KANDIDAT

Ray Mithlesh Kumar (3)

PRØVE

2BIO201 1 Master`s Degree Thesis in Experimental Biotechnology

Emnekode	2BIO201
Vurderingsform	Oppgave
Starttid	11.09.2018 00:00
Sluttid	17.09.2018 12:00
Sensurfrist	--
PDF opprettet	09.04.2019 10:21
Opprettet av	Pål Riiser Berg

ℹ Information

Master thesis <course/code>

Link to practical information about home examination/hand in exams
Link to Studentweb.

You have to answer the questions on the next page.

Your thesis must be uploaded in an PDF format.
The answer paper that is uploaded will automatically be delivered when the submission deadline expires. Until the deadline expires, you can replace your answer paper with new versions.

You need to register the title of your master master's thesis in Studentweb.
This is done by following these steps:
1. Log in to Studentweb.
2. Click on "Active courses" in the top menu, and then select the master’s thesis course.
3. To add a title to the master’s thesis, click on "Edit" on the right side of the page, then click "Show assignment title".
4. In the field “Assignment title in the original language” enter the title in the language in which the thesis is written, and in the field “Assignment title in English” enter the thesis title in English.
5. Click "Save".

If you have any questions, do not hesitate to contact us at name@inn.no

☑ Borrowing and publication agreement form

Publication agreement form

- **AGREEMENT FOR PUBLICATION IN BRAGE**
- **RESTRICTION OF ACCESS TO ASSIGNMENTS**

AGREEMENT FOR BORRWING AND PUBLICATION

Brage is INN University's open institutional archive for academic and scientific work.

For information on what agreeing to borrowing and publishing in Brage involves:

https://eng.inn.no/library/publish

I hereby agree to the publication of the assignment in Brage:

● Yes

○ No

The student has an unlimited right to withdraw her/his consent for publishing.

RESTRICTION OF ACCESS ON THE BASIS OF LEGAL CONFIDENTIALITY REQUIRMENTS / OTHER GROUNDS

To be filled in only if access restriction is applicable:

The access to the assignment is restricted on the following basis:

○ The assignment contains information that is subject to confidentiality – indefinite restriction

● The assignment is to be used in further research – time-limited restriction for up to 5 years

In addition, it shall be stated in the body of the assignment if it is restricted, and the reason for this. Restricted assignments will not be published.

To be filled in for a time-limited restriction:

Grounds for time-limited restriction

Duration of time-limited restriction (up to 5 years): : 2.0

Besvart.

1

Upload your PDF here

✓

Din fil ble lastet opp og lagret i besvarelsen din.

Last ned | ✖ Fjern | Erstatt

Filnavn:	Ray_Mithlesh_master thesis.pdf
Filtype:	application/pdf
Filstørrelse:	2.47 MB
Opplastingstidspunkt:	18.10.2018 11:09
Status:	**Lagret**

Besvart.

INLAND NORWAY
UNIVERSITY
OF APPLIED SCIENCES

Faculty of Applied Ecology, Agricultural Sciences and Biotechnology

Mithlesh Kumar Ray

Master's Thesis

Bioinformatics analysis of epigenetic data from bull and boar semen samples

Master's in Applied and Commercial Biotechnology

2018

Acknowledgement

The present master thesis was carried out at Inland Norway University of Applied Sciences from August 2017 to November 2018 under the supervision of Associate Professor Dr. Rafi Ahmad and Professor Elisabeth Kommisrud.

It is great opportunity for me to express my deepest appreciation for the commendable efforts of my supervisor, Dr. Rafi Ahmad, who has the attitude and the substance of a genius. I can’t thank him enough for such a passionate supervision, encouragement, and guidance during this project. I would also like to pay a sincere gratitude to my co-supervisor Prof. Elisabeth Kommisrud. Her wide knowledge and logical way of thinking has been of great value for me.

It is a genuine pleasure to acknowledge the efforts of Abdul Rahman (Post-doctorate). His understanding, encouragement, and sincere advice throughout the duration of this project have provided a sound base for the successful completion of my thesis.

Lastly, I would like to Thank the whole staff members of Inland Norway university, for making my stay here a cherished memory forever.

Hamar, 18th, Oct 2018

Mithlesh Kumar Ray

Abbreviations

AMZ1	Archaemetzincin-1
AO	Acridine Orange
ART	Assisted Reproductive Technologies
ATM	Ataxia Telangiectasia Mutated
BH	Benjamini-Hochberg
BIRC5	Baculoviral Inhibitor of Apoptosis Repeat-Containing 5
BLAST	Basic Local Alignment Search Tool
cAMP	Cyclic Adenosine Monophosphate
CAMs	Cell Adhesion Molecules
CATSPER4	Cation Channel Sperm-Associated Protein 4
Ccdc42	Coiled-Coil Domain Containing 42
CGIs	CpG Islands
CTSA	Cathepsins A
DFI	DNA Fragmentation Index
DMCs	Differentially Methylated Cytosines
DMRs	Differentially Methylated Regions
DNMT	DNA Methyltransferase
dNTP	Deoxyribonucleotide Triphosphate
FDR	False Discovery Rate
FGF9	Fibroblast Growth Factor 9
FMO1	Flavin Containing Monooxygenase 1
GLRX3	Glutaredoxin 3
GM-CSF	Granulocyte-Macrophage Colony-Stimulating Factor
GO	Gene Ontology
HDACs	Histone Deacetylases
HOPX	HOP Homeobox
HTLV-I	Human T-Lymphotropic Virus-1
IL6	Interleukin-6
INOA	Idiopathic Nonobstructive Azoospermia
IRF4	Interferon Regulatory Factor 4
IVF	In Vitro Fertilization

IVP	In Vitro Production
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KLF5	Kruppel Like Factor 5
LMO4	LIM Domain Only 4
lncRNAs	Long Noncoding RNAs
MBDs	Methylated CpG Binding Protein
MECP2	Methyl-CpG-Binding Protein 2
mRNA	Messenger RNA
NAFLD	Non-Alcoholic Fatty Liver Disease
NCBI	National Center For Biotechnology Information
NGS	Next Generation Sequencing
NRF	Norwegian Red
OPU	Ovum Pickup
ORA	Over Representation Analysis
PCR	Polymerase Chain Reaction
PNMS	Prenatal Maternal Stress
qPCR	Quantitative Polymerase Chain Reaction
ROS	Reactive Oxygen Species
RRBS	Reduced Representation Bisulfite Sequencing
SAM	S-Adenosylmethionine
SCD	Sperm Chromatin Dispersion
SCSA	Sperm Chromatin Structure Assay
SHH	Sonic Hedgehog
SLC7A1	Solute Carrier Family 7 Member 1
SMAD4	SMAD Family Member 4
sncRNA	Small Non-Coding RNA
SOX2	SRY-Box 2
SPATA20	Spermatogenesis-Associated Protein 20
SPRI	Solid Phase Reversible Immobilization
TERT	Telomerase Reverse Transcriptase
TGF-beta	Transforming Growth Factor-B
TSSs	Transcription Start Sites
TUNEL	Terminal Transferase Dutp Nick End Labelling

Table of Contents

Abstract

The study of the correlation between reproduction and epigenetics is a fascinating subject of study. This is due to the transgenerational epigenetic inheritance via the male gametes and its effect on the health and fertility of the offspring. Pig and cattle are important animals for the veterinary industry. Pig is a leading source of meat and an important animal model for biomedical research, while cattle is important for meat production and dairy farming. To have a sustained breeding program, it is critical to study their reproduction and epigenetics. DNA methylation is one of the critical epigenetic mechanism involved in gene expression regulation. Advancements in next-generation sequencing methods and their combination with bioinformatics, helps us to efficiently study epigenetics.

The primary aim of this study was the bioinformatic analysis of epigenetic data obtained using reduced representative bisulfite sequencing (RRBS) in sperm cells with different level of DNA fragmentation index (a phenotypic data) from boar and bull semen samples. To achieve this goal, we established an efficient bioinformatics pipeline for analysis of data from RRBS. The proposed pipeline consists of six steps (i) quality trimming using trim-galore, followed by quality check using fastQC, (ii) preparation of the reference genome, mapping to the reference genome and methylation calling with bismark (iii) finding differentially methylated Cs (DMCs) and regions using methylkit. (v) functional annotation of regions of interests and (vi) pathway and enrichment analysis, using Blast2GO and WebGestalt.

Majority of Cs were differentially hypomethylated between low and high groups (5 samples each). However, increasing DFI resulted in increased rate of hypermethylated Cs between low, medium and high groups (3 samples each). We did not find any consistence effect of DFI on percentage of methylated Cs on CpG. Additionally, nearest annotated TSS and their associated genes to DMCs, were found to be involved in different metabolic processes, some of which are known to be involved in reproduction and spermatogenesis. The KEGG pathway analysis showed that five pathways were common (what is the process in which they are involved) in all the studied groups.

Results from this study might be useful in combination with other ongoing work involving metabolomics and other sperm quality analyses. Taken all together, our findings suggest that sperm cells with different DNA damage could have different DNA methylation and the bismark and methylkit based method for finding DMCs provides more authentic and reliable results that can be compared with previous studies.

1. Background

1.1 Origin of the research

This master thesis is the part of an ongoing Research Council of Norway (RCN) funded project entitled “**Elucidation of underlying factors influencing fertility in modern, efficient livestock production through novel epigenomics and metabolomics**”. The research has been conducted in collaboration with Geno SA, Norsvin, SINTEF and Inland Norway University of Applied Sciences (INN University). Geno is the breeding company for Norwegian Red (NRF) dairy cattle that conduct research and development of the Norwegian Red breed. Its core business is the development, production, and sale of semen and embryos, in Norway and worldwide (Geno SA, 2018). Norsvin is a breeding company owned by Norwegian pig producers. The key tasks of Norsvin are the production of healthy, cost-effective pigs as well as the development, production, and sale of pig genetic engineering as the core focus (Norsvin SA, 2018).

Assisted reproductive technologies (ART), such as *in vitro* fertilization (IVF) and especially intracytoplasmic sperm injection (ICSI), have opened up new approaches for infertility treatment in human. Lately, also in cattle and swine breeding different ART like *in vitro* production (IVP) of embryos including ovum pickup (OPU) and IVF have been introduced as valuable breeding tools (Parrish, 2014). Selection of suitable gamete is the key successful outcome of ART. Study on sperm genetics, epigenetics and metabolomics may play an important role to establish novel marker to improve sustainable health and fertility in cattle and swine.

The correlation between reproduction and epigenetics depict a fascinating subject of study, mainly due to the epigenetic modification of sperm cells and their possible transgenerational effect on health and fertility of offspring. An environmental factor can induce a permanent epigenetic change in the male gamete during embryonic and foetal development that can transmit epigenetic transgenerational inheritance to offspring (Manikkam et al., 2012).

As DNA methylation plays a vital role in gene expression, the bioinformatics analysis of methylation pattern and associated gene in combination with lab verification may be helpful to find out novel biomarker for semen quality. The study can also be beneficial for the health improvement of bulls and boars. Therefore, this thesis seeks to investigate methylation pattern in the different level of DFI and associated genes in bull and boar sperm sample.

2. Introduction

2.1 Sperm cells

A sperm is the smallest, compact cell that is highly adapted for delivering its DNA to an egg during fertilization. Sperm cells are equipped with a strong flagellum to propel them through an aqueous medium as well as unburdened by cytoplasmic organelles such as endoplasmic reticulum, ribosome or Golgi apparatus (Alberts B, 2002). However, the spermatozoon is not only delivering the genome to the oocyte but also plays a central role in the activation of the oocyte arrested at the metaphase of the meiosis II. Sperm usually consists of two morphologically and functionally distinct regions, a head and a flagellum tail (*Figure 1*) (Toshimori, 2009).

Head: The head is spherical or oval shaped containing an acrosome and highly condensed DNA in the nucleus. The nucleus comprises of densely packed chromatin so that it is reduced in volume for transport and transcription stops (Alberts B, 2002; Pesch and Bergmann, 2006). The linker histone proteins in the nucleus have been partially replaced by highly positively charged proteins called protamine, that convey the hyper condensation of sperm nucleus into compact, hydrodynamic shape permitting sperm motility and allowing sperm penetration through egg vestment (Alberts B, 2002; De Jonge et al., 2006). The sperm is haploid, and the nucleus contains half the number of chromosome than the chromosome of the somatic cell of the same species (Alberts et al., 2008). The spermatogonial stem cells differentiate into spermatocytes and the haploid sperm cells as a result of a meiotic cell division of the tetraploid primary spermatocyte (Nishimura and L’Hernault, 2017). The acrosomal vesicle covers the anterior part of the head. The acrosome is bounded by the double-layered an acrosomal membrane cap-like structure, which contains hydrolytic enzymes (*Figure 1*). The hydrolytic enzymes play a significant role in fertilization to penetrate the female egg’s coat (Hafez and Hafez, 2000). During fertilization, in acrosome reaction, the contents of the vesicle are released by exocytosis. This reaction also releases specific proteins to help the sperm to bind tightly to the egg coat (Alberts B, 2002).

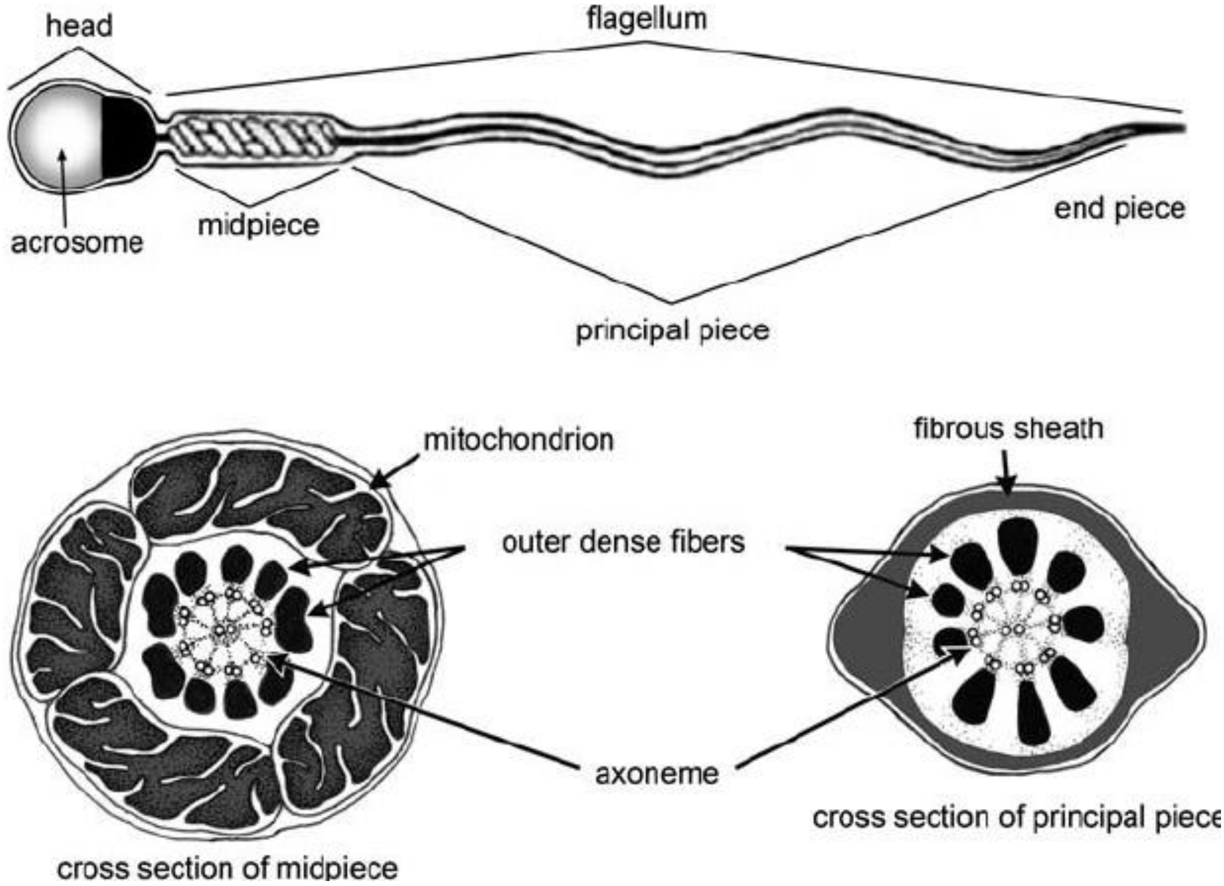


Figure 1. ***Illustration of the human sperm cell structure.*** *The sperm cell consists of two morphologically and functionally distinct regions, a head and a flagellum. Head has two main parts, acrosome and nucleus. The flagellum is divided into three parts including the mid-piece, principal piece and end-piece. Transverse-section through mid-piece and flagellum showing 9+2 axial filament and 9+2 microtubule doublet arrangement of flagellum respectively. Image taken from (Suarez, 2010).*

Midpiece and Tail: The region between the head and tail is the midpiece. The midpiece consists of two central singlet microtubules surrounded by nine evenly spaced microtubule doublets. The 9 + 2 pattern of microtubules is further surrounded by nine outer dense fibres (*Figure 1*) (Hafez and Hafez, 2000). The nine dense fibres are superficially arranged to the axoneme in the proximal part of the flagellum, these all fibres are thickest at the proximal part of midpiece and progressively thinnest part at the tip of tail (Pesch and Bergmann, 2006). The dense fibres are noncontractile, and they are responsible for the flexibility of the flagellum; any defects in these fibres lead to abnormal sperm morphology and movement. The midpiece is the mitochondria rich part of the sperm cell, which provides the source of energy for sperm motility (Alberts et al., 2008). The sperm axoneme is supported by the fibrous sheath which contains protein kinases essential for the final sperm maturation step prior to fertilization. The tail comprises of the end of outer dense fibres and fibrous sheath and axonemal doublets (Storey, 2006).

2.1.1 DNA integrity

Sperm chromatin integrity is one of the essential clinical parameters of male fertility and also for *in vitro* fertilization potential. Incomplete DNA condensation and damage are correlated with infertility(Chu et al., 2006). A reduced natural pregnancy rate and prolonged time to pregnancy have been associated with high sperm DNA fragmentation, in humans (Zini, 2011). The sperm with severe DNA damage may increase the risk of transmitting genetic aberration to the embryo which may lead to miscarriages or birth of offspring with major or minor congenital malformation (Shamsi et al., 2011). A negative correlation has been shown between sperm chromatin integrity and fertility in both *in vivo* and *in vitro* (Evenson and Wixon, 2006; Spanò et al., 2000). The sperm chromatin is highly compact than the nucleus of somatic cells. The important changes in chromatin occur during haploid phase of spermatogenesis (spermiogenesis) (Boissonneault, 2002; Love and Kenney, 1998). A mature spermatozoon contains highly condensed chromatin due to the replacement of somatic histone proteins by protamine and intermediate proteins, and further stabilization ensures through the formation of disulfide bonds during final nuclear maturation (Sergerie et al., 2005). The complete process of chromatin packaging lead to Six-folds more highly condensed sperm DNA than the mitotic chromosome (Ward and Coffey, 1991).

Male infertility is a critical issue in humans (Sironen, 2018). According to WHO in 1999, approximately 50% of infertility can be attributed solely to the male partner (Agarwal and Said, 2003). A semen analysis that measures sperm concentration, motility and morphology have classically been used as the gold standard test for determining a man's fertility, but these parameters do not assess the quality of the sperm nuclear material (Delbès et al., 2010). Over the last decade, many studies have suggested that DNA damage testing should be included in every early fertility checkpoint to provide information about the chromatin structure of the sperm (Lewis, 2013). Highly refined biochemical events that occur during spermatogenesis can be disturbed by environmental stress, gene mutation, and chromosomal abnormalities, which can eventually lead to an abnormal chromatin structure (Agarwal and Said, 2003).

The packaging of sperm chromatin may serve to reprogram the father's genome, thus the appropriate genes from the paternal chromosomes are expressed in the early embryo (Braun, 2001). Hence, a correct chromatin packaging level seems crucial to express the fertilization capacity of sperm fully (Sergerie et al., 2005). The sperm nuclear chromatin instability is correlated with a reduced breeding efficiency of the boars (Evenson et al., 1994). Both in vivo and in vitro, the sperm with nuclear abnormalities can fertilize oocysts while uncompensable sperm nuclear defects may lead to abnormal embryo development. Sperm DNA damage also affects litter size and pregnancy rate in livestock (Evenson, 1999). Moreover, the previous study on bulls showed the sperm chromatin structure significantly correlated with field fertility (Januskauskas et al., 2001).

DNA damage can be induced by various mechanisms illustrated in *Figure 2*. Induction of apoptosis during spermatogenesis and in the seminiferous tubule epithelium. A germ cell may have disruptive nucleus due to chromatin remodelling which can lead to DNA break during spermatogenesis even though the spermatozoon has normal morphology (Agarwal and Said, 2003; Sakkas and Alvarez, 2010). The production of a high level of reactive oxygen species (ROS) by immature sperm can break DNA in mature sperm leading to post-testicular sperm DNA fragmentation (Sakkas and Alvarez, 2010). In general, during sperm migration to the epididymis from seminiferous tubules, oxygen radicle can cause abnormal DNA integrity.

Moreover, ROS can cause DNA damage through direct or indirect activation of sperm caspases and endonucleases (Agarwal and Said, 2003; Sakkas and Alvarez, 2010). DNA fragmentation can be induced by chemotherapy, radiotherapy and environmental toxicant like high level of air pollution (Sakkas and Alvarez, 2010). Apart from this, some other various causes have also been associated with increased levels of sperm DNA damage such as leukocytospermia, smoking, iatrogenic sperm DNA damage, and disease like cancer (Agarwal and Said, 2003).

