, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes", *Nature Genetics*, vol. 34, no. 3, pp. 267-273, 2003.

[92] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander and J. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545-15550, 2005.

[93] A. Sergushichev, "Algorithm for cumulative calculation of gene set enrichment statistic", *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, pp. 956-959, 2016.

[94] B. Giardine, "Galaxy: A platform for interactive large-scale genome analysis", *Genome Research*, vol. 15, no. 10, pp. 1451-1455, 2005.

[95] D. Merkel, "Docker: Lightweight Linux Containers for Consistent Development and Deployment", *Linux Journal*, vol. 2014, no. 239, 2014.

[96] "What is Docker?", *Docker*, 2017. [Online]. Available: https://www.docker.com/what-docker. [Accessed: 08- Oct- 2017].

[97] B. Grüning, M. Beek, B. Batut, J. Chilton, M. ISHII, D. Ryan, E. Afgan, H. Rudolph, K. Ellrott, C. Smith, P. Moreno, H. Hotz, G. Kuster, R. Baertsch, M. Edwards, G. Corguillé, A. Azab, S. Hiltemann, M. Chambers, T. Tanjo, R. Hernández and A. Petkau, "bgruening/docker-galaxy-stable: Galaxy Docker Image 17.05", *Zenodo*, 2017. [Online]. Available: https://zenodo.org/record/583723.

[98] B. Grüning, J. Fallmann, D. Yusuf, S. Will, A. Erxleben, F. Eggenhofer, T. Houwaart, B. Batut, P. Videm, A. Bagnacani, M. Wolfien, S. Lott, Y. Hoogstrate, W. Hess, O. Wolkenhauer, S. Hoffmann, A. Akalin, U. Ohler, P. Stadler and R. Backofen, "The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy", *Nucleic Acids Research*, vol. 45, no. 1, pp. W560-W566, 2017.

[99] Y. Jung, M. Sauria, X. Lyu, M. Cheema, J. Ausio, J. Taylor and V. Corces, "Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes", *Cell Reports*, vol. 18, no. 6, pp. 1366-1382, 2017.

[100] J. Koster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine", *Bioinformatics*, vol. 28, no. 19, pp. 2520-2522, 2012.

[101] R. Dada, M. Kumar, R. Jesudasan, J. Fernández, J. Gosálvez and A. Agarwal, "Epigenetics and its role in male infertility", *Journal of Assisted Reproduction and Genetics*, vol. 29, no. 3, pp. 213-223, 2012.

[102] L. Stuppia, M. Franzago, P. Ballerini, V. Gatta and I. Antonucci, "Epigenetics and male reproduction: the consequences of paternal lifestyle on fertility, embryo development, and children lifetime health", *Clinical Epigenetics*, vol. 7, no. 1, 2015.

[103] C. Ling and L. Groop, "Epigenetics: A Molecular Link Between Environmental Factors and Type 2 Diabetes", *Diabetes*, vol. 58, no. 12, pp. 2718-2725, 2009.

[104] N. Youngson and M. Morris, "What obesity research tells us about epigenetic mechanisms", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1609, pp. 20110337-20110337, 2012.

[105] J. Goecks, A. Nekrutenko, J. Taylor and T. Galaxy Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences", *Genome Biology*, vol. 11, no. 8, p. R86, 2010.

[106] H. Zhang, S. Soiland-Reyes and C. Goble, "Taverna Mobile: Taverna workflows on Android", *EMBnet.journal*, vol. 19, no., p. 43, 2013.

[107] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny and K. Wenger, "Pegasus, a workflow management system for science automation", *Future Generation Computer Systems*, vol. 46, pp. 17-35, 2015.

[108] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes and A. Drummond, "Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data", *Bioinformatics*, vol. 28, no. 12, pp. 1647-1649, 2012.

[109] "Welcome to Planemo's documentation! — Planemo 0.48.0.dev0 documentation", *Planemo.readthedocs.io*, 2017. [Online]. Available: https://planemo.readthedocs.io/en/latest/. [Accessed: 29- Oct- 2017].

[110] D. Blankenberg, G. Von Kuster, E. Bouvier, D. Baker, E. Afgan, N. Stoler, J. Taylor and A. Nekrutenko, "Dissemination of scientific software with Galaxy ToolShed", *Genome Biology*, vol. 15, no. 2, p. 403, 2014.

[111] "Build software better, together", *GitHub*, 2017. [Online]. Available: http://github.com. [Accessed: 26- Oct- 2017].

[112] "Picard Tools - By Broad Institute", *Broadinstitute.github.io*, 2017. [Online]. Available: http://broadinstitute.github.io/picard. [Accessed: 26- Oct- 2017].