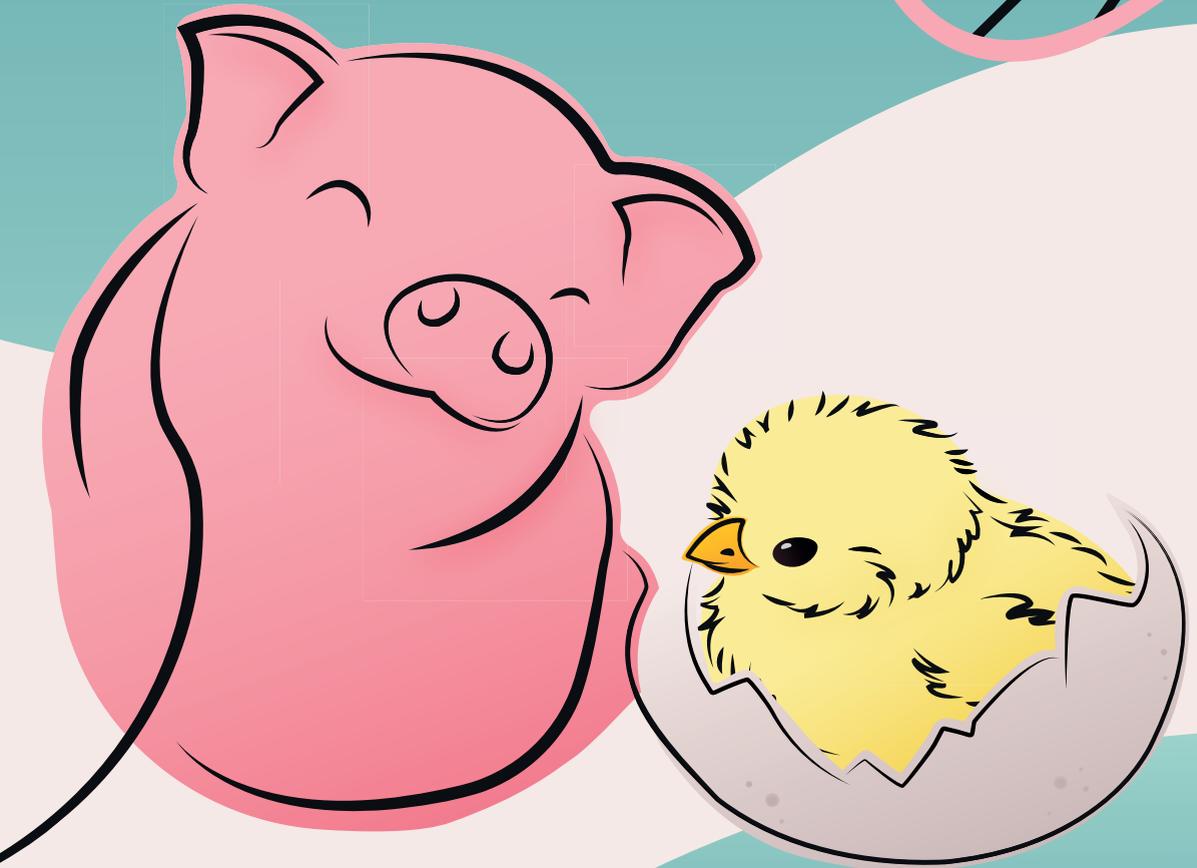


*Disentangling the epigenome
during development of
pig and chicken*



Jani de Vos



Propositions

1. Epigenome maps will become the next 'reference genomes'
(this thesis)
2. Implementation of functional maps into breeding models should be common practice
(this thesis)
3. Journal policies should include publication of all computational code and raw data
4. AI will become our next collaborator in scientific papers
5. Adaptation is crucial in science
6. Social media are essential for communicating science
7. Livestock breeds in developing countries are an essential resource in combating climate change

Propositions belonging to the thesis entitled:

Disentangling the epigenome during development of pig and chicken

Jani de Vos

Wageningen, 18 December 2023

Disentangling the epigenome during development of pig and chicken

Jani de Vos

Promotor

Prof. Dr M.A.M. Groenen

Professor of Animal Breeding and Genomics

Wageningen University & Research

Co-promotor

Dr. O Madsen

Associate professor, Animal Breeding and Genomics Group

Wageningen University & Research

Other members

Dr E.L. Clark, The Roslin Institute, University of Edinburgh, United Kingdom

Dr D.M. Bickhart, Hendrix Genetics, Boxmeer

Prof. Dr K. van Oers, Netherlands Institute of Ecology (NIOO-KNAW) and Wageningen University & Research

Prof. Dr. M.E. Schranz, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Disentangling the epigenome during development of pig and chicken

Jani de Vos

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. D. A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Monday 18 December, 2023

at 4 p.m. in the Omnia Auditorium.

Jani de Vos

Disentangling the epigenome during development of pig and chicken,

260 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)

With references, with summary in English

ISBN: 978-94-6447-974-4

DOI: <https://doi.org/10.18174/641526>

Vir Mamma en Pappa, die de Vosse, Hennings, en my familie.

Abstract

de Vos. (2023). Disentangling the epigenome during development of pig and chicken. PhD thesis, Wageningen University, the Netherlands

The genome provides the entire set of DNA instructions of an organism, while the epigenome involves modifications that do not alter the DNA sequence. The complex interplay between the genome and epigenome, regulates gene expression which is important during the developmental trajectory including processes such as differentiation of organs and tissues. DNA methylation, a type of epigenetic modification, plays a crucial role in the developmental process. However, our understanding of the dynamics of DNA methylation in the developing embryo and fetus is still limited. In this thesis, I investigate the developmental process in both pig and chicken, using a multifaceted approach. Firstly, molecular characteristics of a pig and chicken cell line are investigated using various functional assays, including whole genome sequencing, reduced representation – and whole genome bisulfite sequencing (RRBS and WGBS) to investigate DNA methylation, ChIP-sequencing of histone modifications denoting regulatory elements, ATAC-seq profiling open chromatin, and RNA sequencing to identify gene expression patterns. This provides a comprehensive investigation into cell line utility from an (epi)genomic standpoint. Secondly, further investigation delves into the DNA methylation landscape of seven distinct tissues at three developmental stages in both pig and chicken. Moreover, I identify methylome changes during early and late organogenesis until birth/hatching in pig and chicken. Subsequently, these changes are identified on both a per-site and whole-genome level during development per tissue. Integration of methylation data with transcriptomic profiles uncovers the intricate relationship/interplay between DNA methylation and tissue-specific gene expression patterns. Additionally, a comparative analysis between results from pig and chicken provides insights into mammalian and avian developmental epigenomics. This research provides a glimpse into the complex methylome dynamics that underpin developmental regulation in pig and chicken. By disentangling the complex epigenomic architecture of these species, this study contributes to the broader understanding of developmental processes and the potential application of these insights across diverse species.

Contents

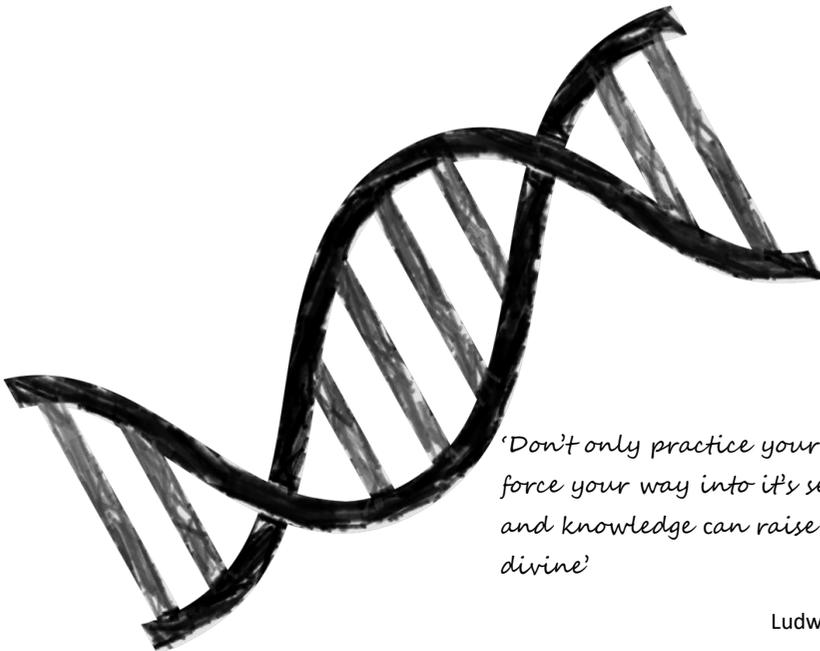
	Abstract	7
Chapter 1	General introduction	11 - 43
Chapter 2	Detailed molecular and epigenetic characterization of the pig IPEC-J2 and chicken SL-29 cell lines	45 - 96
Chapter 3	GSM-pipeline: GENE-SWitCH pipeline for comprehensive bisulfite sequencing analysis	97 - 107
Chapter 4	Methylome dynamics regulating fetal development of seven tissues in pig	109 - 150
Chapter 5	Spatio-temporal methylome dynamics in chicken	151 - 190
Chapter 6	General discussion	191 - 232
	Appendix A: WCGALP short paper	233 - 240
	Summary	241 - 245
	Curriculum Vitae	247 - 251
	Acknowledgements	253 - 259

List of abbreviations

Abbreviation	Description
ATAC	Assay for Transposase-Accessible Chromatin
CADD	Combined Annotation-Dependent Depletion
ChIP-seq	Chromatin immunoprecipitation (ChIP) followed by sequencing
CNV	Copy Number Variation
DEG	Differentially Expressed Genes
Dev	Developmental transition: 30 dpf → 70 dpf; E8 → E15
DMEG	Differentially Methylated Expressed Genes (DMR) associated with DEG
DMR	Differentially Methylated Regions identified from WGBS data
DMS	Differentially Methylated Sites identified from RRBS data
DNMT	DNA methylation by DNA methyltransferases
dpf	Days post fertilization
E15	Embryo at 15 days
E8	Embryo at 8 days
FAANG	Functional Annotation of Animal Genomes
FAETH	Functional-And-Evolutionary Trait Heritability
FMR	Fully Methylated Regions
GBLUP	Genomic best linear unbiased prediction
GO	Gene Ontology
TPM	Transcripts per million
GWAS	Genome-wide association studies
LMR	Lowly Methylated Regions
lncRNA	long non-coding RNA's
Mat	Maturation transition: 70 dpf → newborn; E15 → hatching
PCA	Principal component analysis
QTL	Quantitative Trait Loci
RRBS	Reduced Representation Bisulfite Sequencing
SNP	Single Nucleotide Polymorphism
SNVs	Single Nucleotide Variants
SV	Structural Variants
TET	Ten-Eleven Translocation enzymes
TFBS	Transcription factor binding site
TSS	Transcription Start Site
UMR	Unmethylated Methylated Regions
UTR	Untranslated region
WGBS	Whole genome bisulfite sequencing
WGS	Whole Genome Sequencing

Chapter 1

General introduction



*'Don't only practice your art. But
force your way into it's secrets, for it
and knowledge can raise men into
divine'*

Ludwig van Beethoven

All living organisms are composed of billions of cells, and DNA is the control centre of each cell. That leads to the question: **What is DNA?** It is the abbreviation for deoxyribonucleic acid, commonly known as the molecule that carries the genetic information necessary during an organisms lifetime and basic functions. Simply put it is the **blueprint of life**. Genes are the basic units of this genetic information and are inherited from parents to offspring. This is one of the main reasons we resemble our father, mother, siblings and family members (Calladine and Drew, 1997; *Deoxyribonucleic Acid (DNA)*, 2022; *Introduction: What is DNA? | Learn Science at Scitable*, no date).