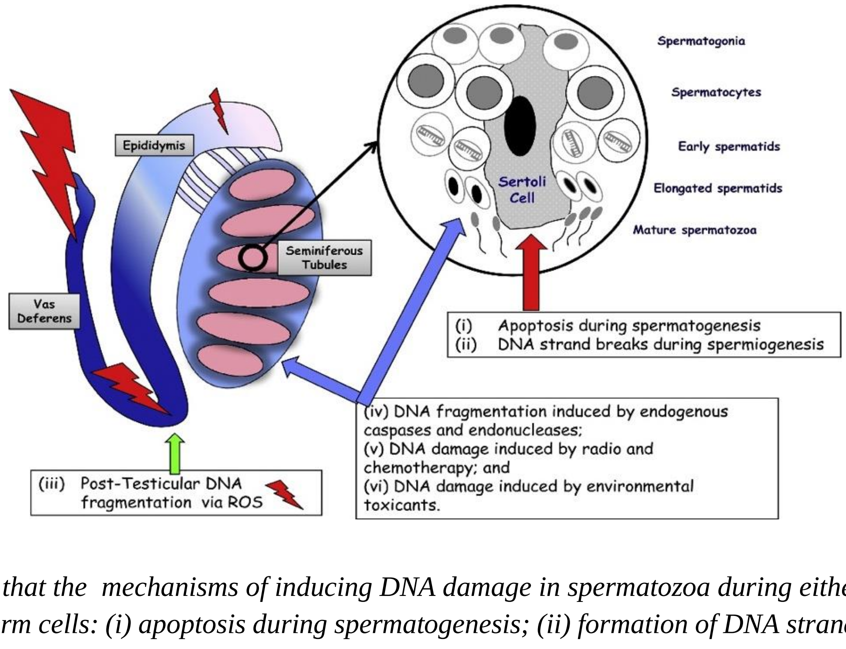

Spermatogonia
Spermatocytes
Early spermatids
Elongated spermatids
Mature spermatozoa
Sertoli Cell
Epididymis
Seminiferous Tubules
Vas Deferens
(i) Apoptosis during spermatogenesis
(ii) DNA strand breaks during spermiogenesis
(iv) DNA fragmentation induced by endogenous caspases and endonucleases;
(v) DNA damage induced by radio and chemotherapy; and
(vi) DNA damage induced by environmental toxicants.
(iii) Post-Testicular DNA fragmentation via ROS

Figure 2. illustrate that the mechanisms of inducing DNA damage in spermatozoa during either the production or the transport of sperm cells: (i) apoptosis during spermatogenesis; (ii) formation of DNA strand breaks during the remodelling of sperm chromatin; (iii) post-testicular DNA fragmentation induced, mainly by ROS, during sperm transport through the reproductive tract (The size of red flashes and gradient darkening in tract indicates the level of DNA damage); (iv) DNA fragmentation induced by endogenous caspases and endonucleases; (v) DNA damage induced by radiotherapy and chemotherapy; and (vi) DNA damage induced by environmental toxicants. Image taken from (Sakkas and Alvarez, 2010).

DNA damage may play a crucial role in embryo development, embryonic cleavage rate, implantation, and spontaneous miscarriages. Poor embryo quality, increased embryo development arrest, and decreased pregnancy rates are associated with higher sperm DNA damage, while good embryo quality, better embryo development, and improved pregnancy rates are associated with low sperm DNA damage (Simon et al., 2014).

The DNA fragmentation index (DFI) is a way to measure the status of chromatin structure. The high value of DFI indicates abnormal chromatin structure (Sergerie et al., 2005). There are several methods to measure DNA damage including sperm chromatin structure assay (SCSA), comet assay, the terminal transferase dUTP nick end labeling (TUNEL) assay, the sperm chromatin dispersion (SCD or Halo) test (Lewis, 2013). The SCSA is one of the most statistically robust tests for sperm DNA fragmentation. The SCSA is a flow cytometry method that utilizes metachromatic properties of acridine orange (AO), which fluoresces green and red when bound to double-stranded and single-stranded DNA, respectively (Evenson, 2016). The SCSA results are further used to calculate DFI (FD Myromslien and TT Zeremichael, 2018).

2.2 Epigenetics

Epigenetics, in a wide sense, is a heritable change in gene expression or cellular phenotype without changing the underlying DNA sequence (Goldberg et al., 2007). The epigenetic mechanisms are essential and natural to many organism functions (Weinhold, 2006). However, epigenetic dysregulation can lead to serious health problems and behavioural effects (Waterland, 2009; Weinhold, 2006).

DNA methylation and histone modification are the crucial epigenetic marks for chromatin activation or inactivation. Post-translational modification of histone can play regulatory roles in gene expression by altering chromatin conformation (Felsenfeld and Groudine, 2003). There are variety of posttranslational modifications including acetylation, methylation and phosphorylation that occur on multiple specific sites of histone (Cheung and Lau, 2005). It has been believed that modified histone can function as a signalling platform by integrating with up-stream signalling pathway to induce transcription activation and repression (Cheung et al., 2000).

Now a days, the epigenetic research involves the study of covalent and noncovalent modification of DNA and histone proteins and the mechanisms by which such modifications influence the complex of proteins (Goldberg et al., 2007). The chromatin can be modified by addition or removal of acetyl groups and some forms of RNA such as microRNAs, small interfering RNAs, and noncoding RNAs (*Figure 3*). This modification alters the chromatin structure to influence gene expression. Mostly, tightly folded chromatin tends to be shut down, or not expressed. However, more open chromatin is functional, or expressed and is crucial for determining time specific and cell-specific manner of gene expression by regulating the DNA wrapped around histone (Schagdarsurengin et al., 2012; Weinhold, 2006).

Small non-coding regulatory RNA play regulatory role in gene expression which can repress gene expression through the induction of DNA methylation and modification in chromatin structure (Godfrey et al., 2007). Long noncoding RNAs are characterized as epigenetic modulator, and it has a regulatory role at almost every stage of gene expression. The lncRNAs guide the catalytic activity of chromatin-modifying proteins at the specific site in the genome by binding with them (Mercer and Mattick, 2013). And sncRNA in gametes may have role in post-fertilization epigenetic reprograming (Gluckman et al., 2007).

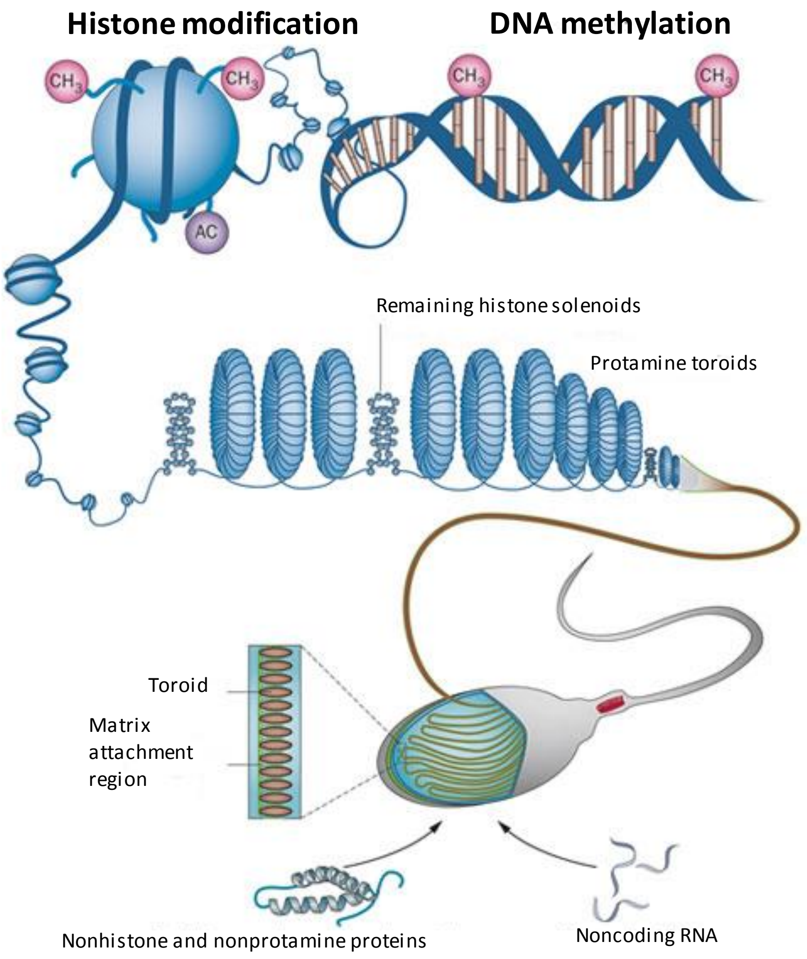


Figure 3. ***The figure illustrates brief outline of sperm epigenetics.*** *During spermatogenesis, DNA-binding protein histones are replaced by protamines. Hence, sperm chromatin consists of nucleohistones coiled into solenoids and nucleoprotamines coiled into toroids by attaching to matrix attachment regions. The remaining histones in sperm are bearing other modification like methylation and highly acetylated, as sperm is known to be transcriptionally inactive cell. Sperm also contains noncoding RNA, silent mRNAs as well as nonhistone and nonprotamine proteins. Image taken from (Schagdarsurengin et al., 2012).*

In males, precise sperm DNA methylation is crucial for fertilization as well as early foetus viability. The relationship between aberrant DNA methylation and male subfertility have been evaluated, in various previous studies, which might be useful to explain male infertility (Laqqan and Hammadeh, 2018). The previous work suggested that sperm DNA methylation abnormalities and variation in mRNA content have been seen in the infertile male. Methylation profile is also associated with motility, as increased hypomethylation was observed in low motile sperm as compared to high motile sperm (Pacheco et al., 2011). The offspring inherits the basic DNA sequence along with the program of gene expression proposed by the parents' epigenetic machinery and an environmental change that results in a change in a sperm-born ncRNA with post- transcriptional gene silencing impact in embryogenesis (Daxinger and Whitelaw, 2012). In such a way, an environmental agent is capable of promoting epigenetic transgenerational changes in the sperm epigenome (Guerrero-Bosagna et al., 2010).

Epigenetics programming has been considered as one of the crucial mechanisms for the long-term effects of exposure to stress in utero (Cao-Lei et al., 2016). Prenatal maternal stress (PNMS) can have negative impact on health outcomes in later life (Beydoun and Saftlas, 2008). The severe prenatal stress can be due to natural and human disasters like famines, environmental pollutants, nutritional factor, earthquakes, maternal depression during pregnancy, and terrorist attacks (Cao-Lei et al., 2016; Eriksson, 2010). For instance, the significant lower birth weight has been observed after World Trade Center attacks in 2001 for those who lived close to the event site (Berkowitz et al., 2003). Additionally, ‘Project Ice Storm’ revealed changes in genome-wide DNA methylation levels triggered by in utero exposure to ice storm mediates the effect of PNMS on child immune and metabolic outcomes (Cao-Lei et al., 2014).

In order to optimize fertilization and assisted reproduction technique in future, an understanding of epigenetic mechanisms involved in spermatogenesis and their impact on embryogenesis is essential (Schagdarsurengin et al., 2012). The sperm cells have distinctive function and morphology to assist fertilization. The genetic material (DNA) is highly condensed to protect the parental genome during transfer to the oocyte. During differentiation to become a mature spermatozoon, sperm cells go through extensive epigenetic modification (Güneş and Kulaç, 2013; Schagdarsurengin et al., 2012). Epidemiology and laboratory studies suggest that environmental factors including parental nutrition, toxic exposure, early environment, paternal age, and phenotypic variation can promote variation in offspring (Curley et al., 2011; Franklin and Mansuy, 2010). Well defined reasons for male-factor infertility include anatomic defects, chromosomal abnormalities, and point mutations. However, these diagnoses represent only a small fraction of patients, and causes remain unexplored for the majority of male-factor infertility cases (Houshdaran et al., 2007). Some studies suggest that the abnormal sperm epigenetic programming may contribute to some cases of male factor infertility (Rousseaux et al., 2005) and also in the production of good quality sperm (Congras et al., 2014).

The healthy viable offspring, in an agriculturally important domestic animal, are associated with epigenetic mechanisms. Therefore, how environmental factors affect epigenetics in both animals and their offspring has been studied (Feeney et al., 2014). For instance, phenotypic transgenerational epigenetic response found on F2 generation with lower fat percentage and higher shoulder muscles percentage in comparison to control, where the experimental group F0 generation male was fed highly methyl-enriched diet (Braunschweig et al., 2012). Previous studies showed that exposed male guinea pigs to different temperature conditions, during spermatogenesis, led to immediate and heritable epigenetic response. The methylation pattern of liver and testes, after heat treatment, was inherited to F2 generation (Weyrich et al., 2016).

2.2.1 DNA methylation

In most higher organisms, DNA is modified after synthesis by the enzymatic conversion of many cytosine residues to 5-methylcytosine known as DNA methylation (Razin and Riggs, 1980). DNA methylation is a crucial element behind epigenetic changes in gene expression in diverse species including mammals (Choi et al., 2015a). From the recent studies, it is believed that 5-methylcytosine is a crucial element in the hierarchy of control mechanisms that govern vertebrate gene function, gene silencing, genomic imprinting, X -chromosome inactivation, cancer progression, embryonic development and differentiation (Choi et al., 2015b; Razin and Riggs, 1980). In mammals, DNA methylation occurs via an enzyme called DNA methyltransferase which attaches a methyl group to the number 5 position of the cytosine base forming 5-methylcytosine. The methyl group donated by S-Adenosylmethionine (SAM) for the chemical modification of DNA and the reaction is catalysed by DNA methyltransferase (DNMT) (*Figure 4*).

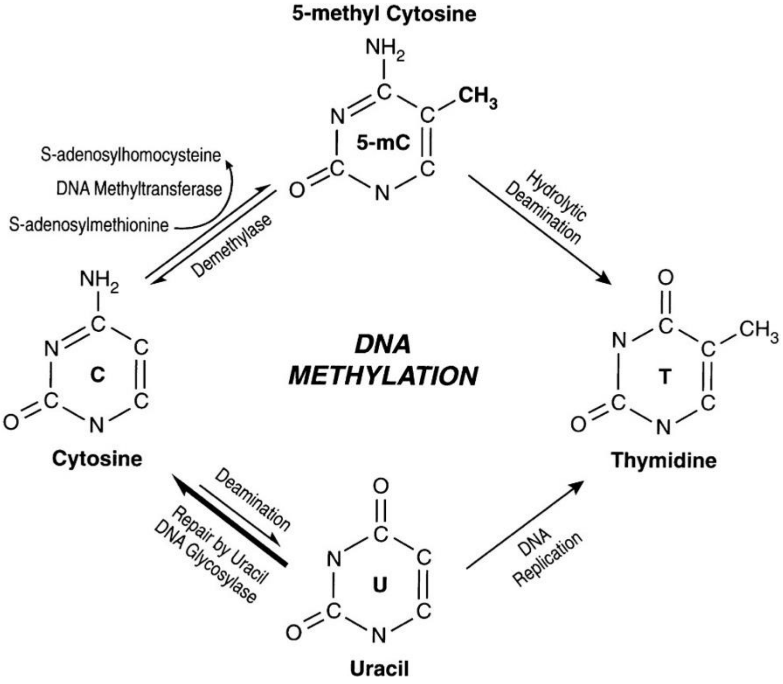


Figure 4. ***The schematic representation shows DNA methylation.*** *The mechanism of DNA methylation, conversion of cytosine into 5'methyl-cytosine with the action of DNMT, demethylation and mutagenesis of cytosine and 5-mC. Figure is taken from (Singal and Ginder, 1999).*

Methylation occurs preferentially at cytosine and guanine dinucleotides which usually inhibits the transcription of the gene, particularly when it occurs in the vicinity of the promoter (*Figure 5*). CpG islands are commonly 0.2 to 2 kb in length and contain a high number of CpG sites. Nearly 50-60 % of promoter regions are associated with CpG islands (Choi et al., 2015a; Ulrey et al., 2005). Phenotypically relevant differences in gene expression may also be associated with variability in the relative methylation levels (Nikolova and Hariri, 2015).

DNA methylation is crucial for mammals' development with the potential role in tissue-specific gene expression. Methylation of a promoter region CpG island can repress transcription initiation by binding with methylated CpG binding protein (MBDs) and transcription repressor including histone deacetylases (HDACs) (Jones and Takai, 2001). For instance, the transcription is repressed in a methylation-dependent manner by the complex formation of MECP2 with HDACs complemented by co-repressor protein Sin3a (Jones et al., 1998). High level of global methylation was observed in swine sperm DNA as compared to other mammals, including humans and mice (Congras et al., 2014). Various current studies have shown that the influence of methylation patterns of livestock phenotypes is associated with disease resistance, milk production, and reproduction. However, DNA methylation patterns in pig placental tissues are also highly associated with litter size (Hwang et al., 2017).

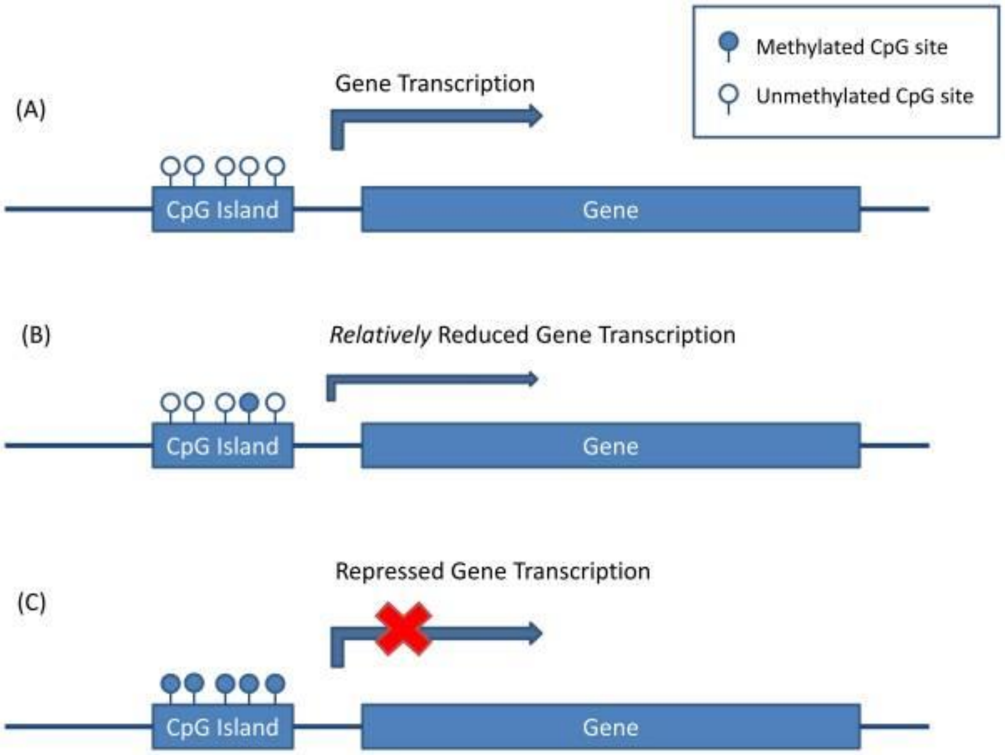


Figure 5. Methylation of CpG island. A. hypomethylation associated with normal gene expression. B. The relatively methylation. C. hypermethylation associated with repression of gene expression. Image taken from (Nikolova and Hariri, 2015).

2.3 A brief overview of the pig genome

The domestic pig (*Sus scrofa*) is a member of Cetartiodactyla order and eutherian mammals which is a clade distinct from the primates and rodents. The haploid genome of the domesticated pig, based on the assembly Sscrofa11.1, is estimated to be 2.80 Gb. The diploid genome is organized in 20 pairs of chromosomes (18 pairs of autosomes, X and Y chromosome) (NCBI, 2018). The statistics obtained, based on Sscrofa11.1 assembly, contains a median total length 2457.91(Mb), median protein count 63577, and median GC% 41.5. The pig is an important biomedical model as its physiology is very similar to that of the humans. The pig has been used as a model for many areas of highly prevalent human disease like diabetes, metabolic disorder, cardiovascular disease, and obesity (Bassols et al., 2014). Short generation interval and large litter size of pigs make them suitable as animal models (Choi et al., 2015b) Further, a comparison between predicted porcine protein sequence with human orthologues gave 112 position where the porcine protein has the same amino acid that is involved in human disease (Groenen et al., 2012). The interest in the pig as an animal model for biomedical research due to an integration of different ‘omics’ data, both from pigs and humans, will come up with a better understanding of biological elements with an impact on production traits (Bassols et al., 2014). As pork is a leading source of protein, it becomes an agriculturally important animal.

2.4 CpG island

The CpG islands (CpGIs) are defined as a short stretch of DNA with the frequency of the CG sequences generally higher than other regions. The location of CpG island is mainly in the 5' regulatory region of all housekeeping genes, and more than 40 % of tissue- specifically expressed genes. It accounts for nearly 1-2% of the genome (Plass, 2001). The criteria that define CpG island, are; the sequence is longer than 200 base pairs, the GC content is above 50%, and the CpG ratio (observed/expected) is above 0.6 (Gardiner-Garden and Frommer, 1987). However, many repeat elements also satisfy these criteria in the genome. To solve the repeat problem more stringent criteria are; the sequence is longer than 500 base pairs, the GC content is above 55%, and the CpG ratio (observed/expected) is above 0.65 (Takai and Jones, 2002). CpG islands are predominantly nonmethylated, and approximately 70% of promoter regions are associated with CGIs. Current work has revealed a large class of CGIs that are far away from transcription start sites (TSSs), but in spite of that they show evidence for promoter function. These findings show a strong correlation between CGIs and transcription initiation (Deaton and Bird, 2011). The sequences up to 2 kb, immediately flanking CpG island are termed as CpG- shores. CpG- shores methylation is also strongly related to gene expression (Irizarry et al., 2009).

2.4.1 Bisulfite sequencing and Reduced Representation Bisulfite Sequencing (RRBS)

Whole genome bisulfite sequencing (WGBS) is a widely used approach to detect methylation patterns. The ideal method for DNA methylation analysis of individual gene starts with random fragmentation for the desired size and library preparation. The library sequence than go through sodium bisulfite treatment to convert unmethylated cytosines to uracil followed by cloning and sequencing (Wreczycka et al., 2017).

However, this method is restricted to DNA methylation, and it is a gold standard because of single base pair resolution. Later, genome-wide high throughput bisulfite sequencing studies at single base pair resolution have been performed especially for small genome (Lu et al., 2015). The advantage of this approach is that it can reach over 90 % of CpG in unbiased representation (Wreczycka et al., 2017).

However, for many reasons, the bisulfite approach has been difficult to be applied on a genome- wide scale. First, DNA methylation is concentrated at repetitive elements, but it is challenging for short sequence reads, corresponding to repeats, to assemble uniquely onto bisulfite-converted genome sequence. Second, the genome complexity can be reduced due to the conversion of unmethylated C to U/T during bisulfite treatment. Third, sequencing the methylome of whole genome is cost- effective (Gu et al., 2011).

Therefore, to avoid the prohibitive cost of large genome bisulfite sequencing, it is more practical to investigate part of the genome in order to obtain methylation patterns of mammals (Lu et al., 2015). Meissner *et al*. developed the RRBS technique in 2005 that examines DNA fragments from a small proportion or reduced representation of the bisulfite-treated genome. The DNA digestion with the methylation-insensitive enzyme such as *Msp1* leads to the genome reduction (Fouse et al., 2010). However, the fragment that incorporates the reduced sequences of genome still includes the majority of promoters as well as regions such as repeated sequences that are difficult to depict with the use of conventional bisulfite sequencing approach (Gu et al., 2010a).

The essential step for accurate determination of methylation pattern is the complete conversion of unmethylated cytosine to uracil; this can be achieved by treating the DNA with a high concentration of bisulfite salt at high temperature and low pH. Generally, these harsh conditions give rise to a high degree of DNA fragmentation, and subsequent DNA loss during purification. Purification is essential to remove bisulfite salt and chemicals used that can inhibit sequencing procedure (Qiagen, 2009b).

2.4.2 Next Generation Sequencing

The field of genomics has changed dramatically since the introduction of next-generation sequencing (NGS). Using different approaches, either creating micro-reactors or the DNA molecules to be sequenced that are attaching to solid surface, the NGS methods defeat the limited measurability of traditional Sanger sequencing, and are allowing for millions of sequencing reaction to occur in parallel (Reis-Filho, 2009). The way researchers think about scientific approaches in basic, applied, and clinical research have changed with the arrival of NGS technologies in the marketplace (Michael, 2009). The current approaches to next-generation sequencing have become widely available, cost-efficient. Additionally, they are standardizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators (Shendure and Ji, 2008). Several methods, involved in sequencing technology, are broadly classified as templet preparation, sequencing and imaging, and data analysis. The fragment templates or mate-pair templates can be created by randomly breaking genomic DNA into smaller size, and the templet is immobilized to a solid surface that allows thousands to billions of sequencing reactions to be performed simultaneously (Michael, 2009).

The read length of Roche 454, after upgrading to 454 FLX Titanium, reached 700bp with 99.9% accuracy and 14 G data output per run within 24 hours. The second NGS system is AB SOLiD System with 85 bp read length, 99.99% accuracy, and 30 G data output per run within 7 days. Finally, Illumina GA/HiSeq System was improved to HiSeq 2000 with 50SE, 50PE, 101PE read length, 98% accuracy, and 600 G per run within 8 days. HiSeq 2000 is globally adopted system and the cheapest in sequencing with $0.02/million bases (Liu et al., 2012).

There was significant improvement in the NGS technologies, by advances in bioinformatics that allowed for increased data storage, and the analysis of very large data sets, and the simultaneous genome wide measurement of multiple epigenetic modification, in conjunction with the transcriptome and genetic variation of same biological sample (Meaburn and Schulz, 2012). Chromatin immunoprecipitation sequencing (ChIP-Seq) can be used to map chromatin modification and protein binding. Bisulfite sequencing has become very popular tool for mapping the genome wide single base resolution of DNA methylation (Xiong et al., 2011).

2.4.3 Illumina Sequencing

The Illumina sequencing technology is innovative and flexible sequencing system for rapid and accurate large-scale sequencing that enables a wide range of applications in genomics, transcriptomics, and epigenetics. The Illumina sequencing, also called as Solexa sequencing, has adopted sequencing by synthesis approach (*Figure 6*); an acrylamide coating on the surface of glass flow cell is used to immobilize colony amplified DNA template. A single fluorescent labeled reversible-terminator dNTP is used during each sequencing cycle (Illumina, 2010; Quail et al., 2012). To initiate the NGS reaction, 100-200 million contiguously separated template cluster can be produced by solid-phage amplification with free ends to which a universal sequencing primer can hybridize (Michael, 2009). Recently, the Illumina NextSeq sequencer was introduced with a high capacity and low cost (Zoll et al., 2016).

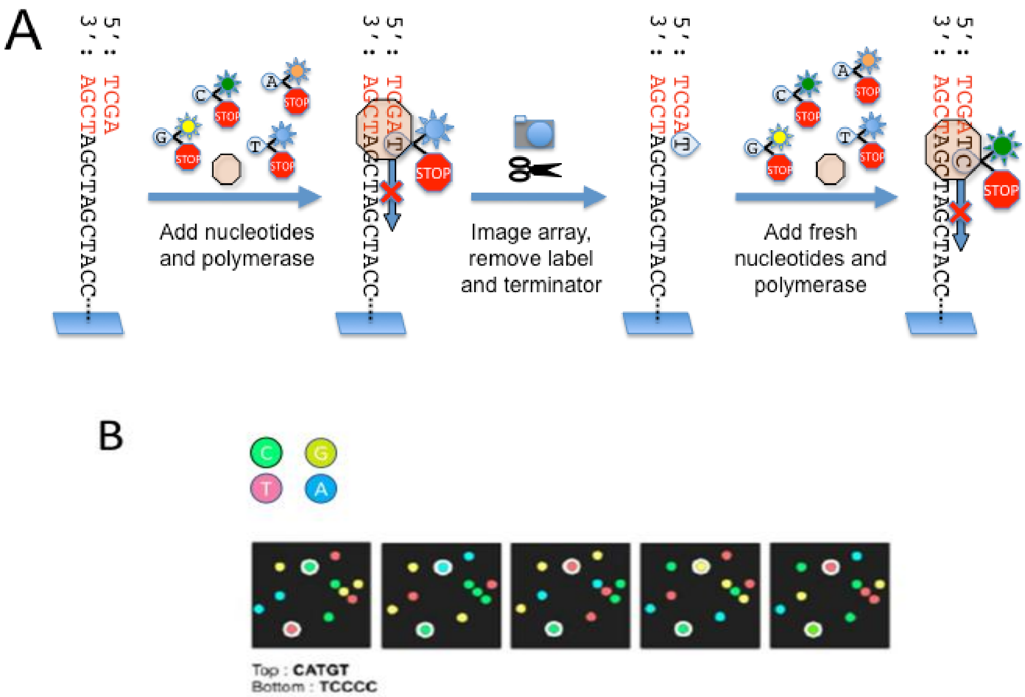

A
5' :: TCGA
3' :: AGCTAGCTAGCTACC...
Add nucleotides and polymerase
Image array, remove label and terminator
Add fresh nucleotides and polymerase
STOP
B
C
G
T
A
Top : CATGT
Bottom : TCCCC

Figure 6. Sequencing by synthesis, Illumina sequencing chemistry. A). A reaction starts with a sequencing primer (red) is annealed to the template sequences linked to the flow cell surface. Then, the addition of DNA polymerase and mixture of fluorescently labeled nucleotides to the flow cell. The nucleotides are designed with a cleavable terminator moiety so that only one nucleotide can be integrated during each cycle of sequencing. After nucleotide integration, fluorescent signals are recorded, and the array is imaged for each cluster. The terminator moiety and fluorescent label are cleaved off and removed, and the next cycle began with addition of fresh nucleotides and polymerase. B). repetition of sequencing cycle in order to determine the sequence of bases in a given fragment. Image courtesy of (Anderson and Schrijver, 2010; Wilantho et al., 2012)

2.5 Bioinformatics tools

2.5.1 CLC Genomic Workbench

CLC Genomic Workbench is a robust tool proposed by scientists for researchers to analyse and visualize next generation sequencing data. Its advanced technology associated with unique features and algorithms that are broadly used by people in industry, researchers and students in academia to overcome challenges associated with data analysis (QIAGEN, 2018). CLC Genomic Workbench comprising set of tools to analyse and compare epigenetic markers. The common problems associated with analysing cytosine methylation data can be solved in one platform, the Bisulfite Sequencing plugin included in CLC Genomic Workbench. CLC Genomic Workbench can share data generated from all high-throughput sequencing platform. It comes with an advantages where, preparation of reference genome, mapping of reads to a known reference, masking, variant calling can be performed in a single platform (QIAGEN, 2017).

2.5.2 Bismark

It has advantage to find a unique alignment by running four alignment process simultaneously. First, bisulfite reads are transformed into a C-to-T and G-to-A version. Second, each of them is aligned to equivalently pre-converted forms of reference genome using four parallel instances of the short-read aligner Bowtie (*Figure 7*). This read mapping enable Bismark to uniquely verify the strand origin of bisulfite read. Finally, Bismark can manage BS-Seq data from both directional and non-directional libraries. Mapping performed in this manner controls partial methylation correctly and unbiased manner, since residual cytosine in the sequencing read are converted *in silico* into a fully bisulfite-converted form before the alignment take place (Krueger and Andrews, 2011).

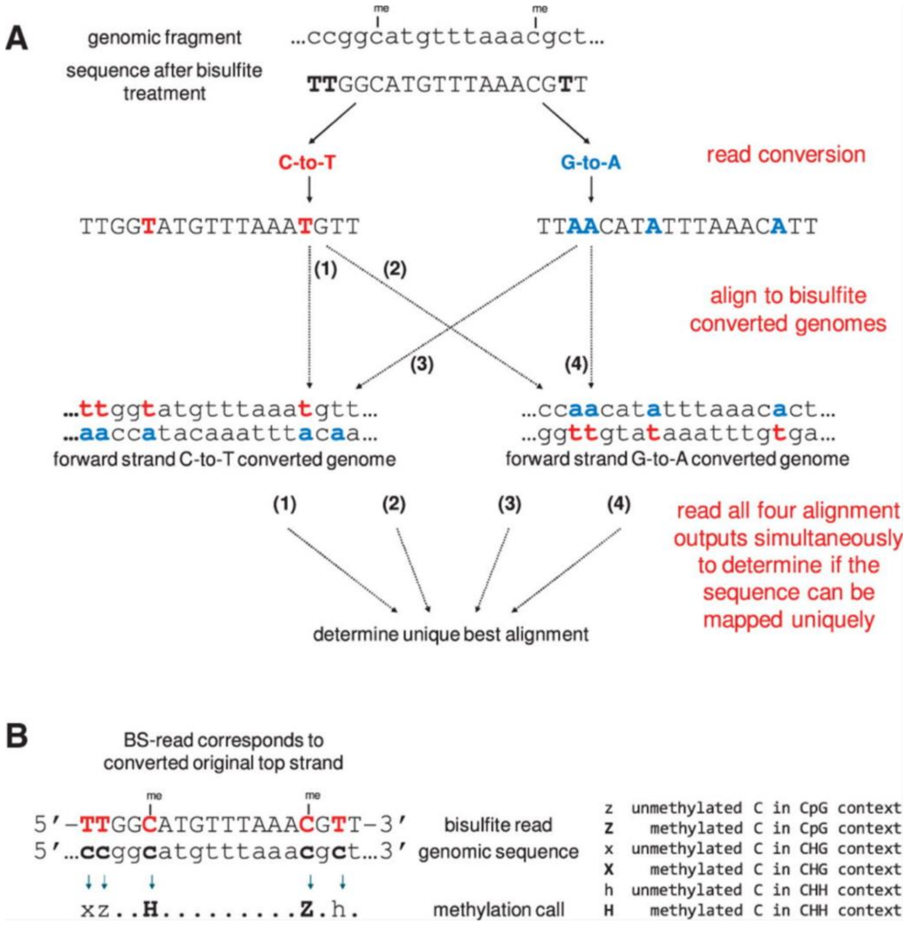

A
genomic fragment
me
me
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment
B
BS-read corresponds to converted original top strand
me
me
5′-TTGGCATGTTTAAACGTT-3′
bisulfite read
5′...ccggcatgtttaaacgct...3′
genomic sequence
xz..H.........Z.h.
methylation call
z unmethylated C in CpG context
Z methylated C in CpG context
x unmethylated C in CHG context
X methylated C in CHG context
h unmethylated C in CHH context
H methylated C in CHH context

Figure 7. Outlook of bisulfite mapping and methylation calling by Bismark. (A) Sequencing reads from a BS-Seq experiment that are converted into a C-to-T and a G-to-A version and are then aligned to similarly converted versions of the reference genome. the best alignment is then determined from the four parallel alignment processes. (B) Determination of the methylation state of positions involve cytosines carried out by comparing the read sequence with the corresponding genomic sequence. Figure taken from (Krueger and Andrews, 2011)

2.5.3 MethylKit

Methylkit, is an R package, facilitated with end to end analysis of RRBS data with comprehensive features and easy to use for detection of differentially methylated CpG sites. It can read DNA methylation information both from a text files as well as alignment files (Akalin et al., 2012b; Stockwell et al., 2014). Methylkit equipped with different feature such as Coverage statistic, methylation statistic, sample correlation and clustering, feature annotation and accessor functions, multiple visualization options, regional and tiling windows analysis, reading methylation calls directly from Bismark (Bowtie/ Bowtie2) alignment files, reading methylation percentage data from generic text files, multithreading support and almost proper documentation (*Figure 8*). And the open-source code can find at https://code.google.com/archive/p/methylkit/. However, methylkit only require methylation score per base for all analyses that can be obtained by two ways, from a text file and SAM format alignments files obtained from Bismark aligner (Akalin et al., 2012b). Firstly, methylkit process the alignment file to obtain % methylation scores, if a SAM file is provided and then reads that information into memory. Furthermore, from SAM files, methylkit users have an option to present methylation information for contexts: CpG, CHG, CHH, since DNA methylation can occur in all these contexts. Read coverage distribution and % methylation can be easily visualized in methylkit as read coverage per base and % methylation per base are the primary information in the methylkit data. If sequencing data suffer from PCR duplication bias can be revealed by the read coverage distribution and % methylation to reveal either the base is high or low methylation. Methylkit is also facilitate with measuring and visualizing similarity between samples by calculating pairwise correlation coefficient between the % methylation. Further visualization can be carried out by plotting scatterplots of the % methylation scores that essential for detecting sample outliers.

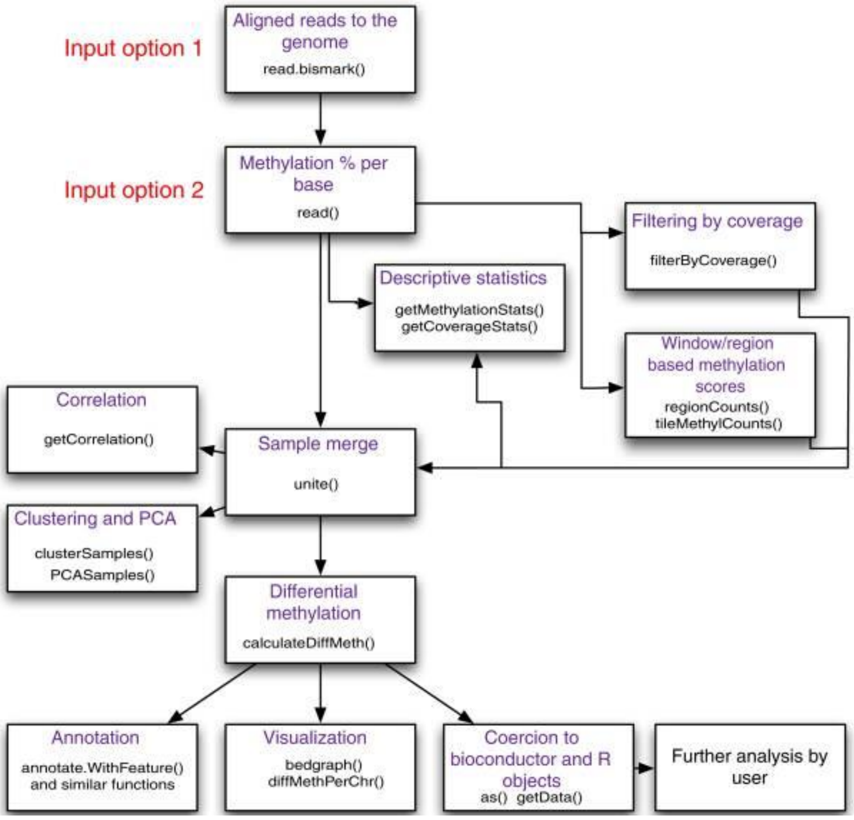


Figure 8. ***Workflow of Methylkit.*** *The possible features of Methylkit as well as the function that could be used for those features are summarized in a flow chart. Figure taken from (Akalin et al., 2012b)*

2.5.4 SeqMonk

SeqMonk is a broadly used bioinformatics tool, developed at Babraham Institute to the visualization and analysis of mapped sequence data where sequence reads are mapped against an annotated genome. SeqMonk work with combination of an annotated genome viewer and a data viewer that allow to visualize high-throughput sequencing datasets (Bioinformatics, 2008). The main features of Seqmonk are; Import of mapped data from mapped data (BAM/SAM/bowtie etc), Creation of data groups for visualisation and analysis, Visualisation of mapped regions against an annotated genome, Flexible quantitation of the mapped data to allow comparisons between data sets, Statistical analysis of data to find regions of interest, Creation of reports containing data and genome annotation (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/).

2.5.5 Blast2GO

Blast2GO is a broad-spectrum bioinformatics tool enable for functional annotation of sequences and data mining on the resulting annotations, principally deployed on the gene ontology (GO) vocabulary that make possible functional analysis to the genomics studies of non-model species. Blast2GO tool supports InterPro, enzyme codes, KEGG pathways, GO direct acyclic graphs (DAGs), and GOSlim as well as its also versatile for genomic research, user friendly and easy to install (Conesa and Götz, 2008). Blast2GO application find homologs to fasta formatted input sequences with the use of Blast and the blast results annotated files are used to map the GO terms. The Blast2GO annotation rule is applied for sequence annotation and the statistics and annotation can be visualized on the GO DAG (Conesa et al., 2005; Gotz et al., 2008). Principally, local or remote Blast searches are used in Blast2GO to find similar sequences to one or several input sequences. By reconstructing the structure of Gene Ontology, GO annotation can be visualized and equivalent GOs mapping gives Enzyme codes which are highlighted on KEGG maps (http://docs.blast2go.com/user-manual/). There are five basic steps; BLASTing, mapping, annotation, statistical analysis and visualization, presented in figure bellow (*Figure 9*).

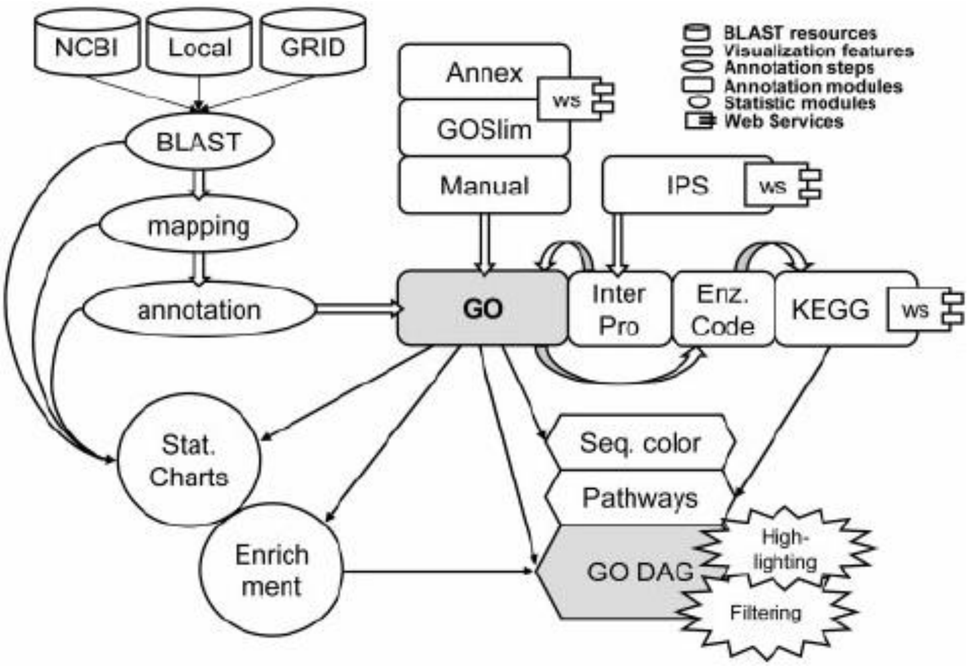


Figure 9. The figure illustrates the schematic representation of the B2G application. Blast, mapping, and annotation: the three step processes through which GO annotation are generated. GO mapping can gives Enzyme code and KEGG pathways map annotations and it also includes highlighting and filtering options in visual tools. Figure taken from (Gotz et al., 2008).

3. Aim of the study

The main goal of the overall project was elucidation of underlying factors influencing fertility in modern, efficient livestock production through novel epigenomics and metabolomics.

The primary aim of my master thesis was to establish an efficient pipeline for bioinformatic analysis of RRBS data from boar and bull semen.

The main aim of the project was achieved through the following sub-goals:

- Compare two different set of bioinformatics tools for analysis of RRBS data, so as to establish the most appropriate tool(s) for future analysis.
- Comparison between different groups of datasets based on the DFI phenotype.
- Identify and annotate differentially methylated regions of interest and perform enrichment analysis.
- Design primer for regions of interest, which could be verified in lab as potential epigenetic markers for semen quality assessment.

4. Material and Methods

An outline of the materials and methods used in the project is described in *Figure 10*. We did not perform the procedure in the box with red outline in the thesis work.