The **genome** refers to the complete DNA sequence which encodes an organisms' genetic composition, while the **epigenome** is defined as (reversible) chemical modifications which do not involve alterations to the DNA sequence itself but which, together with regulatory DNA elements, are involved in regulating gene expression (Francis, 2011; Gonzalez-Recio, Toro and Bach, 2015). Until recently (1990) scientists were primarily interested in protein coding genes which comprise only about 1-3% of the entire genome (ENCODE Project Consortium *et al.*, 2012). The remaining DNA was classified as 'junk' DNA, although this concept started to be questioned by many researchers already throughout the 1990s. After a decade of research by scientists involved in the ENCODE project, ground breaking news emerged in 2012, revealing that previously labelling most of the DNA as junk was incorrect. (Pennisi, 2012; Tragante, Moore and Asselbergs, 2014; Snyder *et al.*, 2020). The discovery revealed that 80% of the human genome holds functional importance, by influencing gene expression in different capacities (Tragante, Moore and Asselbergs, 2014). A main conclusion from the ENCODE research was the **sophisticated complexity of gene regulation**. Many different methodologies and sequencing technologies were developed during and after the pilot phase of the ENCODE project (2007), which enabled investigating the regulatory landscapes of genomes. Thereafter, research into the functional regulatory genome and epigenome became more commonplace (ENCODE Project Consortium *et al.*, 2012; Pennisi, 2012; Snyder *et al.*, 2020).

1.1 The functional regulatory genome

DNA is packaged within the chromosomes in a highly organized structured manner, consisting of DNA bound to histones. The DNA-histone complex forms nucleosomes, which further fold into a 30nm fibre. Through further packaging, chromatid structures are produced (Annunziato, 2008; Klemm, Shipony and Greenleaf, 2019). The degree and process of packaging influence the accessibility of chromosomes to bind non-histone proteins such as transcription factors, which in turn determines the gene regulatory capacity (Annunziato, 2008; Klemm, Shipony and

1 General Introduction

Greenleaf, 2019). Regulatory elements have a paramount role in the timing and degree of gene expression, resulting in differences between tissues and species (**Figure 1**) (Liu *et al.*, 2004; Maston, Evans and Green, 2006; Levo and Segal, 2014). Depending on their regulatory role, non-coding sequences can be classified into various types of regulatory elements, such as promoters, enhancers, silencers, and insulators. Identifying regulatory elements poses greater challenges compared to identifying genes, mainly due to challenges in determining and characterizing promoters, enhancers and transcription factor binding sites (TFBS – see **Box 1**). For instance, the cell type specificity of regulatory elements requires investigating multiple cell types, which is time-consuming and resource-intensive (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014). Although experimental techniques like ChIP-seq, are valuable, they have limitations such as technical artifacts, which can result in possible erroneous predictions impacting the accuracy and reliability of identified regulatory elements (Andersson and Sandelin, 2020). Furthermore, regulatory elements can interact over long genomic distances, forming complex networks of interactions, making them challenging to determine and to understand their interactions and functional consequences (Lenhard, Sandelin and Carninci, 2012; Zaugg *et al.*, 2022).

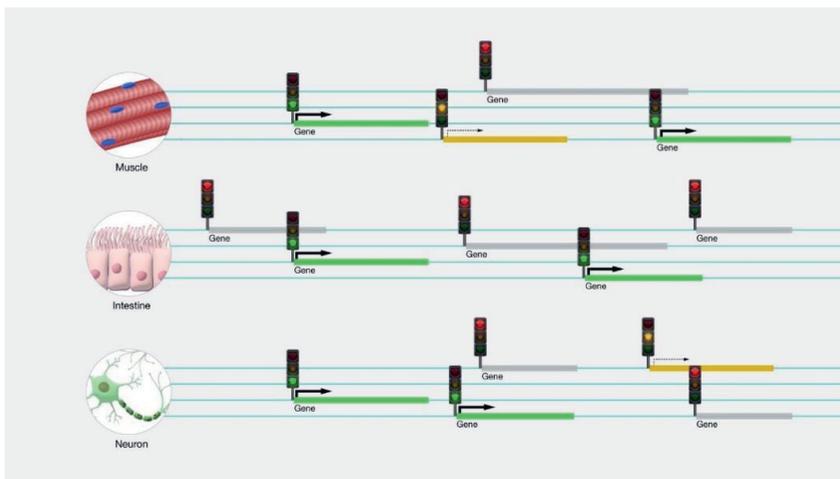


Figure 1 Illustration of gene regulation, which is involved in development and tissue differentiation. Three different tissues are shown in this diagram where genes are switched on (green light), off (red light) and intermediate (yellow light), illustrating the difference in gene expression and regulation between tissues. Arrows indicate direction of transcription of the gene. Adapted from National Human Genome Research Institute [glossary](#) explaining gene regulation.

1.1.1 Promoters

The promoter is a specific DNA sequence located upstream near the start of a gene, facilitating where transcription will start (**Figure 2**). The promoter region serves as a 'docking site' for basic transcriptional machinery such as RNA polymerase, general transcription factors and the transcription pre-initiation complex (Cramer, 2019; Andersson and Sandelin, 2020). The binding to the described elements initializes transcription during which a gene's DNA is used to produce a complementary RNA molecule. The location of the transcription start site (TSS) and the direction of transcription are defined by the promoter (Maston, Evans and Green, 2006; Heintzman and Ren, 2009).

Promoters can be classified into different types based on their structural features, and functionality. These include core-, proximal-, constitutive-, inducible-, tissue-specific-, and bidirectional promoters. Each of these types have unique characteristics, which enable precise control over gene expression in response to specific environmental cues, developmental stages, or cellular conditions. **Core promoters** are typically found ~50 bp around the TSS, and examples thereof includes TATA-box and initiator elements (Maston, Evans and Green, 2006; Haberle *et al.*, 2014; Cramer, 2019; Andersson and Sandelin, 2020). Promoters of this class are diverse in their content and organization. **Proximal promoter** elements are located upstream from the core promoter and many binding sites are available for transcription factor activators (Maston, Evans and Green, 2006; Andersson and Sandelin, 2020). Interestingly CpG islands are located close to or in promoters in 60-70% of vertebrate genes (Antequera, 2003; Illingworth and Bird, 2009; Deaton and Bird, 2011). A CpG island is a section of DNA, typically 500 bp to 2 kb in length, that has a high CG dinucleotide frequency and content in comparison to bulk DNA (Greenberg and Bourc'his, 2019). CpG islands and DNA methylation will be described in greater detail in the next section of this introduction.

BOX1: Transcription Factor Binding Site (TFBS)

A short (6-12bp) DNA sequence usually located within a promoter or enhancer, where transcription factors bind with a high affinity. Transcription factors are non-histone proteins that can bind to DNA and regulate the transcription of nearby genes e.g. activating or repressing gene expression by different mechanisms. TFBS can ultimately influence cellular processes such as development, differentiation, and response to environmental stimuli.

(Klemm, Shipony and Greenleaf, 2019)

1.1.2 Enhancers

Enhancers are typically specialized DNA sequences (100-1000 bp) which can be located further away from the promoter and gene (long-distance transcriptional control elements, **Figure 2**). In some cases enhancer elements can be located several hundred kilobases or even up to millions of bases (Panigrahi and O'Malley, 2021) from a promoter, and are usually located in intergenic and intronic regions of the genome. Enhancers contribute to gene regulation from such long-distances due to DNA looping, which subsequently places the enhancers in close proximity to the promoter and TSS (Smallwood and Ren, 2013; Grosveld, van Staalduinen and Stadhouders, 2021; Panigrahi and O'Malley, 2021). These elements work by attracting a wide variety of transcription factors and chromatin-modifying processes, facilitating gene transcription. Enhancers are responsible for recruiting transcription factors which enable the de-condensation of densely packed chromatin structures, facilitating the assembly of transcriptional machinery at gene promoters, and ultimately enabling transcription (Heintzman and Ren, 2009; Goldberg *et al.*, 2012; Klemm, Shipony and Greenleaf, 2019; Andersson and Sandelin, 2020). Moreover, enhancers play a crucial role in regulating unique gene expression patterns across different cell types. They have been identified as one of the main controllers of cell-type-specific gene expression, with the majority of enhancers exhibiting a strictly constrained developmental window of action (Grosveld, van Staalduinen and Stadhouders, 2021).

Enhancer elements use three main mechanisms to utilise co-factors of transcription factors to enhance transcription and target promoters. One mechanism involves transcription factors modifying the chromatin structure non-covalently, leading to **nucleosome remodelling**. This process potentially expose the promoters, enabling transcription. Another mechanism involves a specific class of cofactors modifying histones at the N-terminal tails (e.g. histones H3 and H4). This induces de-condensation of the packed nucleosome or

BOX2: SUPER ENHANCERS

Clusters of putative enhancers with abnormally high levels of enrichment for the binding of transcriptional coactivators. Generally extensive genomic areas are covered by these elements, with a median size an order of magnitude larger than typical enhancers (super enhancer: ~8000+ bp vs normal: ~700 bp). (Pott and Lieb, 2015; Grosveld, van Staalduinen and Stadhouders, 2021)

attracts additional chromatin remodelling factors. The third mechanism is the recruitment of enhancer elements called mediator complexes which play a role in transcription by acting as

a bridge between the general transcriptional machinery and transcription factors (Heintzman and Ren, 2009; Gasperini, Tome and Shendure, 2020; Panigrahi and O'Malley, 2021).

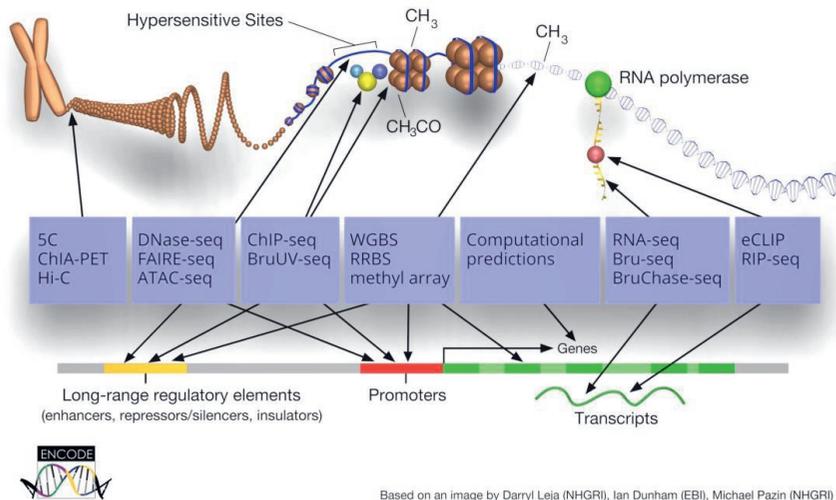


Figure 2 Functional genomic elements identified by the ENCODE project using diverse methods to investigate various types of functional elements. Methods for identifying functional elements are represented as purple boxes (obtained from: <https://www.encodeproject.org/>).

1.1.3 Other elements

Insulators, and silencing elements are other regulatory elements that are frequently found far from the genes they regulate (distal regulatory elements). These elements play a role in the transcriptional regulation of genes in addition to promoters and proximal cis-regulatory elements (Heintzman and Ren, 2009). Silencing elements (silencers) prevent the transcription of genes by binding specific proteins or by changing the structure of the chromatin. Generally, these elements work regardless of the promoter's orientation or distance, although some position-dependent silencers have been observed (Ogbourne and Antalis, 1998; Zhang et al., 2022). Silencers can be located at various positions relative to their target genes e.g. within an intron, 3'-untranslated region (UTR), proximal promoter, distal enhancer, or as an independent distal regulatory module. Silencers can act as a binding site for repressors, which are negative transcription factors, and can require co-repressors. There are cases where an activator switches to a repressor, and these can act from both a close distance to the gene as well as

1 General Introduction

from a longer distance (Kellum and Elgin, 1998; Maston, Evans and Green, 2006; Riethoven, 2010; Bell et al., 2011).