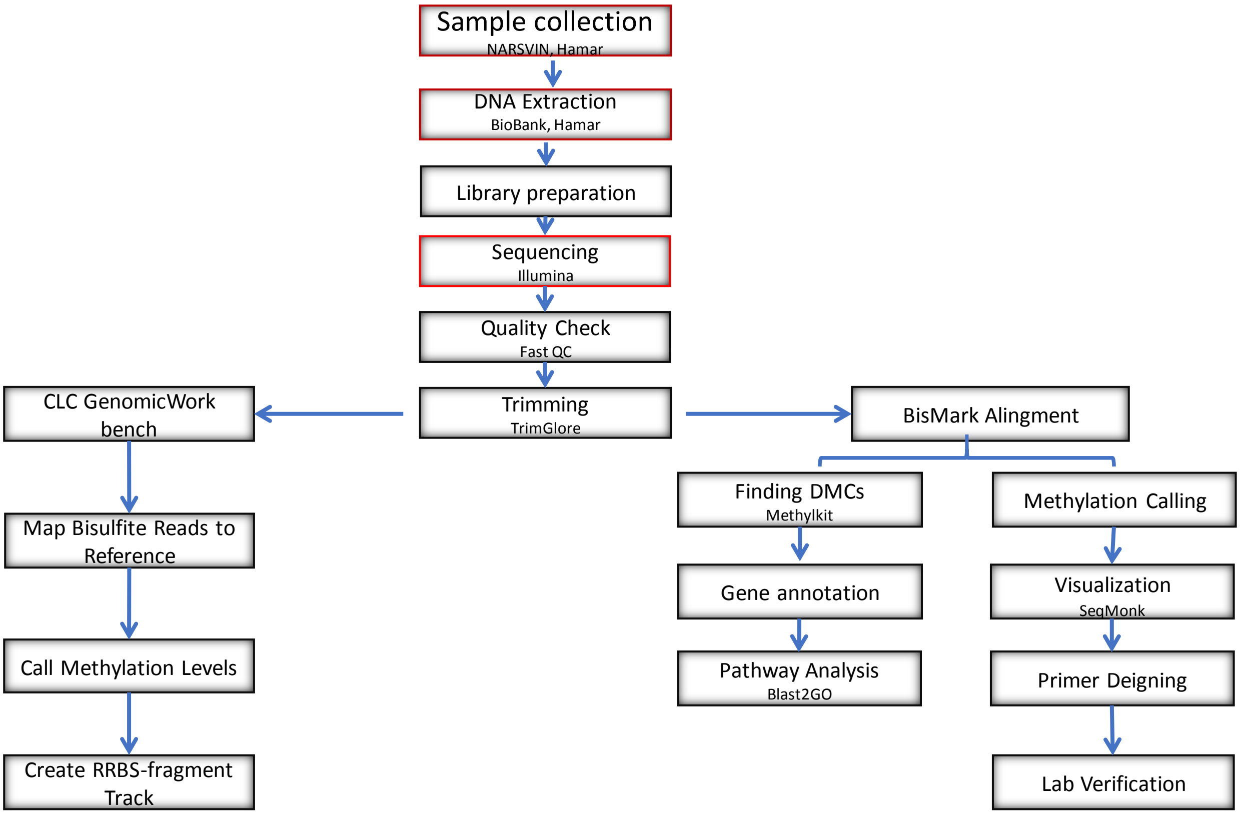


Figure 10. The figure illustrates the brief overview of method used in the project. Where, CLS and Bismark used for reference genome mapping, CLC and Methylkit used to get differential methylation, SeqMonk and Blast2GO were used for visualization and pathway analysis respectively.

4.1 Sample collection and DNA isolation

Semen samples for screening of sperm DNA methylation pattern were collected from a total number of 15 ejaculations taken from nine different individuals of pigs, provided by NORSVIN AS Hamar, Norway, between 2010 and 2013 (**Error! Reference source not found.**). The sperm-rich fraction of ejaculates was collected using the ‘gloved hand’ technique from two different races 2222 is Norwegian Landrace and 6666 is Norwegian Duroc. DNA fragmentation index was analysed in Inland Norway University of applied sciences with the using A Cell Lab Quanta TM SC MPL flow cytometer facilitate with an argon laser with excitation at 488 nm. Genomic DNA was extracted from frozen semen samples using the Maxwell 16 Bench top DNA extraction system in BioBank AS, Hamar. The samples were divided into three groups based on DFI level. The percentage DFI between 0-2 % DFI in low, between 2-10 % DFI in medium, and more than 10 % DFI in high.

The quantification of DNA after library amplification was carried out with the use of PicoGreen method. A minimum DNA concentration of 5.42 ng/ µl was extracted from sample library 738607 and a maximum DNA concentration of 30.45 ng/ µl was retrieved from sample library 738615 (**Error! Reference source not found.**).

Table 1. Sequencing Data from Pig Sperm with % DFI, age and sample collection date. The colour intensity used below corresponds to increasing level of DFI, and is coloured accordingly, High DFI, Medium DFI and Low DFI.

Sample ID	Name	Race	Collection Date	Age	DFI	DNA cons. ng/µl
738618	A	2222	05/03/2013	257	28.39%	26.42
395	B	2222	07/01/2011	523	27.39%	18.83
738619	A	2222	03/04/2013	286	26.05%	18.95
738607	B	2222	10/12/2010	495	21.31%	5.42
393	C	2222	23/08/2010	297	18.48%	13.21
397	D	2222	19/11/2010	472	7.01%	30.56
738616	E	6666	07/06/2010	278	6.81%	64.8
738606	B	2222	26/04/2010	267	2.45%	73.6
738623	F	2222	10/06/2015	287	1.03%	66
738615	G	2222	22/11/2012	340	0.74%	30.45
738610	H	2222	20/08/2010	396	0.63%	7.9
396	H	2222	21/01/2011	550	0.58%	31.88
392	H	2222	12/07/2010	357	0.54%	34.42
738611	H	2222	20/09/2010	427	0.45%	12.74

DOB = Date of birth, 2222 = Norwegian Landrace, and 6666 = Norwegian Duroc.

Bull semen samples were collected from 14 and 17 months old red Norwegian bulls owned by Geno using artificial vagina. Cryopreserved samples after initial sperm quality analyses were subjected to DNA isolation and library preparation.

4.2 RRBS library preparation and sequencing

The purified genomic DNA was digested overnight with *Msp*I and *Taq*$^{\alpha}$I. CG dinucleotide was used to fill the sticky end produced by *Msp*I and *Taq*$^{\alpha}$I digestion, and 3' A overhang was added. The illumine sequencing adapters with 3'T overhangs were ligated to digested DNA, and the ligation products were purified. For RR genome, the size selected 40-220 bp DNA fragments to avoid the formation of adapter dimers, using agarose gel electrophoresis. The size selected DNA was bisulfite-treated with EpiTech ® Bisulfite kit from Qiagen. After bisulfite treatment, the converted DNA was purified using EpiTech spin-columns with 10 µg/ml carrier RNA, which enhance binding of DNA to the column. (Qiagen, 2009a). Library amplification optimization was done with qPCR and, the purified DNA was PCR amplified to enrich for fragment with the adapter. PCR amplified DNA was clean-up with Solid Phase Reversible Immobilization (SPRI) bead, and the final libraries were quantified using the PicoGreen method. For bull samples, we prepared the library according to the protocol of NuGEN (NuGEN, 2017). The construct RRBS libraries were sequenced using platform Illumina NextSeq sequencer at Oslo sequencing center. Eight libraries were sequenced in a single run, and the sequenced single -end reads length was 75 bp.

4.3 Quality control

The RRBS sequencing data from the sequencer first undergo through initial quality control then quality-and adaptor-trimming in order to yields a final set of methylation data (Krueger and Andrews, 2012). Before carrying out the actual read alignment, we subjected all bisulfite converted sequence file to quality and adapter trimming using software Trim Galore (version 0.4.4). Trim Galore works in combination of use of the publicly available adapter trimming tool ‘Cutadapt’ and ‘FastQC’(Bioinformatics, 2013). Trim- Galore has a set of additional parameters for RRBS file which is covered extensively in the trim Galore documentation. The command line used for quality check was*: trim_galore --fastqc --gzip --clip_R1 4 --rrbs --non_directional /path_to_files* by leaving default parameter for adapter (auto-detection of adapter sequence, Phred score: 20,). Trim-Galore was commanded --*gzip* to compress the output file and `clip_R1 4` to remove 4 bp from the 5' end of read 1 (or single-end reads). The `–rrbs` command to specify that the input file was an Msp1 digested RRBS sample. The adapter-trimmed sequence will have a further 2 bp removed from their 3’end to avoid that the filled -in C close to second Msp1 site in a sequence was used for methylation calls. The --*non_directional* option for non-directional RRBS libraries will screen quality-trimmed sequences for CAA or CGA at the start of the read and, if found, removes the first two base pairs as option --rrbs this avoids using cytosine positions that were filled-in during the end-repair step. The detail procedure for quality check can be found on https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ .

Bull sequences were trimmed according to recommended script by manufacture both in Python (v 2.7.5 in Linux) and trim-galore.

4.4 Mapping the reads to a reference genome

After removing adopter, the bisulfite sequencing reads were aligned to the *Sus scrofa* (assembly *Sscrofa* 11.1) reference genome using CLC Genomic Workbench version 11.0 and Bismark (version: v0.19.1) aligner tool.

4.4.1 CLC

Firstly, for all algorithms, the reference genome was indexed. Datasets were imported and mapped to the reference genome using the default parameters (match score: 1, mismatch cost: 2, linear gap cost: insertion: 3, deletion: 3, length fraction: 0.5, similarity fraction: 0.8 and mapping: randomly) which creates read tracks, mapping report and list of unmapped reads. Calling methylation levels for simultaneous detection of differential methylation in two sample types, test (sample with high level DFI) and control (sample with low level DFI). To compare methylation level between the test and control the Fisher exact statistical mode was used in order to get track of methylated cytosines and methylation reports.

4.4.2 Bismark

Bismark aligner tool was used to map the read to reference genome. the reference genome for *Sus scrofa* (assembly Sscrofa11.1) was downloaded from NCBI website, https://www.ncbi.nlm.nih.gov/genome in order to prepare the genome for bisulfite alignment. Firstly, reference genome was prepared using following command line to use in bismark; *bismark_genome_preparation –bowtie2 --verbose /data/genomes,* that create C->T and G->A version of the genome further they were indexed in parallel using the indexer bowtie2-build. Secondly, trimmed files were mapped against indexed reference genome using following command line where, the bismark alignment and methylation calling was carried out by using bowtie2 as default mode, where a multi-seed length of 20bp with 0 mismatches.

bismark -l 20 -n 0 /path to Genome/Sus_scrofa_11_genome/ test_file. Here, Sus_scrofa_11_genome folder contains both C-T and G-A converted genome as well as unconverted genome. Thirdly, the alignment results obtained from bismark were in BAM format, which were converted into SAM format with the use of samtools using command line:

samtools view file.name.bam > file.name.sam

Finally, the SAM files were converted into shorted SAM using command line;

grep -v \'^[[:space:]]\@\' 96_L_bismark_bt2.sam | sort -k3,3 -k4,4n > 96_L_bismark_bt2.sorted.sam.*

4.5 Differential methylation analysis using MethylKit

The methylation percentage calls were calculated from SAM sorted using methylkit (Akalin et al., 2012a). A methylRaw project for CpG methylation was created by using command line: *my.methRaw=processBismarkAln(location=file.list, sample.id =list("test1","test2","test3","ctrl1"," ctrl2","ctrl3"),assembly="Sscrofa11.1", read.context="CpG",nolap = FALSE,mincov = 10,minqual = 20,phred64 = FALSE,treatment = c(1,1,1,0,0,0), save.folder=getwd()).* The basic stats about the methylation data such as coverage and percent methylation were checked using *methylRawList* object which contains methylation information per sample. *getMethylationStats(myobj[[1]], plot = T, both.strands = F.* Similarly, we also plotted the read coverage per base information with command *getCoverageStats(myobj[[1]], plot = T, both.strands = F).* Furthermore, the samples were filtered based on coverage to overcome from PCR bias by discarding bases with very high read coverage (more than99.9th percentile) as well as bases with low read coverage (10X). The command line was; *filtered.myobj <- filterByCoverage(myobj, lo.count = 10, lo.perc = NULL, hi.count = NULL, hi.perc = 99.9).* Further procedure for comparative analysis such as merging, Clustering, sample, PCA analysis. The differential methylated Cs were calculated using function *calculateDiffMeth()* that had q-value 0.01 and percent methylation difference larger than 25%. Additionally, differentially methylated Cs were calculated by comparing methylation levels of low percentage DFI with high and medium percentage DFI samples using logistic regression method on the R package methylkit (Akalin et al., 2012b).

4.6 Visualization

SeqMonk version: 1.41.0 was used to visualize the methylated Cs. The annotated reference sequence of *sus scrofa* was downloaded from Ensemble. As SeqMonk is capable of importing mapping information in a variety of mapping format, Mapping files directly from Bismark and differentially methylated text files were imported by clicking on *File* -> *Import Data* -> *Text (Generic)* ->.

4.7 Primer deigning

We selected 200 bp up and downstream from the target region with no or minimum methylated C in order to design the primer. All ‘G’ residues from up-stream sequence of the target region were converted to ‘A’ and all ‘C’ residues of down-stream sequence from target region were converted to ‘T’ (bisulfite-treated sequences). The primers were designed with the use of the Primer3Plus tool on http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi. The designed primers were blasted against the *S. scrofa* genome.

4.8 KEGG Pathway analysis:

The networking pathway or KEGG pathway analysis is carried out with the help of Blast2GO application. The methylation regions or methylated Cs analysed by both methylkit and CLC Genomic work bench. The gene associated with DMCs/ DMR was obtained from CLC as well as a visualization tool, SeqMonk and fasta sequence of protein of cross ponding genes were easily obtained from https://www.ncbi.nlm.nih.gov/nucleotide/. The fasta sequences of proteins were imported, then blasted, mapped, and annotated respectively. CloudBlast was used to speed up BLASTing with latest NCBI and other standard public databases, while other parameters were remained default. Furthermore, the nearest TSS gene ID list obtained from methylkit was used for enrichment analysis. Enrichment analysis carried out using Over Representation analysis (ORA) method on WebGestalt (http://www.webgestalt.org/option.php). In this KEGG pathway analysis we used Benjamini-Hochberg (BH) for false discovery rate (FDR) and minimum and maximum number of Entrez gene IDs in the categories were 5 and 2000 respectively.

5. Results

5.1 Illumina sequencing datasets

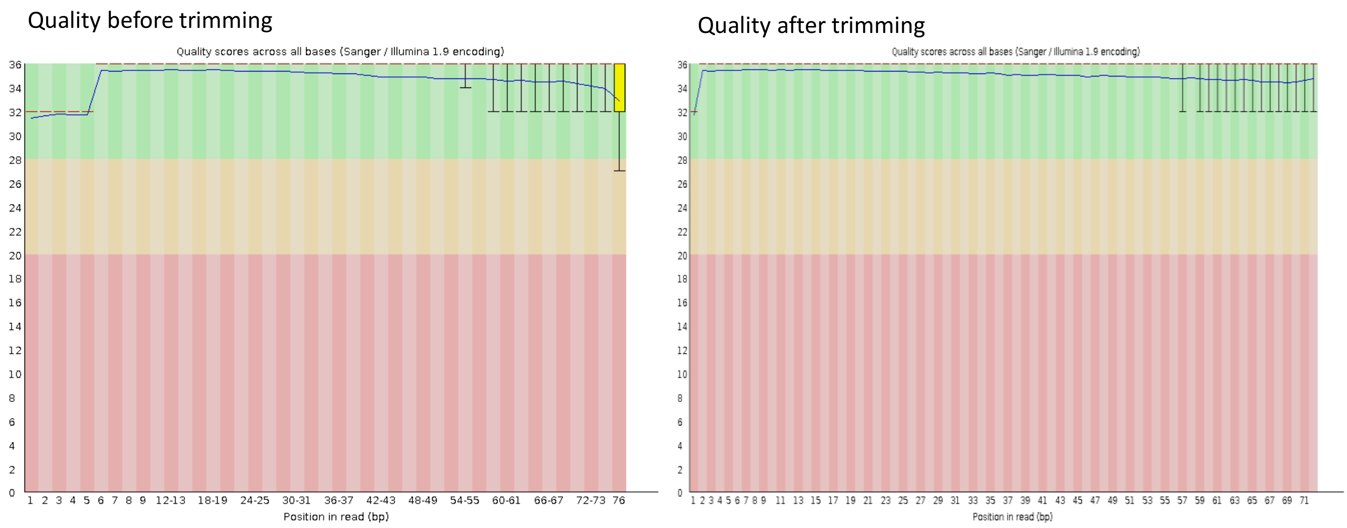
Total 14 sequenced libraries were used in this study. Summary of the sequencing statistics of each library is presented in *Table 2*. The results below show that we generated a minimum and maximum of 11.54 and 21.6 million clean reads corresponding to 11.51 and 20.87 X read coverage respectively.

Table 2. Details of the sequencing datasets of 14 libraries sequenced using Illumina sequencing platform with the number of total reads and clean reads in base pairs, and overall coverage of the reads.

Sample Id	Total Reads (in bp)	Clean Reads (in bp)	Read coverage
738618	17655261	17645947	17.69
395	21075288	20855183	20.11
738607	16767296	16670788	15.94
738619	11580556	11547471	11.51
393	17139869	16928292	16.3
738610	13374736	13313871	13.54
396	21720429	21600109	20.87
738611	16442959	16372832	15.69
392	19490143	19164300	18.44
738615	20489386	20426319	19.75
738623	15442944	15404026	14.56
738616	18530860	18482740	17.86
738602	16983491	16982954	16.84
397	15460504	15025860	14.44

5.2 Quality control reports

In the FastQC results shown below (*Figure 11*), the quality scores are shown on the y-axis and the blue line in the graph shows the mean quality of the reads. The poor quality of proportion of the reads was efficiently removed and the quality phred scores was above 32 (*Figure 11*).


Quality before trimming

Quality after trimming

Figure 11. Per base sequence quality. The figures show an outline of the range of quality values for all bases at each position in the fastQ file before and after trimming.

Adapter trimming was performed to cut out any signs of adapter contamination (*Figure 12*). As, the figure below shows there was a significant effect of adapter contamination on the sequenced reads, especially towards the 3´end of the reads.

Before adapter removal **After adapter removal**

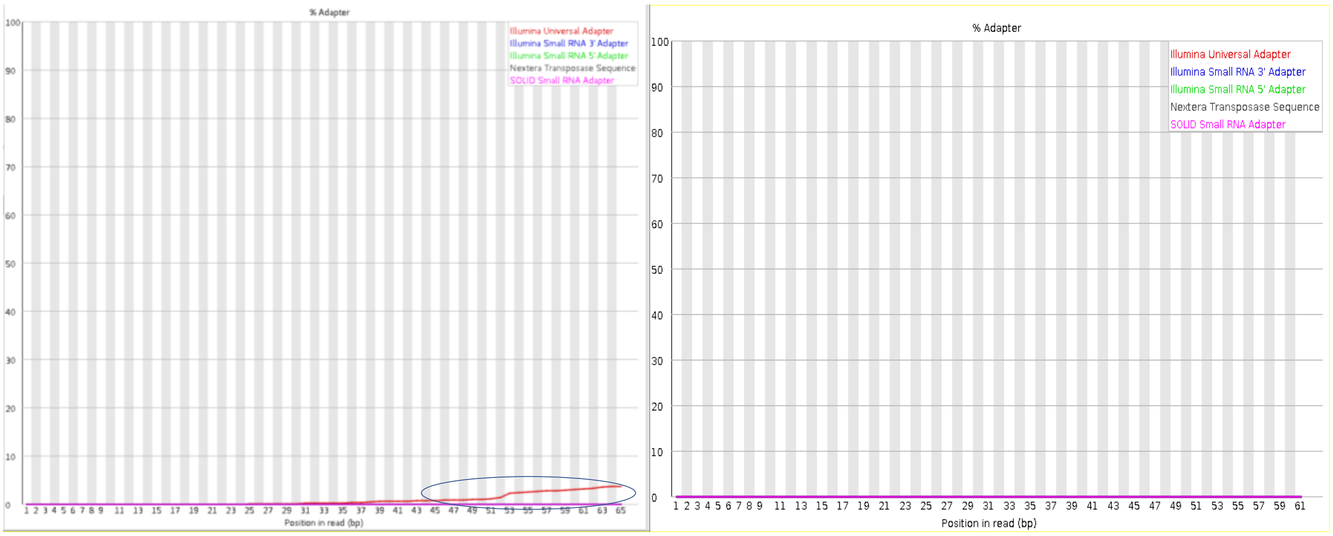


Figure 12. The effect of adapter trimming on RRBS data. The adapter contamination before trimming marked with blue circle, which was completely removed after adapter trimming.

The quality of sequence content across all bases was improved by removing base pairs from the 5' end using --*clip_R1* on Trim Galore to avoid unwanted bias. The sequence length was reduced from 76 bp to 72 bp after clipping as indicated in the graph (*Figure 13*).

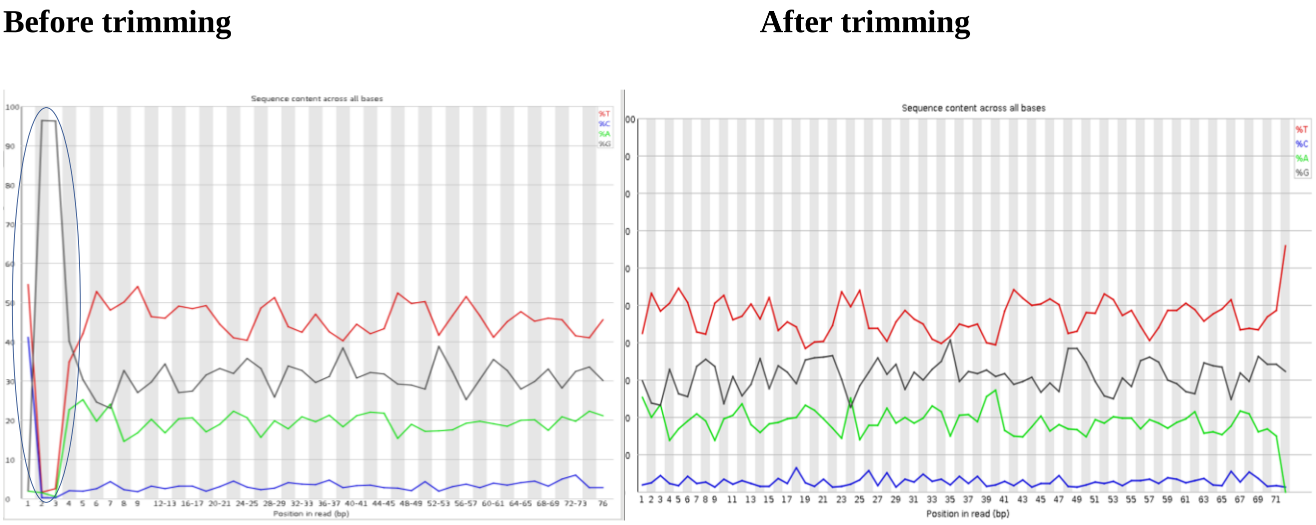

Before trimming **After trimming**

Figure 13. Quality of sequence content across all bases before and after trimming. Unwanted bias at the 5'end in blue circle was removed after trimming.