Insulator elements (boundary elements) offer additional regulation by limiting the expansion of heterochromatin and prevent transcriptional enhancers from activating unrelated promoters. The main function of these elements is preventing genes from being impacted by the transcriptional activity of neighbouring genes (Riethoven, 2010). Insulators typically act in a position-dependent, orientation-independent way, and range in length from 0.5 to 3 kb (Gaszner and Felsenfeld, 2006). CTCF (CCCTC-binding factor) is a highly conserved 11-zinc finger DNA-binding protein that

plays a critical role in regulating gene expression and genome organization. It is involved in diverse cellular processes, such as transcriptional regulation, chromatin architecture, and DNA damage repair (Klenova et al., 1993; Filippova et al., 1996; Bell et al., 2011; Alharbi et al., 2021). CTCF acts as an insulator, mediating long-range chromatin interactions that define and maintain the three-dimensional structure of the genome. CTCF is a sequence-specific DNA-binding protein that recognizes a palindromic DNA motif known as the CTCF binding site, with the consensus sequence CCGCGNGGNGGCG. In this motif, N represents any nucleotide (A, C, G, T). (Lobanenkov et al., 1990; Kim et al., 2007). This motif is found at multiple genomic locations, including promoters, enhancers, and insulators. CTCF binding can prevent the spread of repressive chromatin modifications, block the interaction between enhancers and promoters, and facilitate the formation of higher-order chromatin structures. Recent studies have shown that CTCF is involved in a variety of biological processes, including stem cell differentiation, development, and disease (Heath et al., 2008; Phillips and Corces, 2009; van Ruiten and Rowland, 2021; Dehingia et al., 2022).

1.2 Epigenomic modifications (epigenome)

The field of epigenetics has revolutionized our understanding of how gene expression is regulated beyond the DNA sequence. Since the initial concept of the epigenome in 1939 (Waddington, 1939) it has seen substantial development, and now it is defined as a collection of chemical modifications that occur on DNA and histone proteins that package the DNA (Villota-Salazar, Mendoza-Mendoza and González-Prieto, 2016; John and Rougeulle, 2018;

Box 3:

Heterochromatin

Tightly packed and condensed form of chromatin, which consists of DNA, histone proteins, and other molecules.

Euchromatin

Less condensed and more accessible form of chromatin, which allows for active gene transcription and gene expression.

Klemm, Shipony and Greenleaf, 2019). These modifications play a critical role in regulating gene expression by controlling the accessibility of DNA to the transcriptional machinery, which ultimately determines which genes are turned on or off. The epigenome is dynamic and can be modified in response to internal and external cues, such as environmental factors or developmental signals. As such, it plays a critical role in a wide range of biological processes, including cell differentiation, development, aging, and disease (Bernstein, Meissner and Lander, 2007; Bogdanović, van Heeringen and Veenstra, 2012; Pal and Tyler, 2016; Wang *et al.*, 2022).

1.2.1 DNA methylation

DNA methylation is the addition of a methyl group (-CH₃) to a cytosine. In vertebrates methylation typically takes place at so-called CpG sites where a cytosine is located 5' from a guanine. DNA methylation can be dynamically regulated through the active removal of methyl groups, a process known as DNA demethylation, as well as through DNA methylation by DNA methyltransferases (DNMTs, described in the next paragraph). Active DNA demethylation by TET enzymes, short for Ten-Eleven Translocation enzymes, involve several mechanisms, including enzymatic oxidation of methylcytosine followed by base excision repair or replication-dependent dilution of methylated cytosines. Gene expression is influenced by DNA methylation and it is susceptible to environmental factors (Goldberg *et al.*, 2012; Levo and Segal, 2014; Greenberg and Bourc'his, 2019; Moore *et al.*, 2020; Greenberg, 2021; Nasrullah *et al.*, 2022).

DNMTs are enzymes responsible for adding the methyl groups to DNA. In mammals, the two main types of DNMTs are maintenance DNMTs (DNMT1) and de novo DNMTs (DNMT3A and DNMT3B) (Greenberg, 2021). Homologs of these DNMT types identified in mammals are also found in other eukaryotes such as *Arabidopsis* (Chen and Li, 2004). Maintenance DNMTs ensure that DNA methylation patterns are inherited during DNA replication, while de novo DNMTs establish new methylation patterns during early development or in response to environmental stimuli (Miranda and Jones, 2007; Greenberg and Bourc'his, 2019).

TET enzymes are a family of enzymes that play a key role in the process of active DNA demethylation. TET enzymes are responsible for catalysing the initial steps of DNA demethylation and there are three known TET enzymes in mammals: TET1, TET2, and TET3 (Yin and Xu, 2016; Lio and Rao, 2019; Zeng and Chen, 2019).

Two distinct forms of DNA methylation exist: CpG and non-CpG methylation, which differ in their target sites and patterns. CpG methylation involves the methylation of cytosine residues within CpG dinucleotides (Villota-Salazar, Mendoza-Mendoza and González-Prieto, 2016; John and Rougeulle, 2018; Nasrullah *et al.*, 2022). CpG islands are regions of DNA with a high

1 General Introduction

density of CpG dinucleotides, often located near or in gene promoter regions. CpG islands can be either unmethylated, associated with active gene expression, or methylated, associated with gene silencing. This is the most common and well-studied form of DNA methylation and it plays a critical role in gene regulation, genomic imprinting, X-chromosome inactivation, and maintaining genome stability (Greenberg and Bourc'his, 2019). In normal cells, CpG islands in gene promoter regions are usually unmethylated, allowing for gene expression. Methylation of CpG sites within promoter regions can lead to gene silencing by preventing the binding of transcription factors and other regulatory proteins (Miranda and Jones, 2007; Jin, Li and Robertson, 2011; Greenberg and Bourc'his, 2019; Greenberg, 2021; Nasrullah *et al.*, 2022).

Non-CpG methylation involves the methylation of cytosine residues in contexts other than CpG (Patil, Ward and Hesson, 2014). It is prevalent in certain cell types, including embryonic stem cells, neurons, and germ cells (de Mendoza *et al.*, 2021). Non-CpG methylation plays a role in various biological processes, such as neuronal development, X-chromosome inactivation in pluripotent stem cells, and DNA damage response (Ziller *et al.*, 2011).

DNA methylation can be studied by various methods, and a commonly used method is bisulfite sequencing, whereby unmethylated cytosines are converted to uracils and methylated cytosines remain unchanged. Site specific DNA methylation changes can be detected through this methodology, however application of this method on a whole genome scale (whole genome bisulphite sequencing – WGBS) is costly. Reduced representation bisulphite sequencing (RRBS) applies genome-wide DNA methylation analysis with reduced sequencing of about 2% of the genome, which reduces the costs for investigating DNA methylation (Doherty and Couldrey, 2014). Furthermore, CpG rich regions are preferentially selected and sequenced by RRBS which add to the biological relevance of the methods as CpG rich regions are important for gene regulation. Due to its cost-saving advantages, RRBS has been widely utilized for investigating DNA methylation.

1.2.2 Histone modifications and CHIP-sequencing

Histone modifications are post-translational alterations that occur on the histone proteins associated with DNA in eukaryotic cells. These modifications play a critical role in regulating gene expression, chromosome structure, and various cellular processes (Wolffe and Hayes, 1999; Gardner, Allis and Strahl, 2011). Histone proteins are the building blocks of nucleosomes, which are the basic units of chromatin, the complex of DNA and proteins in the nucleus (Inbar-Feigenberg *et al.*, 2013). The histone core consists of two copies, each of four different types of histone proteins: H2A, H2B, H3 and H4, which form an octamer. Histone modifications can occur through the addition or removal of various chemical groups, such as

acetyl, methyl, phosphoryl, ubiquitin, at specific amino acid residues on the histone protein tails (Kimura, 2013).

Acetylation involves the addition of an acetyl group to lysine residues in the histone tails, and is generally associated with gene activation or transcriptional activation (Wolffe and Hayes, 1999; Kimura, 2013). Acetylated histones create a more open chromatin structure, allowing better access to DNA by the transcription machinery (Zhang, Cooper and Brockdorff, 2015). Histone (H3) methylation can occur on both lysine and arginine residues in histones. Depending on the specific amino acid and the degree of methylation, it can either be associated with gene activation or gene repression (Wolffe and Hayes, 1999; Kimura, 2013; Zhang, Cooper and Brockdorff, 2015). Lysine can be methylated at three different levels, namely monomethylation (me1), dimethylation (me2), and trimethylation (me3) (Kimura, 2013). Histone phosphorylation involves the addition of a phosphate group to serine or threonine residues in histones (Kimura, 2013). Histone phosphorylation is often associated with gene activation and is involved in processes such as DNA repair, cell cycle regulation, and cellular signalling.

These are just a few examples of the wide range of histone modifications that can occur and specific combinations and patterns of histone modifications, which create a "histone code", act as a molecular language, dictating the DNA accessibility, recruitment of protein complexes, and ultimately influencing gene expression and cellular processes (Gardner, Allis and Strahl, 2011). Histone modifications are dynamic and reversible (Pazin and Kadonaga, 1997; Katan-Khaykovich and Struhl, 2002), allowing for precise regulation of gene activity in response to developmental signals, environmental cues, or cellular conditions

In 2007 the first comprehensive genome-wide binding maps implementing chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq) were created (Park, 2009). ChIP-seq is mainly used for the identification of histone modifications, as well as nucleosome modifications due to modifying proteins. Profiling of the DNA-binding proteins such as transcription factors, RNA polymerases and CTCF, at a genome-wide level, is also done using this methodology. The histone modifications are of importance because they regulate the chromatin states which determine the accessibility of DNA for transcription factors to bind which ultimately activates transcription (Park, 2009; Lorzadeh *et al.*, 2016; Andersson and Sandelin, 2020).

1.2.3 Long non-coding RNA

Transcriptome sequencing (RNA-seq) is a powerful method for the determination and characterization of transcripts which in turn aids in the identification of genes (**Figure 2**)

1 General Introduction

(Ozsolak and Milos, 2011). There are many developments in RNA-seq methodology which enable an improved and more complete depiction of the RNA transcripts. RNA is more than the intermediary involved in protein synthesis. RNAs have been found to be essential in controlling many cellular processes including gene expression, transcription and translation. Development of methods for investigating RNA has enabled the subdivision of RNA into coding and non-coding RNA. Although long non-coding RNA's (lncRNA) do not encode proteins, these lncRNA play a pivotal role in the regulation of gene expression through the control of chromatin accessibility, transcription factor occupancy and epigenomic state (see next section) (Mattick and Makunin, 2006; Kaikkonen and Adelman, 2018). Furthermore, non-coding RNAs such as miRNAs regulate gene translation and expression at the level of the mRNA. This leads to more genes encoding regulatory RNA in comparison with protein coding genes (Guil and Esteller, 2015).

1.2.4 Methods for investigating the functional regulatory epigenome

Increasing use of next-generation sequencing and the decrease in costs have enabled new methods which describe the internal states of cells on a genome-wide level. In **Table 1** I summarise techniques used to investigate the epigenome, functional regulatory genome and the transcriptome (Zhao and Garcia, 2015).

A computational method for investigating the functional regulatory genome employs models that combine different data types to annotate regions of the genome into epigenomic states (Ernst and Kellis, 2017). This provides a definition for different regions of the genome that influence gene expression and chromatin structure, thereby regulating cellular processes and determining cellular identity. Epigenomic states can be broadly classified as active or repressive. In active states, genes are accessible for transcription, leading to their active expression (e.g. active TSS). These states are characterized by histone acetylation, low levels of DNA methylation, and accessible genes (regions with open chromatin), allowing the binding of transcription factors and regulatory proteins. In contrast, repressed or silent epigenomic states render genes inaccessible for transcription, leading to their repression. These states are often associated with histone deacetylation, high levels of DNA methylation (especially at CpG islands), and other repressive histone modifications. Repressed genes are typically situated in heterochromatin (**Box 3**), hindering the access of the transcription machinery to DNA.

Table 1 Sequencing methods (used in this thesis and the GENE-SWitCH project) for investigating the epigenome and functional regulatory genome

Molecular assay	Elements	Description
ChIP-seq (Chromatin Immunoprecipitation)	H3K4me3	Epigenetic modification of Histone protein 3, with a trimethylation modification (me3) at the 4 th Lysine residue. This is found at the promoter of all expressed genes.
	H3K4me1	Epigenetic modification of Histone protein 3, with a methylation modification (me3) at the 4 th Lysine residue. This modification typically marks active transcriptional enhancers.
	H3K27me3	Epigenetic modification of Histone protein 3, with a trimethylation modification (me3) at the 27 th Lysine residue. Marks genes that have been silenced through regional modification, as well as identifies enhancers and promoters. Typically, for genes which are active during embryological development and silenced later on during development.
	H3K27Ac	Epigenetic modification of Histone protein 3, with an acetylation modification of the 27 th Lysine residue. Useful for the identification of active enhancer elements.
	CTCF	Transcription factor: highly conserved zinc finger protein. Element that is involved in the 3D structure and folding of the genome and regulates the enhancer-promoter interactions. Sequence specific insulator protein.
RNA-seq	mRNA-seq	Sequencing RNA molecules with a polyA tail
	Small RNA-seq	Method used for the identification of different small RNA such as micro-RNA
	Long RNA-seq (Iso-seq)	Method used to achieve full-length gene transcripts. Used to explore and annotate new alternative transcripts and genes.
DNA methylation	RRBS	Reduced Representation bisulphite sequencing. This technique is used to identify areas with enriched CpG regions.
	WGBS	Whole genome Bisulphite sequencing. Useful for investigating the methylation of the whole genome.
Hi-C	Capture Hi-C	Method for investigation chromatin accessibility, which has a high resolution as well as is high-throughput
ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing)		Used to study the open chromatin, nucleosome positioning, and transcription factor occupancy
WGS (Whole genome sequencing)		Provides a high resolution base-by-base view of entire genomes.

1 General Introduction

This epigenome annotation method has been utilized in species such as pig, chicken, mouse, and human (Pan *et al.*, 2021; van der Velde *et al.*, 2021; T. Wang *et al.*, 2021; Pan *et al.*, 2023). Consortia like the 'human roadmap to the epigenome' showcase the potential of compiling multiple (reference) epigenomic maps, enabling answering fundamental biology questions (Satterlee *et al.*, 2019). **Figure 3** presents an example of epigenomic states in the mouse during development, including among others, active TSS and different types of enhancers as active states, and heterochromatin repressed by polycomb as repressive states.

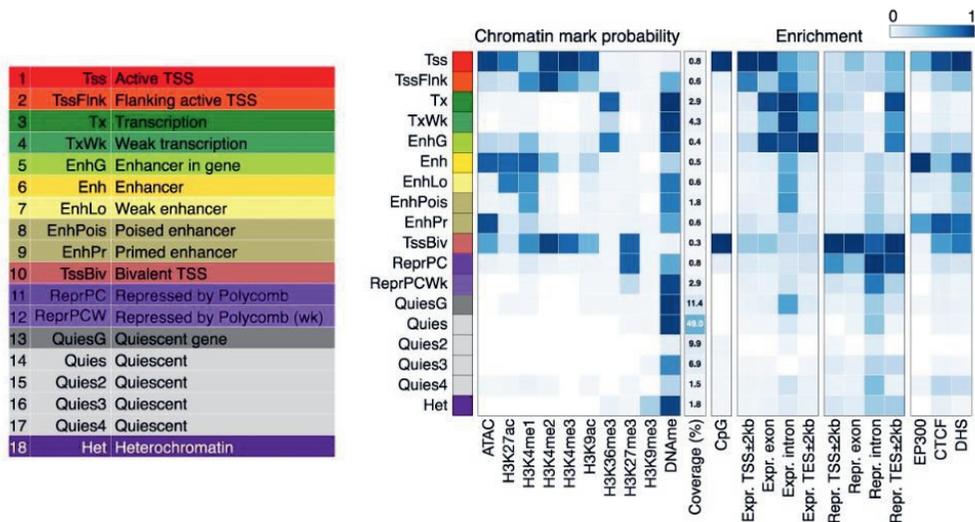


Figure 3 Epigenomic states identified for 66 mouse epigenomes, obtained from van der Velde *et al.*, 2021.

1.3 Investigating the epigenome in two monogastric species

Chickens and pigs are both domesticated animals that have been selectively bred by humans for thousands of years. Chickens were domesticated from a wild ancestor called the red junglefowl (*Gallus gallus*) that was native to South Asia. The domestication of chickens likely began around 10,000 years ago in Southeast Asia (Tixier-Boichard, Bed'hom and Rognon, 2011; Larson and Fuller, 2014; Xiang *et al.*, 2014). Over time, humans selectively bred chickens for traits such as egg production, meat quality, and docile behaviour. Pigs, on the other hand, are thought to have been domesticated around 9,000-10,000 years ago from wild boars (*Sus scrofa*) in East Anatolia and China (Groenen *et al.*, 2012; Groenen, 2016).

Both chicken and pig are primary sources for meat production globally, and they also serve as vital biomedical models and model organisms for evolutionary studies. Chickens are particularly valuable for investigating various biological aspects, such as virology, oncogenesis, and immunology, while their embryos provide a valuable system for studying vertebrate development (Bednarczyk *et al.*, 2021). Pigs, on the other hand, share genetic, anatomical, and physiological similarities with humans, making them valuable for research in human health and medicine. Additionally, both species are readily available and have a short generation interval, making these species ideal for scientific investigations (Meurens *et al.*, 2012).

The first chicken genome was published in 2004 (International Chicken Genome Sequencing Consortium, 2004) and this was the first livestock species to have its genome sequenced. The first genome assembly had a size of 1 billion bp which was later improved and extended to 1.21 Gb (Warren *et al.*, 2016). The pig genome was sequenced by the Swine genome sequencing consortium using a Duroc sow (Groenen *et al.*, 2012) and more recently an improved reference genome was published (Warr *et al.*, 2020). Both these species have high quality reference genomes with initial annotation focussed predominantly on protein coding genes. In recent years it has become evident that differences in gene expression due to variation in regulatory elements (i.e. the functional genome) are important for understanding trait variation (Pai, Pritchard and Gilad, 2015). Therefore, understanding trait variation requires annotation of the functional genome of these two species. Furthermore, many traits are settled during early developmental stages, thus another important aspect is to determine the dynamics of the functional genome during early development as it will provide additional power to selection for important production traits.

1.4. The vertebrate story: mammalian and bird development

Vertebrates are animals belonging to the subphylum Vertebrata, which have a vertebral column with vertebrae (spinal column) from which the name originated. The following groups of species are classified as vertebrates: fish, reptiles, amphibians, birds and mammals (**Figure 4**) (Kardong, 2019). Reptiles, birds and mammals are amniotes, in which a water-filled sac (amnion) surrounds the embryo and protects the embryo from impact (shock absorber), as well as preventing the embryo from drying out (Gilbert and Barresi, 2016; Kardong, 2019).

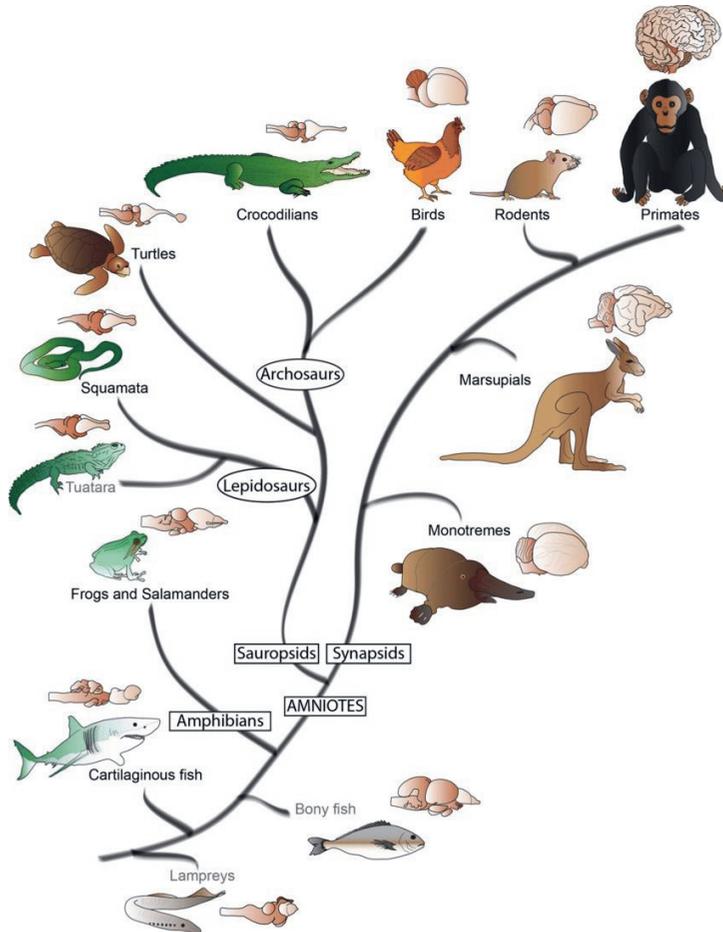


Figure 4 Phylogenetic tree of vertebrates showing the appearance of amniotes. (Obtained from Cárdenas and Borrell, 2020)

Despite the differences in avian and mammalian development, there are also many similarities. Both groups of animals are amniotes, which means that they develop inside an amniotic egg or in the uterus (Gilbert and Barresi, 2016). Similar stages of embryonic and fetal development occur in both avian and mammalian species (chicken and pig). Gastrulation and the process of germ layer formation, which differentiates into organs and tissues, occur similarly in both birds and mammals.

1.4.1 Development of embryological germ layers

Vertebrates have a complex body architecture composed of morphologically unique tissues and organs. During embryonic development, the cells of the fertilized egg divide and differentiate into various cell types that give rise to the different tissues and organs of the body. Gastrulation (**Figure 5**) is a fundamental, evolutionary conserved process during which three primary germ layers called the endoderm, mesoderm, and ectoderm are formed. Formation and differentiation of these germ layers are crucial for the development of a functional and healthy organism (Solnica-Krezel and Sepich, 2012; Kiecker, Bates and Bell, 2016). Organogenesis is the process where the embryonic germ layers further differentiate into various cell types that form specific tissues and organs (Beddington and Robertson, 1999; Eakin and Behringer, 2004; Muhr and Ackerman, 2023).

BOX4:

THE PUNCHLINE

Animal development is a complex cascade of processes such as fertilization, cleavage, gastrulation, organogenesis, metamorphosis and senescence. The complexity of these processes raises questions such as:

“How do cells become organized into functional organs?”