5.3 Pig genome analysis

We calculated the *Msp*I digested RRBS genome size using the CLC Genomic work bench. The total genome size of *Sus scrofa* is 2501.91 Mb (Assembly version 11.1), and the number of CpGs in the genome was 30,619,972. The RRBS genome based on *Msp*I digestion (size range 40-220 bp) was calculated to be 79.58 Mb, which was 3.01 % of the total genome. The total number of fragments of RR genome was 675,240.

5.4 Mapping and analysis of RRBS reads from pig's sperm libraries

A total of 14 libraries were constructed with 40-220 bp insert fragments from different Boar sperm samples. We generated DNA methylation data for 2.7 million to 5.2 million CpG dinucleotide in libraries (*Table 3*). The mapping efficiency with Bismark ranged from 42.3% -63.6% with an average 55.09%. We discarded data with lower than 10x mean CpG coverage, while we obtained minimum

15.49 to 36.31-fold mean CpG coverage. Moreover, the average bisulfite conversion rate was 99.48 %, ranging between 98.87 % and 99.91 % (*Table 3*). In order to establish an efficient pipeline for methylation analysis, we used two alignment tools CLC Genomic workbench and Bismark. The mapping efficiency for CLC was minimum 82.82 % and maximum 88.31%. The percentage of methylated Cs in the CpG context shows a similar pattern for three samples 393, 396 and 392. Notably, the percentage of methylated Cs were nearly doubled in CLC than Bismark for samples 738618, 738607, 738610 and 738611 (*Table 3*).

Table 3. Summary of reduced representation bisulfite sequencing experiments and mapping efficiency of Bismark and CLC tools. MC = methylated C

Sample Id	**Mapping efficiency % (Bismark)**	**CpG covered**	**MCs in CpG% (Bismark)**	**Mean CpG coverage**	**Bisulfite conversion rate %**	**Mapped rate % (clc)**	**MCs in CpG% (clc)**
738618	58.9	3653419	15.5	31.87	98.96	82.85	30.19
395	48.6	3559468	29.7	20.62	99.87	84.69	39.07
738607	61.2	2950999	7.9	36.31	99.82	87.58	22.7
738619	63.6	3427518	48.2	34.26	98.88	82.82	20.64
393	42.3	3421901	35.6	25.66	99.87	84.5	38.41
738610	58.8	2786476	9.8	31.02	99.91	87.25	22.84
396	57	4483066	41.5	25.38	98.87	83.7	43.94
738611	57.7	3040914	10.1	33.59	99.9	88.31	27.75
392	46.7	3926715	42.2	28.25	98.93	85.1	41.7
738615	60.2	3820639	30.5	27.2	99.86	85.67	42.82
738623	59.1	5269572	37.9	15.49	99.92	……	……..
738616	57.3	3127619	8.2	34.92	99.92	……	…….
738602	53.2	3157431	18.8	27.12	99.91	…….	…….
397	46.7	3238911	37.5	22.38	99.86	…….	…….

6. Differential Methylation analysis using %DFI

Further analysis of trimmed reads was carried out by dividing into three groups based on DFI level (**Error! Reference source not found.**).

- In the first group, five samples with the lowest DFI as control and five samples with the highest DFI as a test (Ctrl1 = 738610, ctrl2 = 738611, ctrl3 = 392, ctrl4= 738615, ctrl5 =396 and Test1= 393, Test2 = 738607, Test3 =738618, Test4 =738619, Test5 =395).
- In the second group, three samples with low DFI as control and three samples with high DFI as a test (Ctrl1 = 738611, ctrl2 = 738615, ctrl3= 738623 and Test1= 393, Test2 = 395, Test3 =738618).
- In the third group, three samples with low DFI as control and three samples with medium DFI as a test (Ctrl1 = 738611, ctrl2 = 738615, ctrl3= 738623 and Test1= 397, Test2 = 738602, Test3 =738616).

6.1 High and Low (10 samples)

6.1.1 Finding the differential methylated Cs

The differentially methylated Cs were analysed using methylkit, in ten samples (5 test and 5 controls). Different methylkit parameters were used, including selecting DMCs to 20 % and 25% of methylation difference, and having DMCs to be present in 3,4 or 5 test samples. Table 4 shows that increasing numbers of DMCs were observed in less restricted conditions. A total of 17 DMCs were identified with 25% of methylation difference and present in all five samples. For detail see supplementary file 1. Further down-stream analysis was carried out using DMCs obtained from at least 25% methylation difference in a minimum of four samples.

Table 4. Number of differentially methylated Cs with different set of parameters.

Conditions	**25% DM in 5 samples**	**25% DM in 4 samples**	**25% DM in 3 samples**	**20% DM in 3 samples**
Number of DMCs	17	2322	5546	7280

6.1.2 Pearson’s correlation and clustering

The Pearson’s correlation was carried out for all ten samples of high and low % DFI. The analysis revealed that there was close correlation between samples in term of CpG methylation levels. The most similar distribution of CpG methylation was found between Ctrl3, Ctrl4, Ctrl5 and Test3, Ctrl3, Ctrl4 (Pearson's correlation scores: 0.94-0.95). While least similar was found between Ctrl1vs Test1, Test2, Test3, Test4 and Test5 (Pearson's correlation: 0.78-0.81) (*Figure 14*A). Cluster analysis based on CpG methylation pattern showed Test 1 and 2 and Ctrl 1 and 2 clustering together with maximum height whereas, Test 3, 4 and 5 and Ctrl 3, 4 and 5 clustered separately. However, CpG methylation clustering was not coherent with the DNA fragmentation level and there were no separate clusters observed between low and high %DFI samples (*Figure 14*B).

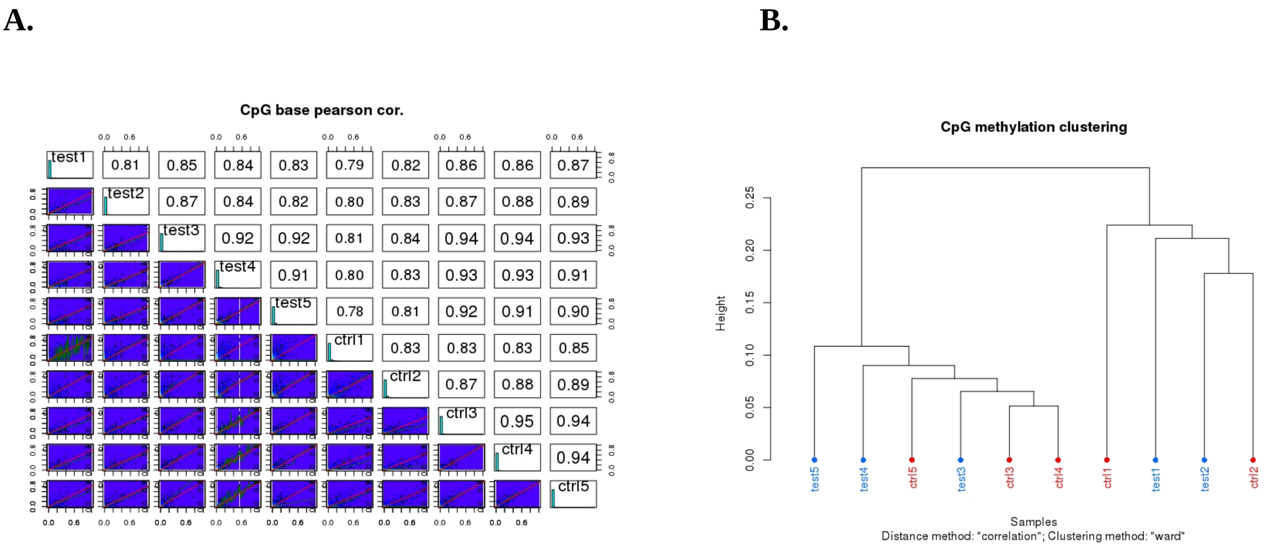
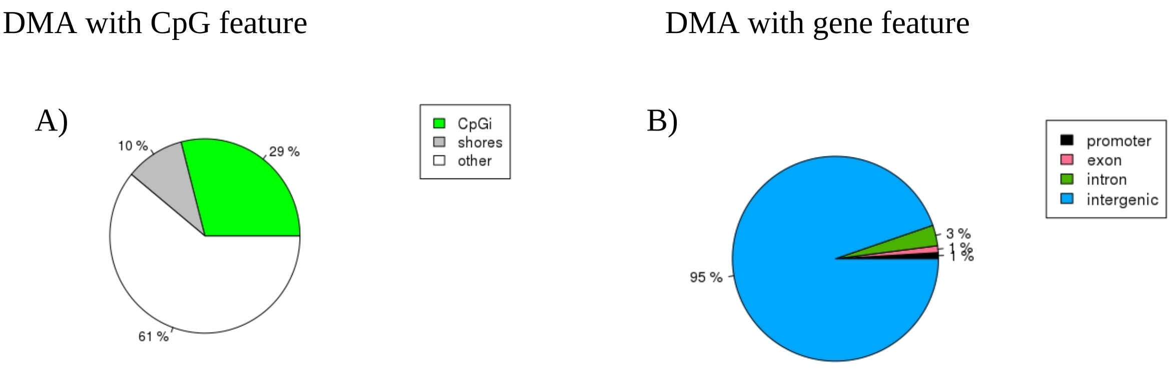


Figure 14. Sample correlation and clustering of low and high %DFI samples. A) Correlation analysis of the CpG methylation patterns among the different samples. Scatter plot of methylation percentage value in ten samples B) Cluster analysis based on methylation levels across the different samples. Distance between samples based on their methylation pattern was estimated by the ward’s minimum variance method.

6.1.3 Regional methylation analyses

The RRBS data from boar sperm cells was analysed to determine the genomic regions elucidating distribution of DMCs for CpG sites annotated as both CpG islands (CpGi) and shores. The data displayed that nearly one-third of DMCs (29%) were present on CpGi and 10% on CpG shores. Remaining 61% DMCs were present on non-CpG regions (*Figure 15*A). The detailed analysis identified the regional distribution of DMCs present across the genome. In this analysis, most abundant DMCs were present at intergenic region (95%), followed by 3%, 1% and 1% at intron, promoter, and exon respectively (*Figure 15*B).



DMA with CpG feature

DMA with gene feature

Figure 15. Proportion of differential methylation on CpG feature and gene feature. A) The pie chart demonstrates the relative proportion of DMCs annotated to CpG islands (bright green), CpG shores (Grey) and regions beyond CpG shores (White). B) The pie chart illustrates the proportions of DMCs annotated to promotor region (Black), exon (magenta), intron (Green) and intergenic region (Blue).

The number of hypo and hyper methylation event per chromosome analysis showed that all chromosomes were mostly hypomethylated with very few hypermethylated Cs. However, chromosome X and Y were totally hypomethylated and hypermethylated respectively (*Figure 16*).

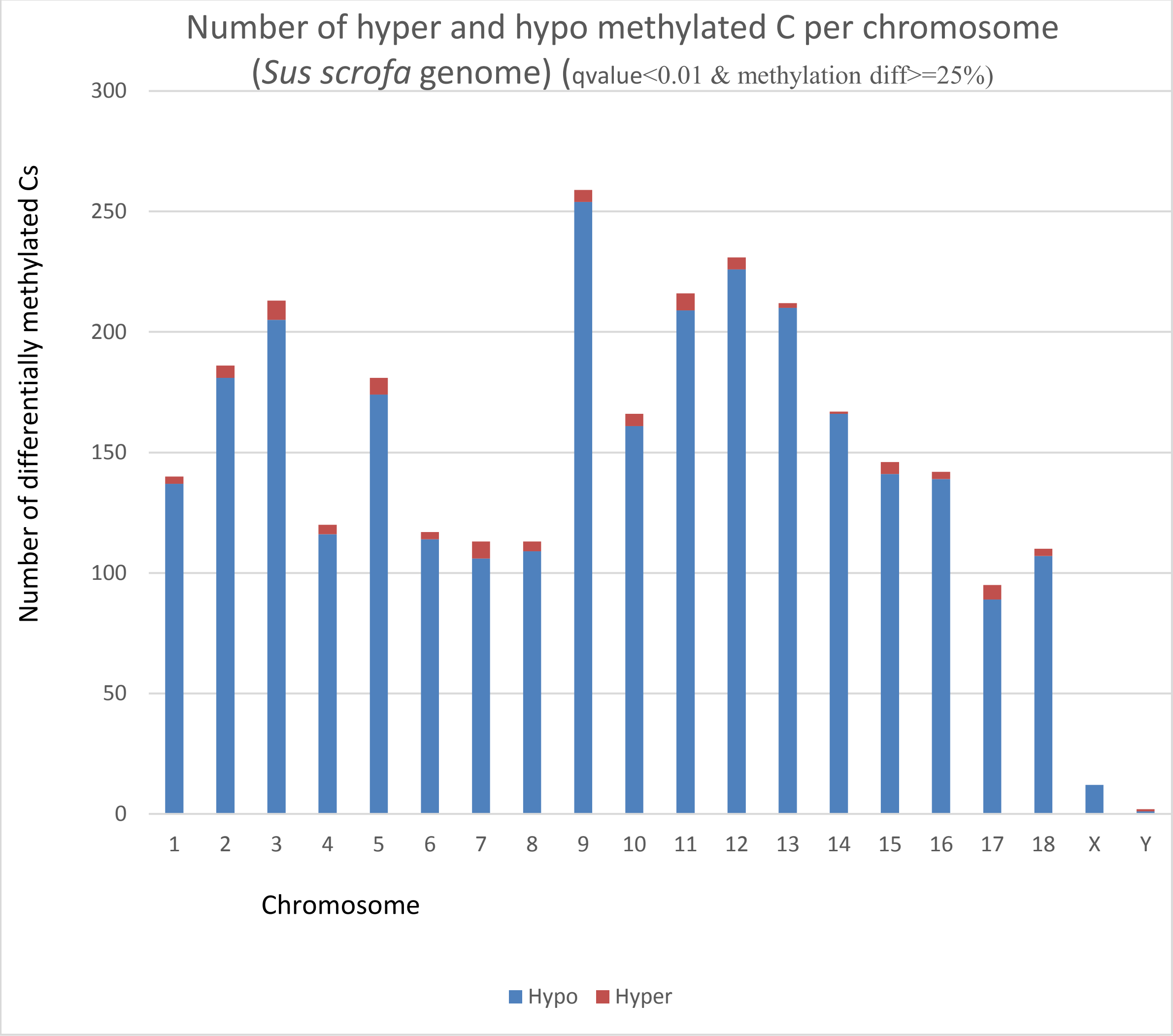


Figure 16. Visualizing differential methylation events (10 samples). The bar graph shows the number of hypo and hyper methylation events per chromosome. A percent of sites with the minimum 10X coverage, methylated CpGs with at least a 25% difference and q-value <0.01, present in how many samples.

6.2 Low vs high and low vs medium (6 samples)

6.2.1 Finding the differential methylated Cs

The differentially methylated Cs were analysed using methylkit in two group (low vs. high, and low vs. medium), six samples in each group. The methylkit parameter was used selecting DMCs to 25% methylation difference and having DMCs to be present in all three samples. We obtained 132 and 189 DMCs in low vs. high and low vs. medium group respectively (supplementary file 2).

6.2.2 Pearson’s correlation and clustering

The Pearson’s correlation was carried out for six samples of low vs. high and low vs. medium % DFI. The analysis revealed that there was similar correlation with in samples in term of CpG methylation levels. The CpG methylation was similar among both scenarios (low vs. high and low vs. medium). The correlation between CpG methylation of different percentage DFI samples was high, and the correlation efficiency was between 0.87 to 0.96 in both group analysis (supplementary file 11). Nevertheless, CpG methylation was not coherent with the DNA fragmentation level, and there were no separate clusters visualised in both groups between low vs. high and between low vs. medium.

6.2.3 Regional methylation profiling

The regional profiling of 6 samples reveals that the majority of DMCs were on intergenic in both cases. High level of methylated Cs on intron and low level of DMCs on promoter region in low vs. high compared to low vs. medium. On the other hand, 5% more DMCs on CpG island in low vs. medium and CpG -shores showed similar results (13%) in both conditions (supplementary file 11).

The number of hypo and hyper methylation event per chromosome analysis showed all chromosomes were mostly hypomethylated with very few hypermethylated Cs. Whereas, chromosome Y was completely hypermethylated in both groups (supplementary file 11).

We also Compare hypo and hyper methylation analysis among two test group medium and high DFI level. Our findings revealed the increased pattern of hypermethylated Cs in high DFI level test group (*Figure 17*).

Total number of DMCs= 190 Total number of DMCs= 136

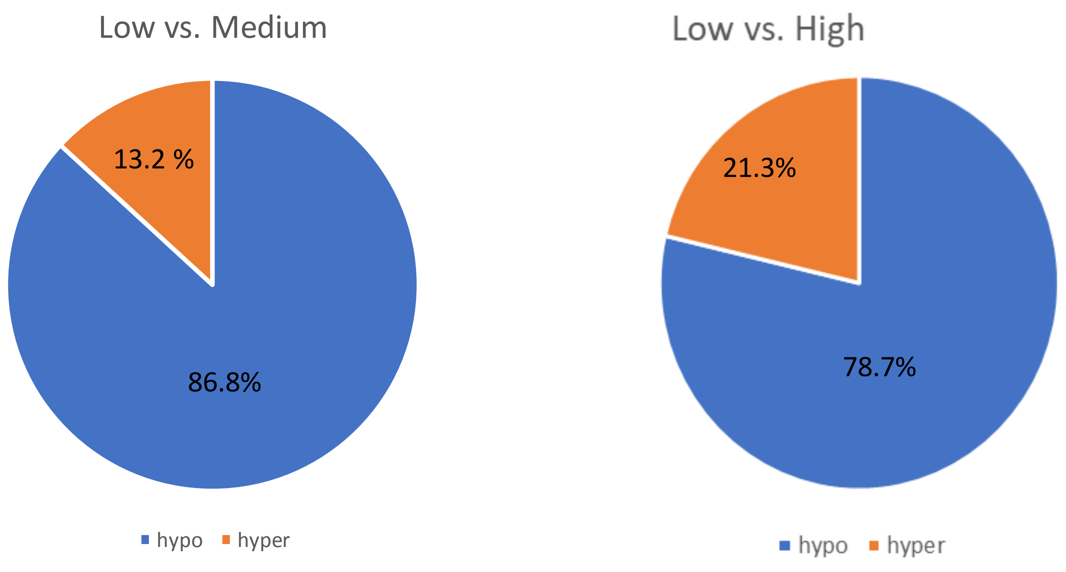


Figure 17. Hypo and hyper DMCs distribution among two test group (High and Medium).

6.3 CLC Results (10 samples)

Furthermore, data from these 10 semen samples was analysed on CLC Genomic Workbench using Fisher exact statistic mode and 7,661 DMRs were identified (supplementary file 3). The DMCs results obtained from methylkit by restricting percentage DM 25 in 4 samples were compared with results from CLC DMRs present in minimum 4 samples, and only 18 DMCs were found to be common between two tools (Table 5).

Table 5. *Details about the common DMCs between CLC and methylkit.*

Chr	Position	strand	pvalue	qvalue	meth.diff
1	274271140	+	7.88E-06	0.0010662	29.60
11	35877089	-	0.0000787	0.0070804	25.63
3	1126590	-	4.02E-22	8.354E-19	46.07
4	121182458	-	1.04E-10	4.682E-08	-34.08
4	121182480	-	1.42E-08	4.071E-06	-29.91
4	121182483	-	2.51E-10	1.042E-07	-28.13
5	40034537	+	2.88E-11	1.463E-08	-42.55

5	40159468	+	2.33E-14	1.971E-11	-44.11
5	40182420	-	0.0000269	0.0029647	-26.39
5	40194444	-	1.1E-07	0.00002527	-33.20
5	40194446	-	0.0000102	0.0013291	-25.43
7	2504153	-	2.72E-20	4.828E-17	45.81
7	2504258	+	3.29E-13	2.357E-10	32.94
7	8960560	-	1.92E-09	6.597E-07	33.72
7	8960583	-	1.92E-07	0.00004188	27.53
9	37285827	-	7.06E-21	1.33E-17	59.62
9	63391332	-	1.89E-08	5.275E-06	-30.63
9	63391360	-	5.05E-06	0.0007282	-25.13

6.4 Validation of bioinformatics results

We selected 12 heavily differentially methylated CpGi to confirm the reliability of the bioinformatics results on methylkit by bisulfite sequencing. The designed primers were blasted against the *S. scrofa* genome and there were no complementary sequences found. The example sequence showed in *Figure 18*. The sequences in the red box are flanking region (*Figure 18*). The detailed information about primer sequences and obtained DNA band can be seen in supplementary file 10.

Chr 3 + -, Region of interest contains 16 Cs with Differentially methylated range from 25 – 46 %

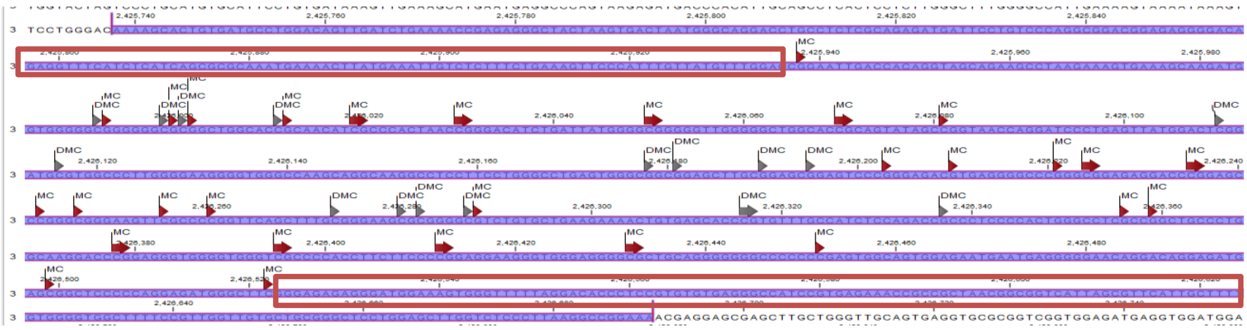

3 TCCTGGGAC
MC
DMC
ACGAGGAGCGAGCTTGCTGGGTTGCAGTGAGGTGCGCGGTCGGTGGAGATGAGGTGGATGGA

Figure 18. The fragment for primer designing to lab verification. The red arrow indicates differentially methylated Cs and gray indicates methylated Cs.