(Gilbert and Barresi, 2016)

“Which regions of the genome play a role in these processes” (this thesis)

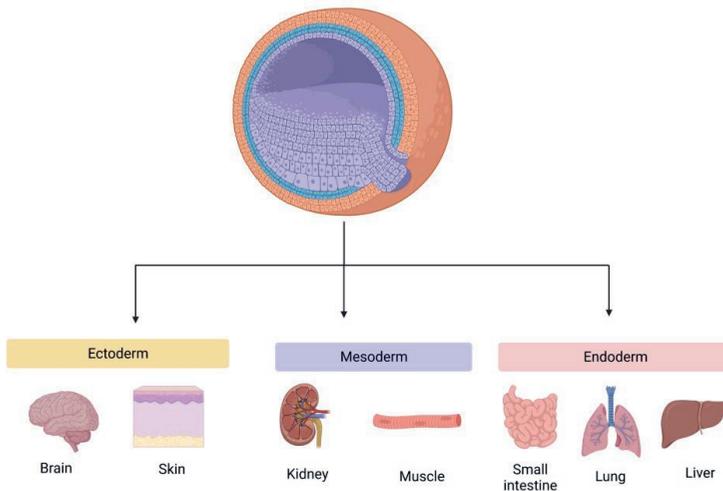


Figure 5 Gastrula with the three embryological germ layers, endoderm, ectoderm and mesoderm which gives rise to different tissues. (Created with BioRender.com)

1.4.1.1 Endoderm

The endoderm is the innermost germ layer, which gives rise to the lining of the digestive tract, respiratory tract, and various organs such as the liver, pancreas, and thyroid gland (Grapin-Botton and Melton, 2000; Bellairs and Osmond, 2005; Nowotschin, Hadjantonakis and Campbell, 2019). During embryonic development, one of the first steps is the formation of the endoderm (Gilbert and Barresi, 2016), which plays an essential role in a few processes such as the exchange of gasses and nutrients. In the amniote, nutrients and oxygen are provided to the fetus from the placenta or yolk sac and the formation of mesodermal organs are induced by the endoderm. Secondly, the linings of tubes such as the digestive tube, and respiratory tract are constructed from the endoderm. These further develop into structures such as the liver buds, pharynx, and pancreatic buds. Lastly the endoderm is responsible for the formation of epithelium in several glands e.g. tonsils and thyroid (Spence, Lauf and Shroyer, 2011; Gilbert and Barresi, 2016). The functions and molecular composition of some tissues will be described in more detail in the following sections.

a. Small intestine

The small intestine is a long, narrow tube that is part of the digestive system and is located between the stomach and the large intestine. It is responsible for most of the nutrient absorption during the digestive process, where it receives partially digested food from the stomach. Through a series of chemical and mechanical processes, the nutrients in the food are broken down further and absorbed into the bloodstream.

Tiny, finger-like projections, called **villi**, cover the wall of the small intestine which increase the surface area of the intestine and facilitate nutrient absorption. The small intestine is divided into three regions: the duodenum, jejunum, and **ileum**. The ileum is the last section of the small intestine, whose primary function is absorption of nutrients (Eckert and Randal, 1983; Frandson and Spurgeon, 1992; Spence, Lauf and Shroyer, 2011; Gilbert and Barresi, 2016). It also plays a role in the immune system, as it contains specialized lymphatic tissue called **Peyer's patches**.

b. Lungs

The lungs play a vital role in gas exchange between inhaled air and the bloodstream. Inside the lungs, air enters into tiny sacs called alveoli, where the exchange of gases take place. Oxygen from the air passes through the walls of the alveoli and into the bloodstream, where it

is transported to the body's tissues. Carbon dioxide from the tissues is transported back to the lungs via the bloodstream and expelled from the body during exhalation.

Lungs are complex organs that are divided into several lobes and are composed of several layers of tissue and structures. **Bronchi** and **bronchioles** are a network of tubes that bring air into the lungs. **Bronchi** are formed from the trachea, where the bronchi branch off into smaller tubes called **alveolar ducts**. (Eckert and Randal, 1983; Frandson and Spurgeon, 1992; Gilbert and Barresi, 2016).

c. Liver

The liver performs many essential functions in the body, including detoxification and metabolism of nutrients, such as carbohydrates, proteins, and fats. Furthermore, the liver produces bile, which helps to digest and absorb fats from the small intestine (Eckert and Randal, 1983; Frandson and Spurgeon, 1992). During early fetal development in eutherian mammals this organ plays a direct role in the synthesis of red blood cells, with this role changing in later life to the destruction of red blood cells (Frandson and Spurgeon, 1992) .

The liver is composed of many lobes, which are divided into smaller units called **lobules** and is one of the most vascularised tissues in the body with the most important being the **hepatic artery and hepatic portal vein**. It contains several types of cells, including **hepatocytes**, which are the primary functional cells of the liver. **Kupffer cells** are a type of immune cell that reside in the liver and help to remove toxins and foreign substances from the blood. **Bile ducts** are a network of tubes that transport bile from the liver to the gallbladder and small intestine (Frandson and Spurgeon, 1992; Kardong, 2019).

1.4.1.2 Mesoderm

One of the major processes during gastrulation is the formation of the mesoderm between the endoderm and ectoderm layers. Initially a paraxial -, intermediate -, and lateral plate mesoderm are formed. The intermediate mesoderm gives rise to the kidney and gonads, while muscles, tendons and cartilage originate from the paraxial mesoderm. The lateral plate mesoderm forms the circulatory system, pelvis, body cavities, and limb bones (Gilbert and Barresi, 2016).

a. Muscles

Muscles are specialized tissues that play a vital role in movement, stability, posture, and body temperature regulation (e.g. shivering). In the body, three types of muscle are found namely, smooth muscle, cardiac muscle, and **skeletal muscle**. Skeletal muscle is the muscle type

1 General Introduction

which constitutes the meat that is consumed from domestic animals and are attached to the bones.

Muscle is composed of specialized cells called **muscle fibers**. Muscles are organized in a structured manner, where bundles of muscle fibers are surrounded by fibrous connective tissue (**endomysium**). A sheath surrounds these bundles of muscle fibers and is known as the **perymysium** and lastly, the **epimysium** is the connective tissue encapsulating the entire muscle. Each muscle fiber contains many **myofibrils**, which are long, cylindrical structures that are responsible for muscle contraction. Myofibrils are made up of two types of protein filaments: thick filaments made of **myosin** and thin filaments made of **actin**. When the muscle is stimulated, the myosin filaments interact with the actin filaments, causing the muscle to contract (Frandsen and Spurgeon, 1992; Gilbert and Barresi, 2016). Upon investigation at a microscopic level, bands of myofibrils are observed where **I-bands** consist only of thin filaments, and **A-bands** have both thick and thin filaments overlapping. A **Z-line** is classified as a line which bisects each I-band, and a segment of myofibril between two Z-lines is known as the **sarcomere** which is the basic unit of contraction (Frandsen and Spurgeon, 1992; Csapo, Gumpenberger and Wessner, 2020; Mukund and Subramaniam, 2020; Z. Wang *et al.*, 2021).

b. Kidney

The kidney plays a role in filtering waste products from the blood, regulating electrolyte balance, and controlling fluid levels in the body. It also helps to regulate blood pressure and produces hormones that stimulate the production of red blood cells. Waste products and excess fluid from the blood is filtered by the kidney, and thereafter is eliminated from the body in the form of urine or uric acid in mammals and birds respectively. Embryonic development of this organ is very similar across vertebrate groups (Kardong, 2019).

The functional unit of the kidney is the **nephron**, which contains many thousands of cells and more than 10 different cell types. During development the kidney goes through three main stages, with the final stage being the only stage contributing to the functional kidney. Two main regions are found in the kidney: the renal cortex and renal medulla. The outer layer of the kidney is called the **renal cortex**, and millions of nephrons are found here. Nephrons filter the blood and reabsorb useful substances. The **renal medulla** is the inner layer of the kidney which contains renal pyramids and collecting ducts. Urine is transported via ducts from the nephrons to the renal pelvis, which connects to the ureter (Eckert and Randal, 1983; Frandsen and Spurgeon, 1992; Gilbert and Barresi, 2016).

1.4.1.3 Ectoderm

The outermost germ layer is known as the ectoderm, and gives rise to the skin, hair, nails, nervous system, and various sensory organs such as the eyes and ears.

a. Skin

Skin is the largest organ of the body and serves several important functions, such as providing a protective barrier that helps to shield the body from physical, chemical, and microbial damage. The skin contains nerve endings that allows perception of touch, pressure, temperature, and pain. Another important function is the regulation of body temperature through the process of shivering and sweating in species with sweat glands. The skin excretes small amounts of waste products, including sweat and sebum. However, birds lack sweat glands, while farm animals including pigs and buffaloes possess fewer sweat glands compared to cattle (Frandsen and Spurgeon, 1992; Ratnakaran *et al.*, 2016; Gourdine *et al.*, 2021; Mota-Rojas *et al.*, 2021; Khan *et al.*, 2023).

The skin is composed of three primary layers, namely the **epidermis**, **dermis**, **melanocytes** and **subcutaneous tissue**. Epidermis is the outermost layer of the skin and is primarily composed of **keratinocytes**, which produce a tough, waterproof protein called **keratin** (Gilbert and Barresi, 2016). **Melanocytes**, which produce the pigment melanin are found within the epidermis. **Langerhans cells**, which are involved in immune response are also found in the epidermis. The dermis is the middle layer of the skin and is composed of **collagen**, **elastin**, and **reticular fibers**. The dermis contains blood vessels, lymphatic vessels, and nerve endings, as well as hair or feather follicles and sebaceous glands (Frandsen and Spurgeon, 1992).

b. Brain

The brain serves as the central control centre of the nervous system, playing a fundamental role in coordinating and regulating various physiological and cognitive functions. It receives and processes sensory information from the body, transforming it into meaningful perceptions of the surrounding environment. Additionally, it actively regulates a multitude of autonomic bodily functions, including digestion, blood pressure regulation, and heart rate control.

Brain composition is complex with several different structures, each with its own unique functions. **Cerebrum** is the largest and most complex structure of the brain and is responsible for consciousness, perception, and voluntary movement. The **cerebellum** is located underneath the cerebrum and is responsible for coordinating movement and balance. At the

1 General Introduction

base of the brain, the **brainstem** is located, and many automatic bodily functions, such as breathing, and heart rate are regulated here. The brain is composed of two major types of specialized cells called neurons and glial cells, that work together to ensure the proper function of the brain. Neurons are responsible for transmitting electrical and chemical signals throughout the brain, while glial cells provide support and protection for the neurons (Frandsen and Spurgeon, 1992).

1.4.2 Differences in development between mammals and birds

Key differences in germ layer and organogenesis in birds and mammals are summarized in **Table 2**. There are several differences between avian and mammalian development, which can be attributed to differences in the way that the two groups of animals evolved.

Table 2 Summary of some key differences in fetal development and birth between mammals (pig) and birds (chicken). Many resources are available discussing this topic and further details can be found in the referenced resources (Gilbert and Barresi, 2016; Kardong, 2019)

Function or tissue	Mammals (eutherian)	Birds
Parity/birth	Viviparity Live birth of offspring that develop inside the womb of the mother.	Oviparity Birds lay eggs that develop outside of the mother's body.
Amniotic composition	Mammalian embryo develops inside a fluid-filled sac called the amniotic sac.	Avian egg has a hard outer shell.
Placenta vs Yolk nutrient supply	Placenta Embryo is nourished by the placenta, an organ that forms during pregnancy and allows the exchange of nutrients and waste products between the mother and the developing fetus. In mammals, early development is slower and more complex, as the embryo must develop specialized structures for exchange of nutrients and waste products with the mother.	Yolk Egg yolk contains all the nutrients needed for the developing embryo, so the early stages of development occur rapidly.

<p>Organ development</p>	<p>- Brain Characterized by a large cerebral cortex.</p> <p>- Lung Alveoli are tiny, grape-like structures within the lungs where gas exchange takes place and the walls thereof are thin, allowing for diffusion of gases between the air and the bloodstream.</p>	<p>- Brain More prominent cerebellum, which is associated with motor coordination.</p> <p>- Lung Avian lungs lack alveoli, and instead, gas exchange occurs across parabronchi, which are tube-like structures in the avian lung.</p>
---------------------------------	--	--

1.5. Aims and outline of this thesis

A major goal of genomics is to understand the functional components of the genome and epigenome. While the ENCODE project has been valuable for the initial identification of the functional elements in the human genome, the identification of functional elements in the genomes of pig and chicken in specific tissues during different phases of fetal development are still limited. The activation status of these functional elements may vary between cells and developmental stages. My research focuses on identifying the regulatory elements that regulate the genome during different stages of development (early and late organogenesis) in pig and chicken. This investigation contributes to identifying the 'switches' that regulate the activation status of these elements during different developmental stages. The developmental stages studied in pig and chicken were aimed to be similar, allowing for a functional comparison and the study of the evolution of these elements. **Figure 6** provides an outline of this thesis. It begins with a molecular characterization of two functionally relevant cell lines in pig and chicken (**Chapter 2**). Next, I describe a bioinformatics pipeline (**Chapter 3**) developed for the analysis of DNA methylation data, which was utilized in my research to identify changes in DNA methylation during development. **Chapters 4** and **5** investigate the dynamics of the DNA methylome at a spatio-temporal level, with a focus on pig fetal development (**Chapter 4**) and chicken embryo development (**Chapter 5**). These chapters identify elements that regulate organ development in chicken and pig, as well as in mammals and birds in general. The thesis concludes with a discussion (**Chapter 6**) of the general findings, the broader impact of this research, and future perspectives.

Chapter 1:
General introduction

Chapter 6:
General discussion

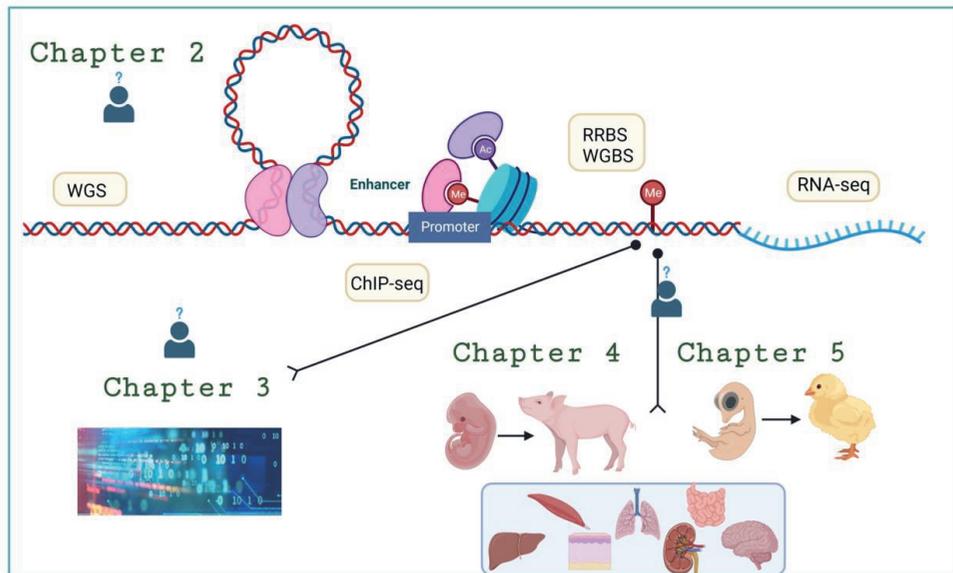


Figure 6 Schematic overview of the chapters in the thesis entitled “Disentangling the epigenome during development of pig and chicken”. (Created with BioRender.com)

1.6. References

- Alharbi, A.B. *et al.* (2021) ‘CTCF as a regulator of alternative splicing: new tricks for an old player’, *Nucleic Acids Research*, 49(14), pp. 7825–7838. Available at: <https://doi.org/10.1093/nar/gkab520>.
- Andersson, R. and Sandelin, A. (2020) ‘Determinants of enhancer and promoter activities of regulatory elements’, *Nature Reviews Genetics*, 21(2), pp. 71–87. Available at: <https://doi.org/10.1038/s41576-019-0173-8>.
- Annunziato, A.T. (2008) ‘DNA Packaging: Nucleosomes and Chromatin’, *Nature Education*, 1(1), p. 26.
- Antequera, F. (2003) ‘Structure, function and evolution of CpG island promoters’, *Cellular and Molecular Life Sciences CMLS*, 60(8), pp. 1647–1658. Available at: <https://doi.org/10.1007/s00018-003-3088-6>.
- Beddington, R.S.P. and Robertson, E.J. (1999) ‘Axis Development and Early Asymmetry in Mammals’, *Cell*, 96(2), pp. 195–209. Available at: [https://doi.org/10.1016/S0092-8674\(00\)80560-7](https://doi.org/10.1016/S0092-8674(00)80560-7).
- Bednarczyk, M. *et al.* (2021) ‘Chicken embryo as a model in epigenetic research’, *Poultry Science*, 100(7), p. 101164. Available at: <https://doi.org/10.1016/j.psj.2021.101164>.

- Bell, O. *et al.* (2011) 'Determinants and dynamics of genome accessibility', *Nature Reviews. Genetics*, 12(8), pp. 554–564. Available at: <https://doi.org/10.1038/nrg3017>.
- Bellairs, R. and Osmond, M. (2005) *The atlas of chick development*. 2nd edn. London, UK: Elsevier.
- Bernstein, B.E., Meissner, A. and Lander, E.S. (2007) 'The Mammalian Epigenome', *Cell*, 128(4), pp. 669–681. Available at: <https://doi.org/10.1016/j.cell.2007.01.033>.
- Bogdanović, O., van Heeringen, S.J. and Veenstra, G.J.C. (2012) 'The epigenome in early vertebrate development', *genesis*, 50(3), pp. 192–206. Available at: <https://doi.org/10.1002/dvg.20831>.
- Calladine, C.R. and Drew, H. (1997) *Understanding DNA: The Molecule and How It Works*. Academic Press.
- Cárdenas, A. and Borrell, V. (2020) 'Molecular and cellular evolution of corticogenesis in amniotes', *Cellular and Molecular Life Sciences*, 77(8), pp. 1435–1460. Available at: <https://doi.org/10.1007/s00018-019-03315-x>.
- Chen, T. and Li, E. (2004) 'Structure and Function of Eukaryotic DNA Methyltransferases', in *Current Topics in Developmental Biology*. Academic Press (Stem Cells in Development and Disease), pp. 55–89. Available at: [https://doi.org/10.1016/S0070-2153\(04\)60003-2](https://doi.org/10.1016/S0070-2153(04)60003-2).
- Cramer, P. (2019) 'Organization and regulation of gene transcription', *Nature*, 573(7772), pp. 45–54. Available at: <https://doi.org/10.1038/s41586-019-1517-4>.
- Csapo, R., Gumpenberger, M. and Wessner, B. (2020) 'Skeletal Muscle Extracellular Matrix – What Do We Know About Its Composition, Regulation, and Physiological Roles? A Narrative Review', *Frontiers in Physiology*, 11, p. 253. Available at: <https://doi.org/10.3389/fphys.2020.00253>.
- Deaton, A.M. and Bird, A. (2011) 'CpG islands and the regulation of transcription', *Genes & Development*, 25(10), pp. 1010–1022. Available at: <https://doi.org/10.1101/gad.2037511>.
- Dehingia, B. *et al.* (2022) 'CTCF shapes chromatin structure and gene expression in health and disease', *EMBO Reports*, 23(9), p. e55146. Available at: <https://doi.org/10.15252/embr.202255146>.
- Deoxyribonucleic Acid (DNA)* (2022) *Genome.gov*. Available at: <https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid> (Accessed: 5 March 2023).
- Doherty, R. and Couldrey, C. (2014) 'Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment', *Frontiers in Genetics*, 5. Available at: <https://www.frontiersin.org/articles/10.3389/fgene.2014.00126> (Accessed: 1 May 2023).
- Eakin, G.S. and Behringer, R.R. (2004) 'Diversity of germ layer and axis formation among mammals', *Seminars in Cell & Developmental Biology*, 15(5), pp. 619–629. Available at: <https://doi.org/10.1016/j.semcdb.2004.04.008>.
- Eckert, R. and Randal, D. (1983) *Animal Physiology: Mechanisms and Adaptations*. 2nd edn. United States of America: W.H. Freeman and Company.

1 General Introduction

- ENCODE Project Consortium *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. Available at: <https://doi.org/10.1038/nature11247>.
- Ernst, J. and Kellis, M. (2017) 'Chromatin-state discovery and genome annotation with ChromHMM', *Nature Protocols*, 12(12), pp. 2478–2492. Available at: <https://doi.org/10.1038/nprot.2017.124>.
- Filippova, G.N. *et al.* (1996) 'An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.', *Molecular and Cellular Biology*, 16(6), pp. 2802–2813.
- Francis, R.C. (2011) *Epigenetics: the ultimate mystery of inheritance*. New York, N.Y.: Norton.
- Frandsen, R.D. and Spurgeon, T.L. (1992) *Anatomy and Physiology of Farm Animals*. 5th edn. Pennsylvania, U.S.A.: Lea & Febiger.
- Gardner, K.E., Allis, C.D. and Strahl, B.D. (2011) 'OPERating ON Chromatin, a Colorful Language where Context Matters', *Journal of Molecular Biology*, 409(1), pp. 36–46. Available at: <https://doi.org/10.1016/j.jmb.2011.01.040>.
- Gasperini, M., Tome, J.M. and Shendure, J. (2020) 'Towards a comprehensive catalogue of validated and target-linked human enhancers', *Nature Reviews Genetics*, 21(5), pp. 292–310. Available at: <https://doi.org/10.1038/s41576-019-0209-0>.
- Gaszner, M. and Felsenfeld, G. (2006) 'Insulators: exploiting transcriptional and epigenetic mechanisms', *Nature Reviews Genetics*, 7(9), pp. 703–713. Available at: <https://doi.org/10.1038/nrg1925>.
- Gilbert, S.F. and Barresi, M.J.F. (2016) *Developmental Biology*. 11th edn. Sunderland, Massachusetts: Sinauer Associates, Inc.
- Goldberg, M.L. *et al.* (2012) *Genetics: from genes to genomes*. New York, NY: McGraw Hill.
- Gonzalez-Recio, O., Toro, M. and Bach, A. (2015) 'Past, present and future of epigenetics applied to livestock breeding', *Frontiers in Genetics*, 6. Available at: <https://www.frontiersin.org/articles/10.3389/fgene.2015.00305> (Accessed: 18 May 2023).
- Gourdine, J.-L. *et al.* (2021) 'The Genetics of Thermoregulation in Pigs: A Review', *Frontiers in Veterinary Science*, 8. Available at: <https://www.frontiersin.org/articles/10.3389/fvets.2021.770480> (Accessed: 14 August 2023).
- Grapin-Botton, A. and Melton, D.A. (2000) 'Endoderm development: from patterning to organogenesis', *Trends in Genetics*, 16(3), pp. 124–130. Available at: [https://doi.org/10.1016/S0168-9525\(99\)01957-5](https://doi.org/10.1016/S0168-9525(99)01957-5).
- Greenberg, M.V.C. (2021) 'Get Out and Stay Out: New Insights Into DNA Methylation Reprogramming in Mammals', *Frontiers in Cell and Developmental Biology*, 8. Available at: <https://www.frontiersin.org/articles/10.3389/fcell.2020.629068> (Accessed: 28 March 2023).