6.5 Pathway Analysis

The corresponding genes to DMCs were obtained by analysing the nearest transcription start site (TSS) from DMCs and their feature name from all three groups of analysis.

In low vs. high (10 samples) group, we obtained 1962 DMCs associated with 521 genes. To investigate the pathway categories of the enzyme-coding genes associated to nearest TSS of DMCs, we performed KEGG pathway analysis with Blast2GO. A total of 57 DMCs were found to be associated with enzyme-coding genes, which are involved in 78 KEGG networking pathway. KEGG pathway analysis revealed these DMCs related genes were mostly involved in Purine Metabolism, Thiamine metabolism, Inositol phosphate metabolism, Drug metabolism- other enzymes and other 33 metabolisms related pathway (supplementary). KEGG pathway analysis revealed that the majority of DMCs associated genes were involved in the pathways related to biosynthesis, degradation and signalling systems. The graphical representation of pathways shown in supplementary file 5.

In low vs. high (6 samples) and low vs. medium (6 samples) groups, 17 and 21 KEEG networking pathways were identified respectively (supplementary file 6 A, B & C). Interestingly, five networking pathways were common between all three groups of analysis. In addition, there were seven pathways common between low vs. high (10 samples) and low vs. high (6 samples). Also,15 pathways were common between low vs. high (10 samples), and low vs. medium (6 samples) were identified (*Table 6*).

Table 6. KEGG pathway of enzyme coding genes associated with DMCs common between Low vs. High, Low vs. High, Low vs. Medium and Medium vs. High (10 samples, 6 samples, and 6 samples respectively).

Common	**Pathway**	**Enzyme codes**
Common in all groups	Purine Metabolism	3.6.1.15, 3.6.1.3, 3.1.5.1, 3.6.1.8
	Thiamine metabolism	3.6.1.15
	Drug metabolism- cytochrome P450	1.14.13.8,2.5.1.18, 1.14.14.1
	N -Glycan biosynthesis	3.2.1.113, 2.4.1.256
	Pyrimidine metabolism	1.3.1.2, 3.6.1.8
	Phosphatidylinositol signaling system	2.7.1.149,3.1.4.11, 2.7.1.137,2.7.7.41, 2.7.1.153

Common in Low vs. High (10 samples), and Low vs. High (6sample)	Biosynthesis of antibiotics	1.14.13.39, 1.4.3.3,2.6.1.13, 1.1.1.27, 1.1.1.35, 1.1.1.2, 3.1.1.31, 2.7.1.2, 2.7.1.1, 4.2.1.2
	Inositol phosphate metabolism	2.7.1.149,3.1.4.11, 2.7.1.137, 2.7.1.153
	Pantothenate and CoA biosynthesis	4.1.1.36, 1.3.1.2
	Pyruvate metabolism	1.1.1.27, 4.2.1.2
	Methane metabolism	1.1.1.284
	Citrate cycle (TCA cycle)	4.2.1.2
Common in Low vs. High (10samples), and Low vs. Medium (6sample)	Metabolism of xenobiotics by cytochrome P450	2.5.1.18, 1.14.14.1
	Tyrosine metabolism	1.11.1.8, 2.1.1.28
	Beta-Alanine metabolism	4.1.1.15, 2.6.1.19, 1.3.1.2
	Fatty acid degradation	1.1.1.35, 1.3.3.6, 1.14.14.1
	Drug metabolism- other enzymes	3.1.1.1, 2.5.1.18, 1.3.1.2
	Phenylpropanoid biosynthesis	1.11.1.7
	Glutathione metabolism	2.5.1.18
	Steroid hormone biosynthesis	1.14.14.1
	Amino sugar and nucleotide sugar metabolism	3.2.1.14, 2.7.1.2, 2.7.1.1, 2.7.1.7, 2.7.1.4
	Retinol metabolism	1.14.14.1
	Porphyrin and chlorophyll metabolism	4.2.1.75
	Tryptophan metabolism	1.1.1.35, 1.14.14.1
	Caffeine metabolism	1.14.14.1
	Linoleic acid metabolism	1.14.14.1
	Arachidonic acid metabolism	1.14.14.1

6.5.1 Methylated Cs present in promoter region

RRBS method mainly emphasises on CpG-rich regions such as promoter and CpG island. Methylation in the promoter plays a vital role in the regulation of gene expression. Hence, we focus on DMCs of 2 kb upstream and downstream of TSS (promoter region), in low vs. high (10 samples). There were 32 DMCs found in the promoter region. These DMCs were visualized with SeqMonk, and we found that the majority of DMCs were present in CpG-shore region, whereas 5 and 6 DMCs were found in CpGi and non-CpG regions respectively. These 32 DMCs were found to be associated with 16 genes (*Table 7*).

Table 7. Differentially methylated Cs present in promoter region.

chr	position	meth.diff	dist.to.feature	feature.name	Asso.Gene	GpGi	CpG Shore	Non-CpG
chr1	107660892	-34.85	1385	NM_001244997	PCLAF		x	
chr1	107660972	-36.94	1465	NM_001244997			x	
chr12	34168637	-31.98	-907	NM_001243919	CUEDC1		x	
chr12	22650666	-29.53	-798	NM_001123164	PNMT			
chr12	22650652	-36.27	-784	NM_001123164				x
chr12	22650643	-33.06	-775	NM_001123164				x
chr12	22650599	-31.28	-731	NM_001123164				x
chr12	53883547	-35.55	383	NM_001244077	CCDC42		x	
chr12	53883537	-37.02	393	NM_001244077			x	
chr12	53883498	-25.6	432	NM_001244077			x	
chr12	53883498	-29.07	432	NM_001244077			x	
chr13	206885478	44.83	930	NM_001143700	AGPAT3	x		
chr15	139558450	-25.69	680	NM_001044529	CAPN10			x
chr17	48081534	-25.07	-946	NM_001243629	CTSA	x		
chr17	48081543	-29.29	-937	NM_001243629		x		
chr17	48081545	-28.9	-935	NM_001243629		x		
chr17	48081551	-33.53	-929	NM_001243629		x		
chr18	40533770	-25.05	-1512	NM_001252232	FKBP9		x	
chr3	1568085	-25.06	670	NM_001244568	NUDT1		x	
chr4	51213198	-29.12	1617	NR_128554	MIR9858		x	
chr4	51213211	-28.63	1630	NR_128554			x	
chr6	83607517	-25.25	-1832	NM_001244280	CATSPER4			x
chr6	55648482	-32.47	95	NM_001195344	KLK7			x
chr6	87949759	-26.95	950	NM_001099931	FABP3		x	

chr6	87949732	-29.22	977	NM_001099931		x
chr6	166529006	-26.55	1658	NM_001032356	TCTEX1D4	x
chr6	166529061	-42.16	1713	NM_001032356		x
chr6	166529084	-36.78	1736	NM_001032356		x
chr6	166529092	-26.85	1744	NM_001032356		x
chr7	46657932	-26.3	-1442	NM_001243379	GSTA4	x
chr7	46657890	-31.46	-1400	NM_001243379		x
chr7	29661061	-25.8	1315	NM_001184893	RGL2	x

6.5.2 Methylated Cs present in CpG region

This study identified 601 DMCs between low and high DFI (10 samples) located on CpGi. Of these, 590 (98%) were hypomethylated, and 11 (2%) were hypermethylated (supplementary file 9). There were 7 and 85 genes close to TSS of hypermethylated and hypomethylated Cs respectively. Of these 92 genes, three genes were found to be involved in spermatogenesis and reproduction (Table 8).

Table 8. Details of genes involved in spermatogenesis and reproduction.

Chr. No.	No. of DMCs	Genes
chr17	4	CTSA
chr12	1	SPATA20
chr11	2	FGF9

We also analysed heavily methylated CpG island (*Table 9*). The molecular function of these genes was analysed with the use of UniProt on https://www.uniprot.org/uniprot. Interestingly, we had found most of the genes were involved in binding activity, majority of them were in DNA binding (KLF5, HOPX, IRF4, TERT), three were in ion binding (LMO4, AMZ1, and BIRC5), two in transfer activity (SLC7A1, GLRX3) and one in flavin adenine dinucleotide binding (FMO1).

Table 9. 10 most DMRs present in CpG.

Chr	feature.name	Asso.Gene
chr11	NM_001097489	KLF5
chr11	NM_001012613	SLC7A1
chr12	NM_214141	BIRC5
chr14	NM_001243896	GLRX3
chr16	NM_001244300	TERT

chr3	NM_001195325	AMZ1
chr4	NM_001112686	LMO4
chr7	NM_001253352	IRF4
chr8	NM_213792	HOPX
chr9	NM_214064	FMO1

6.5.3 The enrichment analysis

Low vs high (10 samples)

In the 10 samples analysis (low vs. high), the methylated Cs nearest to TSS list contains 496 gene IDs of which 304 gene IDs were unambiguously mapped to unique Entrez gene IDs and 191 IDs were mapped to multiple Entrez Gene IDs or could not be mapped to any Entrez Gene ID. The GO Slim summary is based upon the 304 unique Entrez Gene IDs. Among the 304 unique Entrez Gene IDs, 163 IDs were annotated to the selected functional categories and, also in the reference gene list (*S. scrofa*), which was used for the enrichment analysis.

Enrichment analysis to describe property of gene and their products was evaluated by the WebGestalt WEB-based gene set analysis toolkit. The genes were annotated to 49 functional categories, including, 12 biological process, 20 cellular component and 17 molecular functions. In terms of biological process, it was observed that most of the differentially methylated genes were involved in the functional groups of metabolic process and biological regulation, 122 and 110 respectively. While, 20 genes were found to be associated with reproduction. For cellular component, the differentially methylated genes were mostly present in membrane and nucleus, 88 and 62 respectively. For molecular function, the most frequent category was protein binding (75), ion binding (42), and nucleic acid binding (36) (*Figure 19*).

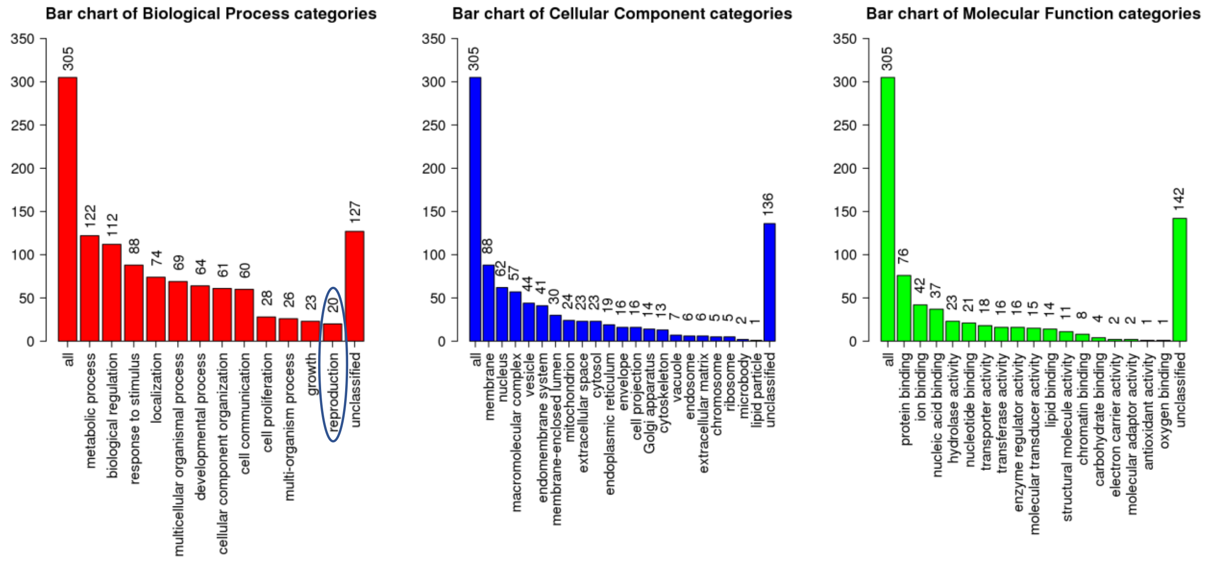


Figure 19. GO Slim summary for the gene IDs that were unambiguously mapped to unique Entrez gene IDs (10 samples). Each biological process, cellular component and molecular function category is represented by a red, blue and green bar respectively. The height of the bar represents the number of genes observed in the category.

In the enrichment genes category, we identified nearly one third of genes were involved in metabolic pathway and majority of gene were associated with different signalling pathway such as TGF-beta signalling, AGE-RAGE signalling in diabetic complications, signalling pathways regulating pluripotency of stem cells and cAMP signalling pathway (*Table 10*). The detailed information of the enriched categories is listed in supplementary file 7.

Table 10. Summary of the enriched categories (10 samples). The table contains the lists of the enriched categories, number of Entrez genes in interested gene list and also in the categories and FDR.

ID	Name - *Sus scrofa* (pig)	Gene	FDR
ssc05012	Parkinson's disease	12	2.10E-02
ssc00190	Oxidative phosphorylation	10	9.44E-02
ssc05016	Huntington's disease	12	9.44E-02
ssc04350	TGF-beta signaling pathway	7	1.11E-01
ssc04933	AGE-RAGE signaling pathway in diabetic complications	8	1.11E-01
ssc04550	Signaling pathways regulating pluripotency of stem cells	9	1.19E-01
ssc05010	Alzheimer's disease	10	1.74E-01
ssc01100	Metabolic pathways	40	1.74E-01
ssc04142	Lysosome	8	1.74E-01

ssc04024	cAMP signaling pathway	11	1.74E-01

Low vs high (6 samples)

The methylated Cs nearest to TSS list contains 60 gene IDs in which 37 gene IDs were unambiguously mapped to unique Entrez gene IDs and 37 gene IDs were mapped to multiple Entrez gene IDs or could not mapped to any Entrez gene IDs. The GO slim summary are based upon 37 unique Entrez gene IDs. Among the 37 unique Entrez gene IDs, 25 IDs were annoteted to the selected functional catagories and also in the reference gene list, which are used to enrichment analysis.

The maximum numbers of differentially methylated Cs related genes were identified in the functional groups of metabolic process and biological regulation 13 and 11 respectively out of 37 whiles, only one gene was found to be associated with reproduction in biological process categories. In cellular component categories, most of the genes identified associated with membrane and macromolecular complex (15 and 5 respectively). Moreover, binding functional groups such as protein binding, ion binding, nucleotide binding shared a higher number of genes, in molecular function categories (supplementary file11**Error! Reference source not found.**).

In the enrichment category analysis, we identified four genes were involved in the cAMP signalling pathway and the majority of the gene were associated with different infectious diseases such as Staphylococcus aureus infection, leishmaniasis, HTLV-I infection, and hepatitis C (*Table 11*). The detailed information of the enriched categories is listed in supplementary file 8.

Table 11. Summary of the enriched categories low vs high (6 samples). The table contains the lists of the enriched categories, number of Entrez genes in interested gene list and also in the categories and FDR.

ID	Name - *Sus scrofa* (pig)	Gene	FDR
ssc04024	cAMP signaling pathway	4	0.82
ssc04670	Leukocyte transendothelial migration	3	0.82
ssc05030	Cocaine addiction	2	0.82
ssc05150	Staphylococcus aureus infection	2	0.82
ssc05140	Leishmaniasis	2	1
ssc04610	Complement and coagulation cascades	2	1
ssc05166	HTLV-I infection	3	1

ssc00770	Pantothenate and CoA biosynthesis	1	1
ssc04080	Neuroactive ligand-receptor interaction	3	1
ssc05160	Hepatitis C	2	1

Low vs. medium(6 samples)

The methylated Cs nearest to TSS list contains 102 gene IDs in which 67 gene IDs were unambiguously mapped to unique Entrez gene IDs and 35 gene IDs were mapped to multiple Entrez gene IDs or could not be mapped to any Entrez gene IDs. The GO slim summary based upon 67 unique Entrez gene IDs. Among the 67 unique Entrez gene IDs, 40 IDs were annotated to the selected functional categories and also in the reference gene list, which are used to enrichment analysis.

The maximum numbers of differentially methylated Cs related genes were identified in the functional groups of metabolic process and biological regulation 23 and 21 respectively out of 67 while, two gene were found to be associated with reproduction in biological process categories. In cellular component categories, most of the genes identified associated with membrane and nucleus (17 and 13 respectively). Moreover, binding functional groups such as protein binding, nucleic acid binding, and nucleotide binding shared a higher number of genes, in molecular function categories (supplementary file 11).

In the enrichment category analysis, we identified most of the genes were involved in different signalling pathways such as AGE-RAGE signalling pathway, and Hippo signalling pathway. Then in metabolisms like tyrosine metabolism and oxidative phosphorylation (*Table 12*). The detailed information of the enriched categories is listed in supplementary file 8.

Table 12. Summary of the enriched categories low vs medium (6 samples). The table contains the lists of the enriched categories, number of Entrez genes in interested gene list and also in the categories and FDR.

ID	**Name - *Sus scrofa* (pig)**	**Gene**	**FDR**
ssc05215	Prostate cancer	3	1.00E+00
ssc00350	Tyrosine metabolism	2	1.00E+00
ssc04933	AGE-RAGE signaling pathway in diabetic complications	3	1.00E+00
ssc05030	Cocaine addiction	2	1.00E+00
ssc00520	Amino sugar and nucleotide sugar metabolism	2	1.00E+00
ssc00190	Oxidative phosphorylation	3	1.00E+00

ssc00750	Vitamin B6 metabolism	1	1.00E+00
ssc04390	Hippo signaling pathway	3	1.00E+00
ssc05012	Parkinson's disease	3	1.00E+00
ssc05210	Colorectal cance	2	1.00E+00

There were 20 DMCs, which were identified to be common between low vs medium and low vs high (6 samples). Six of these DMCs were present in CpGi and CpG-shore (*Table 13*). Further, the enrichment categories analysis of these common DMCs showed three genes were involved in metabolic process, two in Alzheimer’s disease and other different eight enriched categories shares remaining genes (supplementary file 15).

Table 13 Common DMCs between low vs. medium and low vs. high (6 samples).

Chr	**position**	**meth.diff**	**Gene overlap**	**CpGi**	**CpG shore**	**CpG shelf**	**Non CpG region**
1	691280	-30.64	x	x			
1	174395679	-34.62					x
10	15089172	-33.17					
10	24967232	-29.38	x				x
10	24967245	-27.73	x				x
12	42298045	-45.41	x				x
13	197528565	-35.14					x
13	204576682	-34.75					x
14	5977080	52.16					x
15	139317558	-27.63		x			
15	139317594	-25.44		x			
17	12618750	-26.11					x
18	51311076	40.25	x		x		
5	40193124	-40.08					x
5	39908450	-29.25					x
5	39927482	-28.08					x
5	39908458	-26.19					x
6	168489151	-32.06					x
Y	40208807	31.53			x		
Y	40209331	27.63			x		

7. Bull datasets

RRBS sequencing data from the bull semen sample were trimmed with the use of trim-galore, then mapped to the *Bos taurus* (assembly 3.11). We obtained maximum 31.1 mapping efficiency (*Table 14*).

Table 14. Mapping report of bull's sperm sequencing data.

Sample ID	Mapping efficiency	total sequence	unique best hit	Percentage of Methylated Cs in CpG
1129187	30.1	26829279	8077298	42.3
1129189	9.7	16697176	1611443	45.8
1129191	26.1	20173518	5269989	38.7
1129193	31.1	19851091	6169187	45.6
1129188	25.5	20980163	5343666	36
1129190	25.3	19902375	5043307	37.4
1129192	26.4	20499976	5409239	38.6
1129180	23.6	22431108	5293518	40.2
1129182	23.7	22065066	5236640	37.9
1129184	28.4	19902460	5659445	43.1

8. Discussion

The thesis work was started with the aim of performing a bioinformatic analysis of epigenetic data from bull and boar semen samples. RRBS as a cost-efficient method has become the method of choice for analysis of methylation profile of targeted region. It has led to an increase in the demand for well-established bioinformatic tools to facilitate subsequent data analysis. Here we were able to establish a quick and efficient pipeline to analyze RRBS sequence data with the help of different available bioinformatics tools.

Quality check and trimming

Quality control is crucial for high throughput experiments to work as expected. In our work, we obtained good quality of sequencing data after quality control with the use of FastQC (*Figure 11*). The higher the score the better the base call (Andrews, 2010) and we achieved quality phred score more than 32, above 20 is evaluate as good (*Figure 11*). Additionally, our base calls fell into green background in the graph which indicates good quality (Andrews, 2010). The adapter contamination may lead to misalignment and incorrect methylation calls because of too many mismatches in the alignment process. If the read does not align properly, a lower mapping efficiency arises as a consequence of adapter contamination (Bioinformatics, 2013). This was eliminated by adapter trimming as shown in *Figure 12*.

Reference reduced representation (RR) genome (*S. scrofa*) analysis

The percentage of RR genome is higher by 0.41% and whole genome size lowered by 0.3 Gb compared to the previous finding by *Choi et al.* in RRBS analysis of pig genome using five different tissue (Choi et al., 2015b) . This might be due to a newer genome assembly version as the *Choi et al.* results were based on assembly 10.2, while our results are based on assembly 11.1.

CpG coverage

We prepared RRBS library for pig sperm samples with the use of *Msp1* digestion enzyme combined with another methyl insensitive restriction endonuclease *TaqαI*. The more accurate analysis of the average methylation level can be achieved by double enzyme digestion which increases the CpG coverage of genomic regions (Wang et al., 2013). Further improvement in genome coverage was also observed in double digestion (Choi et al., 2015b).

The read coverage is an important parameter when working with NGS data. The reason for having a relatively high average read is to resemble long adjacent reads accurately and to make sure accuracy of the final sequence. In this study, we were able to obtain read coverage between 11.51 to 20.87 X, and at least 10X coverage was used to call methylation difference at single CG site level. The recommended coverage range for DMR identification using WBGS has been reported to be between 5x to 10x (Ziller et al., 2015). We used the CG site with more than 10X coverage to calculate the methylation level and even analyze at the region level (CpGi, promoter, exon and intron). However, 5x coverage had been used for methylation analysis at region level in cattle sperm. Previously, even lower than 5x coverage had been used before for the methylation analysis at the region level (Zhou et al., 2018).

Bisulfite conversion

The conversion rate of unmethylated Cs to Ts after bisulfite treatment has significant effect on analysis of methylation of given sequencing reads. The conversion rate ideally should be as close to 100% as possible for a high-quality experiment. More than 99.5% of conversion rate is the ideal value for good experiment (Wreczycka et al., 2017), and we were able to meet this typical value (*Table 3*). We were able to obtain quite high percentage of average bisulfite conversion rate (99.48 %) as compared to the experiments carried out by Schachtschneider *et. al*. (94.27%)S on six tissue samples from pig (Schachtschneider et al., 2015). Similarly, bisulfite conversion rate of > 99% was observed in RRBS analysis of bovine somatic tissue (Zhou et al., 2016). Also, similar bisulfite conversion rate has been observed in studies done on clinical samples and mouse embryonic stem cells where the conversion rate was >99% (Gu et al., 2010b). Therefore, using our bisulfite conversion protocol, we were able to achieve similar conversion rate to previously published paper.

Mapping efficiency

In this study, the RRBS data were processed and aligned against the complete Pig genome using the Bismark and CLC Genomics Workbench and compared the mapping performance of both the tools. Bismark aligns the reads to reference genome using bowtie 2. Bowtie2 has become a method of choice with an advantage of speed, accuracy, sensitivity, quality value awareness, and can align in both local and end-to-end modes. The alignment report reveals a difference in mapping efficiency, and a relatively high mapping efficiency was obtained with CLC > 82 % (*Table 3*) as compared to Bismark. The high mapping efficiency from CLC might be due to algorithm used in CLC. The mapping efficiency obtained from Bismark was on average 55 % (*Table 3*), and the similar mapped rate was obtained from the previous study on pig methylome analysis of different tissues using BS-seeker (Choi et al., 2015b).

One of the primary goals of bisulfite sequencing is to investigate differential methylation. We used two approaches to identify differentially methylated Cs, CLC Genomics Workbench and methylKit. We only found 18 DMCs common between these two methods. Moreover, methylation patterns were totally different between these two approaches. We decided to carry on further investigation using methylkit because many previous works on analysing methylation were done by methylkit (Legendre et al., 2015)but not that many with CLC.

Pearson's correlation study of all three categories of analysis revealed that the correlation between samples within the same group (test or control) and between different groups (control and test) was high (> 0.8). We did not find any previous work done on methylation analysis of level of percentage DFI to compare our findings.

Methylation analysis

Overall, methylation patterns on CpG were very similar between high and low percentage DFI sperm samples (27.38% and 26.82% in test and control respectively). The results of this study are not in agreement with the results of previous studies that have shown the significant lower mean level of global sperm DNA methylation in men with severe DNA fragmentation (Montjean et al., 2015). Similarly, in recent studies on patients with reproductive failures, a negative correlation had been found between global methylated Cs level and DNA fragmentation but positive correlation with healthy men groups (Olszewska et al., 2017). Our finding did not show consistent % DFI effect on methylation which might be due to the difference in methods from previous work or possibly less samples in each group in the present study. Similarly, no association between sperm methylation and DFI was found in previous study on human (Benchaib et al., 2005). However, our findings revealed that hyper methylated DMCs were increased by increasing DFI level (*Figure 17*). The results of this study are in agreement with the results of previous studies that have shown the hypermethylation pattern of DNA related to poor sperm parameters, idiopathic male infertility, and even in pregnancy failure (Benchaib et al., 2005; Pacheco et al., 2011).

The number of hypo and hyper methylation events per chromosome revealed that most of the DMCs were hypomethylated (nearly 87%). Increased CpG hypomethylation has been observed in the low motile sperm in human (Pacheco et al., 2011).

Regional analysis of DMCs

The pattern of differentially methylated Cs on CpG context shows higher proportion of DMCs on CpGi than CpG shore in 10 samples analysis group (low vs high) (*Figure 15*A). However, more DMCs on CpG shore than CpGi were identified in both low vs. medium and low vs. high (6 samples) groups (supplementary file 11). This uneven pattern of DMCs distribution might be due to the different number of samples used.

The annotation of differentially methylated Cs on CpG context shows about three times lower DMCs on CpG shores than CpGi in 10 samples analysis (low vs high). However, 6 samples analysis (low vs high) shows a higher distribution of DMCs on CpG shores, and similar distribution of DMCs was observed in 6 samples analysis (low vs medium). This uneven pattern of DMCs distribution might be due to the different number of samples used.

Moreover, we compared DMCs results from two test groups namely, medium DFI level and high DFI level. The distribution of DMCs in high DFI level test group was increased by 19% and 3% in intron and exon respectively while 2% less DMCs identified on promoter. Furthermore, 6 % less DMCs were identified on CpGi in high DFI level test group (Supplementary file 11).

The annotation of differentially methylated Cs on gene feature shows higher number of DMCs in exon and intron than the promoter region. Similar pattern of distribution on gene feature was observed in both 10 samples group (*Figure 15*B) and 6 sample group (low vs high) (supplementary file 11). There were nearly five times more DMCs identified in exon and intron than promotor. Similar DMCs distribution was identified between promoter and gene body (intron and exon) in low vs medium percentage DFI analysis. The DMCs were mainly concentrated in the intergenic regions (>90%) with only small proportion distributed in the gene body and promoter in 10 sample group (low vs high). Similar pattern of DMCs distribution was observed in all three condition of analysis (10 samples low vs high, 6 samples low vs high, and 6 sample low vs medium).

Primer deigning

Validation of bioinformatics results of DNA methylation for a candidate region of interest using PCR based methods is the popular approach. An advantage of this approach is a detailed analysis of a specific region of the genome with a lower burden of false discoveries(Hernández et al., 2013). In this study, to validate bioinformatics results, we designed primer for ten regions of interest. Some primers work successfully, and we had got an expected band (supplementary file 10) whereas, for other primers method optimization is needed.

Article review for genes involved in spermatogenesis and reproduction

Further analysis of DMCs on promoter (*Table 7*), in low vs. high (10 samples) group, revealed two hypomethylated Cs to nearest TSS related genes associated with reproduction namely, *Ccdc42* and *CATSPER4.* The gene *Ccdc42*, which has been described to play an important role for proper sperm development and male fertility in mouse. It was also observed that mutation in Ccdc42 is associated with malformation of the mouse sperm flagella (Pasek et al., 2016). In human, *Catsper4* promotes sperm hyperactive motility thereby facilitating the entry of sperm into the zona pellucida. The gene was expressed primarily in testes and rete testes, and the gene transcript was also detected in ejaculated sperm. The gene was also found to be conserved among various species (Song et al., 2011). Additionally, infertility has been seen in *Catsper3* and *Catsper4* knockout male mice due to lack of hyperactivated sperm motility (Jin et al., 2007).

The analysis of DMCs in CpGi revealed three genes involved in spermatogenesis and reproduction (Table 8). CTSA (Cathepsins A) identified on CpG island on promoter has grabbed increasing attention for pig performance traits and as potential marker for meat quality (Balatsky et al., 2016). Moreover, the genes with importance in spermatogenesis and reproduction were obtained the analysis of methylated Cs in CpGi (Table 8), in low vs. high (10 samples) group. SPATA20 (Spermatogenesis-associated protein 20) has been found to be expressed abundantly in human testis, hence proposing possible roles of the gene in spermatogenesis (Bonilla and Xu, 2008). FGF9 is expressed in gonads of both sexes in mice prior to sex determination. It has been identified to play crucial role in male development (Kim et al., 2006) , testicular embryogenesis, sex determination, and development of reproductive system in many species (Colvin et al., 2001).

In low vs. medium (6 samples) group, the identified gene ‘SHH’, corresponding to nearest TSS of DMC, has important role in embryonic development. The protein Sonic Hedgehog produced from the gene ‘SHH’ is an essential chemical signal for embryonic development (Seppala et al., 2017). It also play roles in cell growth, development of brain, teeth, long bone, and many other parts of the body as well as normal shaping of the body (Lee et al., 2016).

In low vs. high (6 samples) group, the gene identified ‘ATM’, corresponding to nearest of DMC has important role in spermatogenesis. In human, ATM (ataxia telangiectasia mutated) is essential for normal spermatogenesis. It has been demonstrated that the functional variant of ATM gene rs 189037 to contributes an increased risk of idiopathic nonobstructive azoospermia (INOA) (Li et al., 2013).

Notably, our study identified hypermethylated Cs on CpG-shore on chromosome Y associated with '*SRY*' gene in both low vs. high and low vs. medium (6 samples) categories analysis. In mammalian species, *SRY* gene is believed to be the principal factor in sex determination (Yang et al., 1993). This result incorporated with transcriptomic analysis and pig outlet results from the same semen sample may be helpful to evaluate an epigenetic effect on sex determination.

Pathway analysis

In the pathway analysis, the majority of DMCs associated genes were involved in the metabolic process (supplementary file 4) (**Error! Reference source not found.**). Low vs. high (10 samples) and low vs. medium (6 samples) showed 10 and 3 genes involved in oxidative phosphorylation respectively (*Table 10* and *Table 12*). The sperm cells have distinguished metabolic and signalling pathways in specific regions of the cell composed to function in a localized fashion (Travis et al., 2001). For instance, oxidative phosphorylation and glycolysis provide energy to sperm motility organized in fibrous sheath of flagellum. Further, different signalling pathways have been involved in hyperactivation (Suarez and Ho, 2003) and capacitation (Visconti et al., 2002) of sperm. Such as the expression of hyperactivity and capacitation have been associated with increased intracellular cAMP (Ho and Suarez, 2001; Visconti et al., 2002). Furthermore, two genes identified to be involved in tyrosine metabolism in low vs. medium (6 samples) (*Table 12*). A cAMP- dependent protein kinase appear to play regulatory role in tyrosine phosphorylation and capacitation (Visconti et al., 1995).

The enrichment categories analysis of low vs. high (10 samples) showed seven genes involved in the TGF-beta signalling pathway. TGF- β, originating in the male seminal vesicle, is a crucial male- female signalling agent that regulates the female immune response to seminal fluid at coitus (Sharkey et al., 2012). TGF- β also helps in pregnancy stability via inducing expression of two important cervical cell cytokines GM-CSF (Robertson, 2007) and IL6 (Prins et al., 2012).

The enrichment categories analysis of low vs. high (10 samples) showed nine genes involved in the Signalling pathways regulating pluripotency of stem cells (*Table 10*). In the previous study, the expression level of genes like SMAD4 and SOX2 have been associated to cell signalling to maintain pluripotency in porcine epiblast (Hall et al., 2009).

The analysis of RRBS sequencing data from bull semen sample showed very low mapping efficiency. We also mapped this sequencing data with some other reference genomes like pig, rat and human in order to find the cause of low mapping efficiency. Nearly 30% mapping efficiency was obtained with rat genome (data not shown). Then we looked at length of reads and found very short reads (up to 8 bp) in our sequencing files. Finally, based on BioAnalyzer image from Norwegian Sequencing Center we got the information that there were issues with size selection during library preparation and are in the process of preparing the library again.

9. Conclusion

The current study revealed that sperm cells with different level of DNA damage could have different DNA methylation signatures. In this study, we identified those differences using two different pipelines namely, CLC and Bismark/methylkit. The Bismark/methylkit based method and further downstream analyses provided more authentic and reliable results that can be compared with previous studies and we were able to amplify some of the regions of interest that contained differentially methylated Cs using PCR. Downstream analyses revealed important pathways involved in some important biological phenomena such as reproduction and developmental process.

The results obtained from this thesis could be correlated with other sperm quality parameters like motility and morphology. However, DNA methylation analysis is an expensive and relatively time-consuming method. Furthermore, the bioinformatic pipeline and the parameters used have an impact on the results of the final analyses. Another issue regarding sperm DNA analyses is the sperm population, as a single ejaculation could have different sperm populations and even each single sperm cell might have different genome because of crossing over during spermatogenesis.

In order to have a better understanding of sperm DNA methylation this study in combination with the ongoing work on metabolomics etc. could be used to get a better picture for identification of potential epigenetics markers for semen quality assessment.

10. Future perspective

Further studies on a large number of severe DNA damage sperm samples which can gives more reliable results to develop diagnostic kits for the selection of good quality sperm.

Separation of different population of sperm cells (for example motile and non -motile) and perform the DNA methylation analyses.

Expression study of the corresponding genes during embryo development in offspring. Transcriptomics analysis so as to co-relate the methylation data with gene expression levels.

In addition, further sequencing of amplified PCR product and comparing the methylation status of region of interest with illumine results need to be considered.

11. References

Agarwal, A., and T.M. Said. 2003. Role of sperm chromatin abnormalities and DNA damage in male infertility. *Human Reproduction Update*. 9:331-345.

Akalin, A., M. Kormaksson, S. Li, F.E. Garrett-Bakelman, M.E. Figueroa, A. Melnick, and C.E. Mason. 2012a. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*. 13:R87.

Akalin, A., M. Kormaksson, S. Li, F.E. Garrett-Bakelman, M.E. Figueroa, A. Melnick, and C.E. Mason. 2012b. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*. 13:R87-R87.

Alberts B, J.A., Lewis J, et al. 2002. Molecular Biology of the Cell, New York: Garland Science;.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. 2008. Molecular Biology of the Cell: DVD-ROM. *Garland Science*.

Anderson, M.W., and I. Schrijver. 2010. Next Generation DNA Sequencing and the Future of Genomic Medicine. *Genes*. 1:38-69.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data.

Balatsky, V., I. Bankovska, R.N. Pena, A. Saienko, T. Buslyk, S. Korinnyi, and O. Doran. 2016. Polymorphisms of the porcine cathepsins, growth hormone-releasing hormone and leptin receptor genes and their association with meat quality traits in Ukrainian Large White breed. *Molecular Biology Reports*. 43:517-526.

Bassols, A., C. Costa, D. Eckersall, J. Osada, J. Sabrià, and J. Tibau. 2014. The pig as an animal model for human pathologies: A proteomics perspective.

Benchaib, M., V. Braun, D. Ressnikof, J. Lornage, P. Durand, A. Niveleau, and J.F. Guérin. 2005. Influence of global sperm DNA methylation on IVF results. *Human Reproduction*. 20:768-773.

Berkowitz, G.S., M.S. Wolff, T.M. Janevic, I.R. Holzman, R. Yehuda, and P.J. Landrigan. 2003. The World Trade Center disaster and intrauterine growth restriction. *Jama*. 290:595-596.

Beydoun, H., and A.F. Saftlas. 2008. Physical and mental health outcomes of prenatal maternal stress in human and animal studies: a review of recent evidence. *Paediatric and perinatal epidemiology*. 22:438-466.

Bioinformatics, B. 2008. <SeqMonk Course Manual.pdf>.

Bioinformatics, B. 2013. Reduced Representation Bisulfite-Seq

A Brief Guide to RRBS.

Boissonneault, G. 2002. Chromatin remodeling during spermiogenesis: a possible role for the transition proteins in DNA strand break repair. *FEBS letters*. 514:111-114.

Bonilla, E., and E.Y. Xu. 2008. Identification and characterization of novel mammalian spermatogenic genes conserved from fly to human. *MHR: Basic science of reproductive medicine*. 14:137-142.

Braun, R.E. 2001. Packaging paternal chromosomes with protamine. *Nature genetics*. 28:10.

Braunschweig, M., V. Jagannathan, A. Gutzwiller, G. Bee, and T. Shioda. 2012. Investigations on Transgenerational Epigenetic Response Down the Male Line in F2 Pigs (Transgenerational Epigenetic Response in Pig). *PLoS ONE*. 7:e30583.

Cao-Lei, L., D.P. Laplante, and S. King. 2016. Prenatal maternal stress and epigenetics: review of the human research. *Current Molecular Biology Reports*. 2:16-25.

Cao-Lei, L., R. Massart, M.J. Suderman, Z. Machnes, G. Elgbeili, D.P. Laplante, M. Szyf, and S. King. 2014. DNA methylation signatures triggered by prenatal maternal stress exposure to a natural disaster: Project Ice Storm. *PloS one*. 9:e107653.

Cheung, P., C.D. Allis, and P. Sassone-Corsi. 2000. Signaling to chromatin through histone modifications. *Cell*. 103:263-271.

Cheung, P., and P. Lau. 2005. Epigenetic Regulation by Histone Methylation and Histone Variants. *Molecular Endocrinology*. 19:563-573.

Choi, M., J. Lee, M.T. Le, D.T. Nguyen, S. Park, N. Soundrarajan, K.M. Schachtschneider, J. Kim, J.-K. Park, and J.-H. Kim. 2015a. Genome-wide analysis of DNA methylation in pigs using reduced representation bisulfite sequencing. *DNA research*. 22:343-355.

Choi, M., J. Lee, M.T. Le, D.T. Nguyen, S. Park, N. Soundrarajan, K.M. Schachtschneider, J. Kim, J.-K. Park, J.-H. Kim, and C. Park. 2015b. Genome-wide analysis of DNA methylation in pigs using reduced representation bisulfite sequencing. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*. 22:343-355.

Chu, D.S., H. Liu, P. Nix, T.F. Wu, E.J. Ralston, J.R. Yates, 3rd, and B.J. Meyer. 2006. Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. *Nature*. 443:101-105.

Colvin, J.S., R.P. Green, J. Schmahl, B. Capel, and D.M. Ornitz. 2001. Male-to-female sex reversal in mice lacking fibroblast growth factor 9. *Cell*. 104:875-889.

Conesa, A., and S. Götz. 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics*. 2008:619832.

Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 21:3674-3676.

Congras, A., M. Yerle-Bouissou, A. Pinton, F. Vignoles, L. Liaubet, S. Ferchaud, and H. Acloque. 2014. Sperm DNA methylation analysis in swine reveals conserved and species-specific methylation patterns and highlights an altered methylation at the GNAS locus in infertile boars. *Biology of reproduction*. 91:137.

Curley, J.P., R. Mashoodh, and F.A. Champagne. 2011. Epigenetics and the origins of paternal effects. *Hormones and behavior*. 59:306-314.

Daxinger, L., and E. Whitelaw. 2012. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet*. 13:153-162.

De Jonge, C.J., C.J.D. Jonge, and C.L.R. Barratt. 2006. The sperm cell : production, maturation, fertilization, regeneration. Cambridge University Press, Cambridge, UK ;.

Deaton, A.M., and A. Bird. 2011. CpG islands and the regulation of transcription. *Genes & development*. 25:1010-1022.

Delbès, G., B.F. Hales, and B. Robaire. 2010. Toxicants and human sperm chromatin integrity. *MHR: Basic science of reproductive medicine*. 16:14-22.

Eriksson, J.G. 2010. Early programming of later health and disease: factors acting during prenatal life might have lifelong consequences. *Diabetes*. 59:2349-2350.

Evenson, D., and R. Wixon. 2006. Meta-analysis of sperm DNA fragmentation using the sperm chromatin structure assay. *Reproductive BioMedicine Online*. 12:466-472.

Evenson, D.P. 1999. Loss of livestock breeding efficiency due to uncompensable sperm nuclear defects. *Reproduction, fertility, and development*. 11:1-15.

Evenson, D.P. 2016. The Sperm Chromatin Structure Assay (SCSA®) and other sperm DNA fragmentation tests for evaluation of sperm nuclear DNA integrity as related to fertility. *Animal Reproduction Science*. 169:56-75.

Evenson, D.P., L. Thompson, and L. Jost. 1994. Flow cytometric evaluation of boar semen by the sperm chromatin structure assay as related to cryopreservation and fertility. *Theriogenology*. 41:637-651.

FD Myromslien, NH Tremoen, I Andersen-Ranberg, R Fransplass,a, EB Stenseth, , and M.v.S. TT Zeremichael, E Grindflek2, AH Gaustad. 2018. Sperm DNA integrity in Landrace and Duroc boar semen and its relationship to litter size I.N.U.o.A.S. Department of Biotechnology, NO-2318 Hamar, Norway, editor.

Feeney, A., E. Nilsson, and M.K. Skinner. 2014. Epigenetics and transgenerational inheritance in domesticated farm animals. *Journal of Animal Science and Biotechnology*. 5:48.

Felsenfeld, G., and M. Groudine. 2003. Controlling the double helix. *Nature*. 421:448.

Fouse, S.D., R.P. Nagarajan, and J.F. Costello. 2010. Genome-scale DNA methylation analysis. *Epigenomics*. 2:105-117.

Franklin, T.B., and I.M. Mansuy. 2010. Epigenetic inheritance in mammals: evidence for the impact of adverse environmental effects. *Neurobiology of disease*. 39:61-65.

Gardiner-Garden, M., and M. Frommer. 1987. CpG Islands in vertebrate genomes. *Journal of Molecular Biology*. 196:261-282.

Geno SA. 2018. Om Geno.

Gluckman, P.D., M.A. Hanson, and A.S. Beedle. 2007. Non-genomic transgenerational inheritance of disease risk. *Bioessays*. 29:145-154.

Godfrey, K.M., K.A. Lillycrop, G.C. Burdge, P.D. Gluckman, and M.A. Hanson. 2007. Epigenetic Mechanisms and the Mismatch Concept of the Developmental Origins of Health and Disease. *Pediatric Research*. 61:5R.

Goldberg, A.D., C.D. Allis, and E. Bernstein. 2007. Epigenetics: A Landscape Takes Shape. *Cell*. 128:635-638.

Gotz, S., J.M. Garcia-Gomez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talon, J. Dopazo, and A. Conesa. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 36:3420-3435.