- Greenberg, M.V.C. and Bourc'his, D. (2019) 'The diverse roles of DNA methylation in mammalian development and disease', *Nature Reviews Molecular Cell Biology*, 20(10), pp. 590–607. Available at: <https://doi.org/10.1038/s41580-019-0159-6>.
- Groenen, M.A.M. *et al.* (2012) 'Analyses of pig genomes provide insight into porcine demography and evolution', *Nature*, 491(7424), pp. 393–398. Available at: <https://doi.org/10.1038/nature11622>.
- Groenen, M.A.M. (2016) 'A decade of pig genome sequencing: a window on pig domestication and evolution', *Genetics Selection Evolution*, 48(1), p. 23. Available at: <https://doi.org/10.1186/s12711-016-0204-2>.
- Grosveld, F., van Staaldouin, J. and Stadhouders, R. (2021) 'Transcriptional Regulation by (Super)Enhancers: From Discovery to Mechanisms', *Annual Review of Genomics and Human Genetics*, 22(1), pp. 127–146. Available at: <https://doi.org/10.1146/annurev-genom-122220-093818>.
- Guil, S. and Esteller, M. (2015) 'RNA-RNA interactions in gene regulation: the coding and noncoding players', *Trends in Biochemical Sciences*, 40(5), pp. 248–256. Available at: <https://doi.org/10.1016/j.tibs.2015.03.001>.
- Haberle, V. *et al.* (2014) 'Two independent transcription initiation codes overlap on vertebrate core promoters', *Nature*, 507(7492), pp. 381–385. Available at: <https://doi.org/10.1038/nature12974>.
- Heath, H. *et al.* (2008) 'CTCF regulates cell cycle progression of $\alpha\beta$ T cells in the thymus', *The EMBO Journal*, 27(21), pp. 2839–2850. Available at: <https://doi.org/10.1038/emboj.2008.214>.
- Heintzman, N.D. and Ren, B. (2009) 'Finding distal regulatory elements in the human genome', *Current Opinion in Genetics & Development*, 19(6), pp. 541–549. Available at: <https://doi.org/10.1016/j.gde.2009.09.006>.
- Illingworth, R.S. and Bird, A.P. (2009) 'CpG islands – “A rough guide”', *FEBS Letters*, 583(11), pp. 1713–1720. Available at: <https://doi.org/10.1016/j.febslet.2009.04.012>.
- Inbar-Feigenberg, M. *et al.* (2013) 'Basic concepts of epigenetics', *Fertility and Sterility*, 99(3), pp. 607–615. Available at: <https://doi.org/10.1016/j.fertnstert.2013.01.117>.
- International Chicken Genome Sequencing Consortium. (2004) 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.', *Nature*, 432, pp. 695–716. Available at: <https://doi.org/10.1038/nature03154>.
- Introduction: What is DNA? | Learn Science at Scitable* (no date). Available at: <https://www.nature.com/scitable/topicpage/introduction-what-is-dna-6579978/> (Accessed: 5 March 2023).
- Jin, B., Li, Y. and Robertson, K.D. (2011) 'DNA Methylation', *Genes & Cancer*, 2(6), pp. 607–617. Available at: <https://doi.org/10.1177/1947601910393957>.
- John, R.M. and Rougeulle, C. (2018) 'Developmental Epigenetics: Phenotype and the Flexible Epigenome', *Frontiers in Cell and Developmental Biology*, 6. Available at: <https://www.frontiersin.org/articles/10.3389/fcell.2018.00130> (Accessed: 1 May 2023).

1 General Introduction

- Kaikkonen, M.U. and Adelman, K. (2018) 'Emerging Roles of Non-Coding RNA Transcription', *Trends in Biochemical Sciences*, 43(9), pp. 654–667. Available at: <https://doi.org/10.1016/j.tibs.2018.06.002>.
- Kardong, K.V. (2019) *Vertebrates: Comparative Anatomy, Function, Evolution*. 8th edn. New York, NY, U.S.A: McGraw-Hill Education.
- Katan-Khaykovich, Y. and Struhl, K. (2002) 'Dynamics of global histone acetylation and deacetylation in vivo: rapid restoration of normal histone acetylation status upon removal of activators and repressors', *Genes & Development*, 16(6), pp. 743–752. Available at: <https://doi.org/10.1101/gad.967302>.
- Kellum, R. and Elgin, S.C.R. (1998) 'Chromatin boundaries: Punctuating the genome', *Current Biology*, 8(15), pp. R521–R524. Available at: [https://doi.org/10.1016/S0960-9822\(07\)00337-5](https://doi.org/10.1016/S0960-9822(07)00337-5).
- Khan, R.U. *et al.* (2023) 'Physiological dynamics in broiler chickens under heat stress and possible mitigation strategies', *Animal Biotechnology*, 34(2), pp. 438–447. Available at: <https://doi.org/10.1080/10495398.2021.1972005>.
- Kiecker, C., Bates, T. and Bell, E. (2016) 'Molecular specification of germ layers in vertebrate embryos', *Cellular and Molecular Life Sciences*, 73(5), pp. 923–947. Available at: <https://doi.org/10.1007/s00018-015-2092-y>.
- Kim, T.H. *et al.* (2007) 'Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome', *Cell*, 128(6), pp. 1231–1245. Available at: <https://doi.org/10.1016/j.cell.2006.12.048>.
- Kimura, H. (2013) 'Histone modifications for human epigenome analysis', *Journal of Human Genetics*, 58(7), pp. 439–445. Available at: <https://doi.org/10.1038/jhg.2013.66>.
- Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) 'Chromatin accessibility and the regulatory epigenome', *Nature Reviews Genetics*, 20(4), pp. 207–220. Available at: <https://doi.org/10.1038/s41576-018-0089-8>.
- Klenova, E.M. *et al.* (1993) 'CTCF, a Conserved Nuclear Factor Required for Optimal Transcriptional Activity of the Chicken c-myc Gene, Is an 11-Zn-Finger Protein Differentially Expressed in Multiple Forms', *Molecular and Cellular Biology*, 13(12), pp. 7612–7624. Available at: <https://doi.org/10.1128/mcb.13.12.7612-7624.1993>.
- Larson, G. and Fuller, D.Q. (2014) 'The Evolution of Animal Domestication', *Annual Review of Ecology, Evolution, and Systematics*, 45(1), pp. 115–136. Available at: <https://doi.org/10.1146/annurev-ecolsys-110512-135813>.
- Lenhard, B., Sandelin, A. and Carninci, P. (2012) 'Metazoan promoters: emerging characteristics and insights into transcriptional regulation', *Nature Reviews Genetics*, 13(4), pp. 233–245. Available at: <https://doi.org/10.1038/nrg3163>.
- Levo, M. and Segal, E. (2014) 'In pursuit of design principles of regulatory sequences', *Nature Reviews Genetics*, 15(7), pp. 453–468. Available at: <https://doi.org/10.1038/nrg3684>.

- Lio, C.-W.J. and Rao, A. (2019) 'TET Enzymes and 5hmC in Adaptive and Innate Immune Systems', *Frontiers in Immunology*, 10. Available at: <https://www.frontiersin.org/articles/10.3389/fimmu.2019.00210> (Accessed: 28 July 2023).
- Liu, Y. *et al.* (2004) 'Eukaryotic Regulatory Element Conservation Analysis and Identification Using Comparative Genomics', *Genome Research*, 14(3), pp. 451–458. Available at: <https://doi.org/10.1101/gr.1327604>.
- Lobanenkov, V.V. *et al.* (1990) 'A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene'.
- Lorzadeh, A. *et al.* (2016) 'Nucleosome Density ChIP-Seq Identifies Distinct Chromatin Modification Signatures Associated with MNase Accessibility', *Cell Reports*, 17(8), pp. 2112–2124. Available at: <https://doi.org/10.1016/j.celrep.2016.10.055>.
- Maston, G.A., Evans, S.K. and Green, M.R. (2006) 'Transcriptional Regulatory Elements in the Human Genome', *Annual Review of Genomics and Human Genetics*, 7(1), pp. 29–59. Available at: <https://doi.org/10.1146/annurev.genom.7.080505.115623>.
- Mattick, J.S. and Makunin, I.V. (2006) 'Non-coding RNA', *Human Molecular Genetics*, 15(suppl_1), pp. R17–R29. Available at: <https://doi.org/10.1093/hmg/ddl046>.
- de Mendoza, A. *et al.* (2021) 'The emergence of the brain non-CpG methylation system in vertebrates', *Nature Ecology & Evolution*, 5(3), pp. 369–378. Available at: <https://doi.org/10.1038/s41559-020-01371-2>.
- Meurens, F. *et al.* (2012) 'The pig: a model for human infectious diseases', *Trends in Microbiology*, 20(1), pp. 50–57. Available at: <https://doi.org/10.1016/j.tim.2011.11.002>.
- Miranda, T.B. and Jones, P.A. (2007) 'DNA methylation: The nuts and bolts of repression', *Journal of Cellular Physiology*, 213(2), pp. 384–390. Available at: <https://doi.org/10.1002/jcp.21224>.
- Moore, J.E. *et al.* (2020) 'Expanded encyclopaedias of DNA elements in the human and mouse genomes', *Nature*, 583(7818), pp. 699–710. Available at: <https://doi.org/10.1038/s41586-020-2493-4>.
- Mota-Rojas, D. *et al.* (2021) 'Efficacy and Function of Feathers, Hair, and Glabrous Skin in the Thermoregulation Strategies of Domestic Animals', *Animals*, 11(12), p. 3472. Available at: <https://doi.org/10.3390/ani11123472>.
- Muhr, J. and Ackerman, K.M. (2023) 'Embryology, Gastrulation', in *StatPearls*. Treasure Island (FL): StatPearls Publishing. Available at: <http://www.ncbi.nlm.nih.gov/books/NBK554394/> (Accessed: 3 May 2023).
- Mukund, K. and Subramaniam, S. (2020) 'Skeletal muscle: A review of molecular structure and function, in health and disease', *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 12(1), p. e1462. Available at: <https://doi.org/10.1002/wsbm.1462>.
- Nasrullah *et al.* (2022) 'DNA methylation across the tree of life, from micro to macro-organism', *Bioengineered*, 13(1), pp. 1666–1685. Available at: <https://doi.org/10.1080/21655979.2021.2014387>.

1 General Introduction

- Nowotschin, S., Hadjantonakis, A.-K. and Campbell, K. (2019) 'The endoderm: a divergent cell lineage with many commonalities', *Development*, 146(11), p. dev150920. Available at: <https://doi.org/10.1242/dev.150920>.
- Ogbourne, S. and Antalis, T.M. (1998) 'Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes.', *Biochemical Journal*, 331(Pt 1), pp. 1–14.
- Ozsolak, F. and Milos, P.M. (2011) 'RNA sequencing: advances, challenges and opportunities', *Nature Reviews Genetics*, 12(2), pp. 87–98. Available at: <https://doi.org/10.1038/nrg2934>.
- Pai, A.A., Pritchard, J.K. and Gilad, Y. (2015) 'The Genetic and Mechanistic Basis for Variation in Gene Regulation', *PLOS Genetics*, 11(1), p. e1004857. Available at: <https://doi.org/10.1371/journal.pgen.1004857>.
- Pal, S. and Tyler, J.K. (2016) 'Epigenetics and aging', *Science Advances*, 2(7), p. e1600584. Available at: <https://doi.org/10.1126/sciadv.1600584>.
- Pan, Z. *et al.* (2021) 'Pig genome functional annotation enhances the biological interpretation of complex traits and human disease', *Nature Communications*, 12(1), p. 5848. Available at: <https://doi.org/10.1038/s41467-021-26153-7>.
- Pan, Z. *et al.* (2023) 'An atlas of regulatory elements in chicken: A resource for chicken genetics and genomics', *Science Advances*, 9(18), p. eade1204. Available at: <https://doi.org/10.1126/sciadv.ade1204>.
- Panigrahi, A. and O'Malley, B.W. (2021) 'Mechanisms of enhancer action: the known and the unknown', *Genome Biology*, 22(1), p. 108. Available at: <https://doi.org/10.1186/s13059-021-02322-1>.
- Park, P.J. (2009) 'ChIP–seq: advantages and challenges of a maturing technology', *Nature Reviews Genetics*, 10(10), pp. 669–680. Available at: <https://doi.org/10.1038/nrg2641>.
- Patil, V., Ward, R.L. and Hesson, L.B. (2014) 'The evidence for functional non-CpG methylation in mammalian cells', *Epigenetics*, 9(6), pp. 823–828. Available at: <https://doi.org/10.4161/epi.28741>.
- Pazin, M.J. and Kadonaga, J.T. (1997) 'What's Up and Down with Histone Deacetylation and Transcription?', *Cell*, 89(3), pp. 325–328. Available at: [https://doi.org/10.1016/S0092-8674\(00\)80211-1](https://doi.org/10.1016/S0092-8674(00)80211-1).
- Pennisi, E. (2012) 'ENCODE Project Writes Eulogy for Junk DNA', *Science*, 337(6099), pp. 1159–1161. Available at: <https://doi.org/10.1126/science.337.6099.1159>.
- Phillips, J.E. and Corces, V.G. (2009) 'CTCF: Master Weaver of the Genome', *Cell*, 137(7), pp. 1194–1211. Available at: <https://doi.org/10.1016/j.cell.2009.06.001>.
- Ratnakaran, A. *et al.* (2016) 'Behavioral Responses to Livestock Adaptation to Heat Stress Challenges', *Asian Journal of Animal Sciences*, 11, pp. 1–13. Available at: <https://doi.org/10.3923/ajas.2017.1.13>.