Groenen, M.A.M., A.L. Archibald, H. Uenishi, C.K. Tuggle, Y. Takeuchi, M.F. Rothschild, C. Rogel-Gaillard, C. Park, D. Milan, H.-J. Megens, S. Li, D.M. Larkin, H. Kim, L.A.F. Frantz, M. Caccamo, H. Ahn, B.L. Aken, A. Anselmo, C. Anthon, L. Auvil, B. Badaoui, C.W. Beattie, C. Bendixen, D. Berman, F. Blecha, J. Blomberg, L. Bolund, M. Bosse, S. Botti, Z. Bujie, M. Bystrom, B. Capitanu, D. Carvalho-Silva, P. Chardon, C. Chen, R. Cheng, S.-H. Choi, W. Chow, R.C. Clark, C. Clee, R.P.M.A. Crooijmans, H.D. Dawson, P. Dehais, F. De Sapio, B. Dibbits, N. Drou, Z.-Q. Du, K. Eversole, J. Fadista, S. Fairley, T. Faraut, G.J. Faulkner, K.E. Fowler, M. Fredholm, E. Fritz, J.G.R. Gilbert, E. Giuffra, J. Gorodkin, D.K. Griffin, J.L. Harrow, A. Hayward, K. Howe, Z.-L. Hu, S.J. Humphray, T. Hunt, H. Hornshøj, J.-T. Jeon, P. Jern, M. Jones, J. Jurka, H. Kanamori, R. Kapetanovic, J. Kim, J.-H. Kim, K.-W. Kim, T.-H. Kim, G. Larson, K. Lee, K.-T. Lee, R. Leggett, H.A. Lewin, Y. Li, W. Liu, J.E. Loveland, Y. Lu, J.K. Lunney, J. Ma, O. Madsen, K. Mann, L. Matthews, S. McLaren, T. Morozumi, M.P. Murtaugh, J. Narayan, D. Truong Nguyen, P. Ni, S.-J. Oh, S. Onteru, F. Panitz, E.-W. Park, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 491:393.

Gu, H., C. Bock, T.S. Mikkelsen, N. Jager, Z.D. Smith, E. Tomazou, A. Gnirke, E.S. Lander, and A. Meissner. 2010a. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods*. 7:133-136.

Gu, H., C. Bock, T.S. Mikkelsen, N. Jäger, Z.D. Smith, E. Tomazou, A. Gnirke, E.S. Lander, and A. Meissner. 2010b. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature methods*. 7:133.

Gu, H., Z.D. Smith, C. Bock, P. Boyle, A. Gnirke, and A. Meissner. 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols*. 6:468.

Guerrero-Bosagna, C., M. Settles, B. Lucker, and M.K. Skinner. 2010. Epigenetic transgenerational actions of vinclozolin on promoter regions of the sperm epigenome. *PloS one*. 5:e13100.

Güneş, S., and T. Kulaç. 2013. The role of epigenetics in spermatogenesis. *Turkish journal of urology*. 39:181-187.

Hafez, E.S.E., and B. Hafez. 2000. Reproduction in farm animals. Lippincott Williams & Wilkins, Philadelphia.

Hall, V.J., J. Christensen, Y. Gao, M.H. Schmidt, and P. Hyttel. 2009. Porcine pluripotency cell signaling develops from the inner cell mass to the epiblast during early development. *Developmental dynamics: an official publication of the American Association of Anatomists*. 238:2014-2024.

Hernández, H.G., M.Y. Tse, S.C. Pang, H. Arboleda, and D.A. Forero. 2013. Optimizing methodologies for PCR-based DNA methylation analysis. *Biotechniques*. 55:181-197.

Ho, H.-C., and S.S. Suarez. 2001. Hyperactivation of mammalian spermatozoa: function and regulation. *Reproduction*. 122:519-526.

Houshdaran, S., V.K. Cortessis, K. Siegmund, A. Yang, P.W. Laird, and R.Z. Sokol. 2007. Widespread epigenetic abnormalities suggest a broad DNA methylation erasure defect in abnormal human sperm. *PloS one*. 2:e1289.

Hwang, J.H., S.M. An, S. Kwon, D.H. Park, T.W. Kim, D.G. Kang, G.E. Yu, I.-S. Kim, H.C. Park, J. Ha, and C.W. Kim. 2017. DNA methylation patterns and gene expression associated with litter size in Berkshire pig placenta.(Research Article)(Report). *PLoS ONE*. 12:e0184539.

Illumina, I. 2010. Illumina Sequencing Technology.

Irizarry, R.A., C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J.B. Potash, S. Sabunciyan, and A.P. Feinberg. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*. 41:178.

Januskauskas, A., A. Johannisson, and H. Rodriguez-Martinez. 2001. Assessment of sperm quality through fluorometry and sperm chromatin structure assay in relation to field fertility of frozen-thawed semen from Swedish AI bulls. *Theriogenology*. 55:947-961.

Jin, J., N. Jin, H. Zheng, S. Ro, D. Tafolla, K.M. Sanders, and W. Yan. 2007. Catsper3 and Catsper4 Are Essential for Sperm Hyperactivated Motility and Male Fertility in the Mouse1. *Biology of reproduction*. 77:37-44.

Jones, P.A., and D. Takai. 2001. The role of DNA methylation in mammalian epigenetics. *Science*. 293:1068-1070.

Jones, P.L., G.C. Jan Veenstra, P.A. Wade, D. Vermaak, S.U. Kass, N. Landsberger, J. Strouboulis, and A.P. Wolffe. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*. 19:187.

Kim, Y., A. Kobayashi, R. Sekido, L. DiNapoli, J. Brennan, M.-C. Chaboissier, F. Poulat, R.R. Behringer, R. Lovell-Badge, and B. Capel. 2006. Fgf9 and Wnt4 act as antagonistic signals to regulate mammalian sex determination. *PLoS biology*. 4:e187.

Krueger, F., and S.R. Andrews. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 27:1571-1572.

Krueger, F., and S.R. Andrews. 2012. Quality Control, trimming and alignment of Bisulfite-Seq data (Prot 57). Epigenesys.

Laqqan, M., and M. Hammadeh. 2018. Aberrations in sperm DNA methylation patterns of males suffering from reduced fecundity. *Andrologia*. 50:e12913.

Lee, R.T.H., Z. Zhao, and P.W. Ingham. 2016. Hedgehog signalling. *Development*. 143:367.

Legendre, C., G.C. Gooden, K. Johnson, R.A. Martinez, W.S. Liang, and B. Salhia. 2015. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clinical Epigenetics*. 7:100.

Lewis, S.E.M. 2013. The place of sperm DNA fragmentation testing in current day fertility management. *Middle East Fertility Society Journal*. 18:78-82.

Li, Z., J. Yu, T. Zhang, H. Li, and Y. Ni. 2013. rs189037, a functional variant in <em>ATM</em> gene promoter, is associated with idiopathic nonobstructive azoospermia. *Fertility and Sterility*. 100:1536-1541.e1531.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. 2012. Comparison of next-generation sequencing systems. *BioMed Research International*. 2012.

Love, C.C., and R.M. Kenney. 1998. The relationship of increased susceptibility of sperm DNA to denaturation and fertility in the stallion. *Theriogenology*. 50:955-972.

Lu, J., H. Zhang, L. Zhang, and C. Luo. 2015. Chapter 11 - Bioinformatics and Biostatistics in Mining Epigenetic Disease Markers and Targets A2 - Zheng, Y. George. *In* Epigenetic Technological Applications. Academic Press, Boston. 219-244.

Manikkam, M., C. Guerrero-Bosagna, R. Tracey, M.M. Haque, and M.K. Skinner. 2012. Transgenerational actions of environmental compounds on reproductive disease and identification of epigenetic biomarkers of ancestral exposures. *PloS one*. 7:e31901.

Meaburn, E., and R. Schulz. 2012. Next generation sequencing in epigenetics: insights and challenges. *In* Seminars in cell & developmental biology. Vol. 23. Elsevier. 192-199.

Mercer, T.R., and J.S. Mattick. 2013. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural &Amp; Molecular Biology*. 20:300.

Michael, L.M. 2009. Sequencing technologies — the next generation. *Nature Reviews Genetics*. 11:31.

Montjean, D., A. Zini, C. Ravel, S. Belloc, A. Dalleac, H. Copin, P. Boyer, K. McElreavey, and M. Benkhalifa. 2015. Sperm global DNA methylation level: association with semen parameters and genome integrity. *Andrology*. 3:235-240.

NCBI. 2018. Genome Assembly and Annotation report for Sus scrofa 11.1. https://www.ncbi.nlm.nih.gov/genome?term=sus%20scrofa.

Nikolova, Y.S., and A.R. Hariri. 2015. Can we observe epigenetic effects on human brain function? *Trends in cognitive sciences*. 19:366-373.

Nishimura, H., and S.W. L'Hernault. 2017. Spermatogenesis. *Current Biology*. 27:R988-R994.

Norsvin SA. 2018. Om Oss.

NuGEN. 2017. Ovation® RRBS Methyl-Seq

System 1–16. https://www.nugen.com/sites/default/files/M01394_v4_User_Guide%3A_Ovation_RRBS_Methyl-Seq_System_1285.pdf.

Olszewska, M., M.Z. Barciszewska, M. Fraczek, N. Huleyuk, V.B. Chernykh, D. Zastavna, J. Barciszewski, and M. Kurpisz. 2017. Global methylation status of sperm DNA in carriers of chromosome structural aberrations. *Asian Journal of Andrology*. 19:117-124.

Pacheco, S.E., E.A. Houseman, B.C. Christensen, C.J. Marsit, K.T. Kelsey, M. Sigman, and K. Boekelheide. 2011. Integrative DNA methylation and gene expression analyses identify DNA packaging and epigenetic regulatory genes associated with low motility sperm. *PLoS One*. 6:e20280.

Parrish, J.J. 2014. Bovine in vitro fertilization: in vitro oocyte maturation and sperm capacitation with heparin. *Theriogenology*. 81:67-73.

Pasek, R.C., E. Malarkey, N.F. Berbari, N. Sharma, R.A. Kesterson, L.L. Tres, A.L. Kierszenbaum, and B.K. Yoder. 2016. Coiled-coil domain containing 42 (Ccdc42) is necessary for proper sperm development and male fertility in the mouse. *Developmental biology*. 412:208-218.

Pesch, S., and M. Bergmann. 2006. Structure of mammalian spermatozoa in respect to viability, fertility and cryopreservation. *Micron*. 37:597-612.

Plass, C. 2001. CpG Islands. Springer Science & Business Media B.V. / Books. 227-229.

Prins, J.R., N. Gomez-Lopez, and S.A. Robertson. 2012. Interleukin-6 in pregnancy and gestational disorders. *Journal of Reproductive Immunology*. 95:1-14.

Qiagen. 2009a. EpiTech ® Bisulfite Handbook.

Qiagen. 2009b. EpiTect Bisulfite Handbook.

QIAGEN. 2018. CLC Genomics Workbench 11. https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/.

QIAGEN. 2017. Bisulfite Sequencing

Plugin

USER MANUAL. http://resources.qiagenbioinformatics.com/manuals/bisulfite-sequencing/current/BisulfiteSequencing_Plugin_User_Manual.pdf.

Quail, M., M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 13.

Razin, A., and A.D. Riggs. 1980. DNA methylation and gene function. *Science*. 210:604-610.

Reis-Filho, J.S. 2009. Next-generation sequencing. *Breast Cancer Research*. 11:S12.

Robertson, S.A. 2007. GM-CSF regulation of embryo development and pregnancy. *Cytokine & Growth Factor Reviews*. 18:287-298.

Rousseaux, S., C. Caron, J. Govin, C. Lestrat, A.-K. Faure, and S. Khochbin. 2005. Establishment of male-specific epigenetic information. *Gene*. 345:139-153.

Sakkas, D., and J.G. Alvarez. 2010. Sperm DNA fragmentation: mechanisms of origin, impact on reproductive outcome, and analysis. *Fertility and sterility*. 93:1027-1036.

Schachtschneider, K.M., O. Madsen, C. Park, L.A. Rund, M.A.M. Groenen, and L.B. Schook. 2015. Adult porcine genome-wide DNA methylation patterns support pigs as a biomedical model. *BMC Genomics*. 16:743.

Schagdarsurengin, U., A. Paradowska-Dogan, and K. Steger. 2012. Analysing the sperm epigenome: Roles in early embryogenesis and assisted reproduction.

Seppala, M., G.J. Fraser, A.A. Birjandi, G.M. Xavier, and M.T. Cobourne. 2017. Sonic Hedgehog Signaling and Development of the Dentition. *Journal of Developmental Biology*. 5:6.

Sergerie, M., G. Laforest, L. Bujan, F. Bissonnette, and G. Bleau. 2005. Sperm DNA fragmentation: threshold value in male fertility. *Human Reproduction*. 20:3446-3451.

Shamsi, M.B., S.N. Imam, and R. Dada. 2011. Sperm DNA integrity assays: diagnostic and prognostic challenges and implications in management of infertility. *Journal of Assisted Reproduction and Genetics*. 28:1073-1085.

Sharkey, D.J., A.M. Macpherson, K.P. Tremellen, D.G. Mottershead, R.B. Gilchrist, and S.A. Robertson. 2012. TGF-β mediates proinflammatory seminal fluid signaling in human cervical epithelial cells. *The Journal of Immunology*:1200005.

Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. *Nature Biotechnology*. 26:1135.

Simon, L., K. Murphy, M. Shamsi, L. Liu, B. Emery, K. Aston, J. Hotaling, and D. Carrell. 2014. Paternal influence of sperm DNA integrity on early embryonic development. *Human Reproduction*. 29:2402-2412.

Singal, R., and G.D. Ginder. 1999. DNA Methylation. *Blood*. 93:4059-4070.

Sironen, A. 2018. Molecular genetics of the iMMotile short tail sperM defect.

Song, C., B. Gao, H. Wu, Y. Xie, X. Wang, B. Li, G. Chen, and J. Mao. 2011. Molecular cloning, spatial and temporal expression analysis of CatSper genes in the Chinese Meishan pigs. *Reproductive Biology and Endocrinology*. 9:132.

Spanò, M., J.P. Bonde, H.I. Hjøllund, H.A. Kolstad, E. Cordelli, and G. Leter. 2000. Sperm chromatin damage impairs human fertility‡‡The Danish First Pregnancy Planner Study is a collaborative follow-up study on environmental and biological determinants of fertility. The project is coordinated by the Steno Institute of Public Health, Aarhus University and is undertaken in collaboration with the Department of Growth and Reproduction, National University Hospital, Copenhagen. The team includes Jens Peter E. Bonde, Niels Henrik I. Hjøllund, Tina

Kold Jensen, Tine Brink Henriksen, Henrik A. Kolstad, Erik Ernst, Aleksander Giwercman, Niels Erik Skakkebæk, and Jørn Olsen. *Fertility and Sterility*. 73:43-50.

Stockwell, P.A., A. Chatterjee, E.J. Rodger, and I.M. Morison. 2014. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics*. 30:1814-1822.

Storey, B.T. 2006. The Sperm Cell. Production, Maturation, Fertilization, Regeneration edited by Christopher De Jonge and Christopher Barratt. 359 pp., United Kingdom: Cambridge University Press; 2006. ISBN: 0- 521- 85397- 4. Cost: $95.00. *Journal of Andrology*. 27:707-707.

Suarez, S. 2010. How do sperm get to the egg? Bioengineering expertise needed! *Experimental mechanics*. 50:1267-1274.

Suarez, S., and H. Ho. 2003. Hyperactivation of mammalian sperm. *CELLULAR AND MOLECULAR BIOLOGY-PARIS-WEGMANN-*. 49:351-356.

Takai, D., and P.A. Jones. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*. 99:3740-3745.

Toshimori, K. 2009. Dynamics of the mammalian sperm head : modifications and maturation events from spermatogenesis to egg activation. Springer, Berlin.

Travis, A.J., C.J. Jorgez, T. Merdiushev, B.H. Jones, D.M. Dess, L. Diaz-Cueto, B.T. Storey, G.S. Kopf, and S.B. Moss. 2001. Functional relationships between capacitation-dependent cell signaling and compartmentalized metabolic pathways in murine spermatozoa. *Journal of Biological Chemistry*. 276:7630-7636.

Ulrey, C.L., L. Liu, L.G. Andrews, and T.O. Tollefsbol. 2005. The impact of metabolism on DNA methylation. *Human Molecular Genetics*. 14:R139-R147.

Visconti, P.E., G.D. Moore, J.L. Bailey, P. Leclerc, S.A. Connors, D. Pan, P. Olds-Clarke, and G.S. Kopf. 1995. Capacitation of mouse spermatozoa. II. Protein tyrosine phosphorylation and capacitation are regulated by a cAMP-dependent pathway. *Development*. 121:1139-1150.

Visconti, P.E., V.A. Westbrook, O. Chertihin, I. Demarco, S. Sleight, and A.B. Diekman. 2002. Novel signaling pathways involved in sperm acquisition of fertilizing capacity. *Journal of Reproductive Immunology*. 53:133-150.

Wang, J., Y. Xia, L. Li, D. Gong, Y. Yao, H. Luo, H. Lu, N. Yi, H. Wu, X. Zhang, Q. Tao, and F. Gao. 2013. Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. *BMC Genomics*. 14:11-11.

Ward, W.S., and D. Coffey. 1991. DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. *Biology of reproduction*. 44:569-574.

Waterland, R.A. 2009. Is epigenetics an important link between early life events and adult disease? *Hormone Research in Paediatrics*. 71:13-16.

Weinhold, B. 2006. Epigenetics: The Science of Change. *Environmental Health Perspectives*. 114:A160-A167.

Weyrich, A., D. Lenz, M. Jeschek, T.H. Chung, K. Rübensam, F. Göritz, K. Jewgenow, and J. Fickel. 2016. Paternal intergenerational epigenetic response to heat exposure in male Wild guinea pigs. *Molecular Ecology*. 25:1729-1740.

Wilantho, A., O. Praditsap, W. Charoenchim, S. Kulawonganunchai, A. Assawamakin, and S. Tongsima. 2012. Next generation sequencing (NGS) technologies and their applications in omics-research.

Wreczycka, K., A. Gosdschan, D. Yusuf, B. Grüning, Y. Assenov, and A. Akalin. 2017. Strategies for analyzing bisulfite sequencing data. *Journal of biotechnology*. 261:105-115.

Xiong, M., Z. Zhao, J. Arnold, and F. Yu. 2011. Next-generation sequencing. *Journal of BioMed Research*. 2010.

Yang, H., R. Fries, and G. Stranzinger. 1993. The sex- determining region Y (SRY) gene is mapped to p12- p13 of the Y chromosome in pig (Sus scrofa domestica) by in situ hybridization. *Animal genetics*. 24:297-300.

Zhou, Y., E.E. Connor, D.M. Bickhart, C. Li, R.L. Baldwin, S.G. Schroeder, B.D. Rosen, L. Yang, C.P. Van Tassell, and G.E. Liu. 2018. Comparative whole genome DNA methylation profiling of cattle sperm and somatic tissues reveals striking hypomethylated patterns in sperm. *GigaScience*. 7:giy039.

Zhou, Y., L. Xu, D.M. Bickhart, E.H. abdel Hay, S.G. Schroeder, E.E. Connor, L.J. Alexander, T.S. Sonstegard, C.P. Van Tassell, H. Chen, and G.E. Liu. 2016. Reduced representation bisulphite sequencing of ten bovine somatic tissues reveals DNA methylation patterns and their impacts on gene expression. *BMC Genomics*. 17:779.

Ziller, M.J., K.D. Hansen, A. Meissner, and M.J. Aryee. 2015. Coverage recommendations for methylation analysis by whole genome bisulfite sequencing. *Nature methods*. 12:230-232.

Zini, A. 2011. Are sperm chromatin and DNA defects relevant in the clinic? *Systems biology in reproductive medicine*. 57:78-85.

Zoll, J., E. Snelders, P.E. Verweij, and W.J.G. Melchers. 2016. Next-Generation Sequencing in the Mycology Lab. *Current Fungal Infection Reports*. 10:37-42.

SUPPLEMENTARY FILES

1. **S1-Differentially methylated Cs from different set of parameters**

   Description: The excel file contains DMCs obtained from methylkit in different sets of parameters including 25% DM in 5 samples, 25% DM in 4 samples, 25% DM in 3 samples, and 20% DM in 3 samples

2. **S2- DMCs from low vs. high and low vs. medium (6 samples)**

   Description: The excel file contains DMCs obtained from low vs. high and low vs. medium (6 samples) with SeqMonk analysis and common DMCs between both groups.

3. **S3- DMRs obtained from CLC**

   Description: The excel file contains DMs obtained from CLC and common DMCs between CLC and methylkit.

4. **S4- Blast2GO, KEGG pathway analysis with enzyme code and name (10 samples)**

   Description: The excel file contains Blast2GO analysis of genes obtained from nearest TSS of DMCs, KEGG pathway analysis of enzyme coding genes and enzyme code with name.

5. **S5- Graphical representation of KEGG pathways (10 samples)**

   Description: The PDF file contains the graphical representation of KEGG networking pathways for enzyme coding genes (10 samples).

6. **S6A- Graphical representation of KEGG pathways of low vs. high (6 samples)**

   Description: The PDF file contains the graphical representation of KEGG networking pathways for enzyme coding genes in low vs. high (6 samples).

7. **S6B- Graphical representation of KEGG pathways of low vs. medium (6 samples)**

   Description: The PDF file contains the graphical representation of KEGG networking pathways for enzyme coding genes in low vs. medium (6 samples).

8. **S6C- Blast2GO, KEGG pathway analysis with enzyme code and name for low vs. med and low vs. high (6 samples)**

9. **Description:** The excel file contains Blast2GO analysis of genes obtained from nearest TSS of DMCs, KEGG pathway analysis of enzyme coding genes and enzyme code with name for low vs. med and low vs. high (6 samples).

10. **S7- The details of enriched categories analysis for 10 samples group**

    Description: The excel file contains details of the genes involved in enriched categories with Entrez gene ID.

11. **S8- The details of enriched categories analysis for 6 samples groups (Low vs. high and low vs. med)**

Description: The excel file contains details of the genes involved in enriched categories with Entrez gene ID.

12. **S9- SeqMonk analysis of 10 samples group**

    Description: The excel file contains visualization report of DMCs from 10 samples group analysis based on CpG features.

13. **S10- The primer sequences.**

    Description: The file contains list of primer sequences.

14. **S11- List of figures**

    Description: The file contains list of images of 6 samples group (low vs. high and low vs. medium) obtained from methylkit.

15. **S12- The enrichment categories analysis of DMCs common between low vs. High and low vs. Medium (6 samples)**

    Description: The file includes enrichment categories analysis of common DMCs found between low vs. high and low vs. medium (6 samples) groups with number of genes and Entrez gene ID.

16. **All results from Medium vs. high (6 samples) group analysis.**