- Riethoven, J.-J.M. (2010) 'Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators', in I. Ladunga (ed.) *Computational Biology of Transcription Factor Binding*. Totowa, NJ: Humana Press (Methods in Molecular Biology), pp. 33–42. Available at: https://doi.org/10.1007/978-1-60761-854-6_3.
- van Ruiten, M.S. and Rowland, B.D. (2021) 'On the choreography of genome folding: A grand pas de deux of cohesin and CTCF', *Current Opinion in Cell Biology*, 70, pp. 84–90. Available at: <https://doi.org/10.1016/j.ceb.2020.12.001>.
- Satterlee, J.S. *et al.* (2019) 'The NIH Common Fund/Roadmap Epigenomics Program: Successes of a comprehensive consortium', *Science Advances*, 5(7), p. eaaw6507. Available at: <https://doi.org/10.1126/sciadv.aaw6507>.
- Smallwood, A. and Ren, B. (2013) 'Genome organization and long-range regulation of gene expression by enhancers', *Current Opinion in Cell Biology*, 25(3), pp. 387–394. Available at: <https://doi.org/10.1016/j.ceb.2013.02.005>.
- Snyder, M.P. *et al.* (2020) 'Perspectives on ENCODE', *Nature*, 583(7818), pp. 693–698. Available at: <https://doi.org/10.1038/s41586-020-2449-8>.
- Solnica-Krezel, L. and Sepich, D.S. (2012) 'Gastrulation: Making and Shaping Germ Layers', *Annual Review of Cell and Developmental Biology*, 28(1), pp. 687–717. Available at: <https://doi.org/10.1146/annurev-cellbio-092910-154043>.
- Spence, J.R., Lauf, R. and Shroyer, N.F. (2011) 'Vertebrate intestinal endoderm development', *Developmental Dynamics*, 240(3), pp. 501–520. Available at: <https://doi.org/10.1002/dvdy.22540>.
- The FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* (2014) 'A promoter-level mammalian expression atlas', *Nature*, 507(7493), pp. 462–470. Available at: <https://doi.org/10.1038/nature13182>.
- Tixier-Boichard, M., Bed'hom, B. and Rognon, X. (2011) 'Chicken domestication: From archeology to genomics', *Comptes Rendus Biologies*, 334(3), pp. 197–204. Available at: <https://doi.org/10.1016/j.crv.2010.12.012>.
- Tragante, V., Moore, J.H. and Asselbergs, F.W. (2014) 'The ENCODE Project and Perspectives on Pathways', *Genetic Epidemiology*, 38(4), pp. 275–280. Available at: <https://doi.org/10.1002/gepi.21802>.
- van der Velde, A. *et al.* (2021) 'Annotation of chromatin states in 66 complete mouse epigenomes during development', *Communications Biology*, 4(1), pp. 1–15. Available at: <https://doi.org/10.1038/s42003-021-01756-4>.
- Villota-Salazar, N.A., Mendoza-Mendoza, A. and González-Prieto, J.M. (2016) 'Epigenetics: from the past to the present', *Frontiers in Life Science*, 9(4), pp. 347–370. Available at: <https://doi.org/10.1080/21553769.2016.1249033>.
- Waddington, C.H. (1939) *An introduction to modern genetics*. New York: The MacMillan Company. Available at: <https://www.cabdirect.org/cabdirect/abstract/19390101030> (Accessed: 1 May 2023).

1 General Introduction

- Wang, K. *et al.* (2022) 'Epigenetic regulation of aging: implications for interventions of aging and diseases', *Signal Transduction and Targeted Therapy*, 7(1), pp. 1–22. Available at: <https://doi.org/10.1038/s41392-022-01211-8>.
- Wang, T. *et al.* (2021) 'Integrative Epigenome Map of the Normal Human Prostate Provides Insights Into Prostate Cancer Predisposition', *Frontiers in Cell and Developmental Biology*, 9. Available at: <https://www.frontiersin.org/articles/10.3389/fcell.2021.723676> (Accessed: 18 May 2023).
- Wang, Z. *et al.* (2021) 'The molecular basis for sarcomere organization in vertebrate skeletal muscle', *Cell*, 184(8), pp. 2135-2150.e13. Available at: <https://doi.org/10.1016/j.cell.2021.02.047>.
- Warr, A. *et al.* (2020) 'An improved pig reference genome sequence to enable pig genetics and genomics research', *GigaScience*, 9(6), p. g1aa051. Available at: <https://doi.org/10.1093/gigascience/g1aa051>.
- Warren, W.C. *et al.* (2016) 'A New Chicken Genome Assembly Provides Insight into Avian Genome Structure', *G3: Genes|Genomes|Genetics*, 7(1), pp. 109–117. Available at: <https://doi.org/10.1534/g3.116.035923>.
- Wolffe, A.P. and Hayes, J.J. (1999) 'Chromatin disruption and modification', *Nucleic Acids Research*, 27(3), pp. 711–720. Available at: <https://doi.org/10.1093/nar/27.3.711>.
- Xiang, H. *et al.* (2014) 'Early Holocene chicken domestication in northern China', *Proceedings of the National Academy of Sciences*, 111(49), pp. 17564–17569. Available at: <https://doi.org/10.1073/pnas.1411882111>.
- Yin, X. and Xu, Y. (2016) 'Structure and Function of TET Enzymes', in A. Jeltsch and R.Z. Jurkowska (eds) *DNA Methyltransferases - Role and Function*. Cham: Springer International Publishing (Advances in Experimental Medicine and Biology), pp. 275–302. Available at: https://doi.org/10.1007/978-3-319-43624-1_12.
- Zaugg, J.B. *et al.* (2022) 'Current challenges in understanding the role of enhancers in disease', *Nature Structural & Molecular Biology*, 29(12), pp. 1148–1158. Available at: <https://doi.org/10.1038/s41594-022-00896-3>.
- Zeng, Y. and Chen, T. (2019) 'DNA Methylation Reprogramming during Mammalian Development', *Genes*, 10(4), p. 257. Available at: <https://doi.org/10.3390/genes10040257>.
- Zhang, T., Cooper, S. and Brockdorff, N. (2015) 'The interplay of histone modifications – writers that read', *EMBO reports*, 16(11), pp. 1467–1481. Available at: <https://doi.org/10.15252/embr.201540945>.
- Zhang, Y. *et al.* (2022) 'Long-Distance Repression by Human Silencers: Chromatin Interactions and Phase Separation in Silencers', *Cells*, 11(9). Available at: <https://doi.org/10.3390/cells11091560>.
- Zhao, Y. and Garcia, B.A. (2015) 'Comprehensive Catalog of Currently Documented Histone Modifications', *Cold Spring Harbor Perspectives in Biology*, 7(9), p. a025064. Available at: <https://doi.org/10.1101/cshperspect.a025064>.

Ziller, M.J. *et al.* (2011) 'Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types', *PLOS Genetics*, 7(12), p. e1002389. Available at: <https://doi.org/10.1371/journal.pgen.1002389>.

Chapter 2

Detailed molecular and epigenetic characterization of the pig IPEC-J2 and chicken SL-29 cell lines

Jani de Vos¹, Richard P.M.A. Crooijmans¹, Martijn F.L. Derks¹, Susan L. Kloet², Bert Dibbits¹, Martien A.M. Groenen¹, Ole Madsen¹

¹ Animal Breeding and Genomics, Wageningen University & Research, Wageningen, 6708PB, The Netherlands

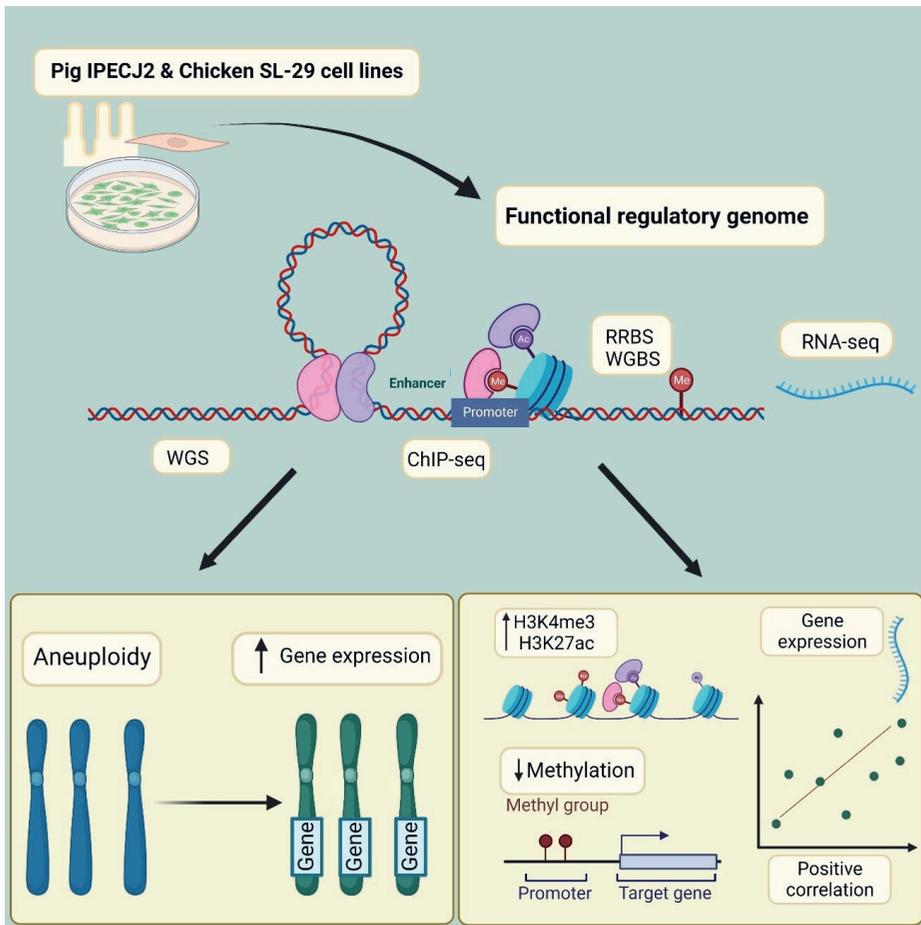
² Human Genetics, Leids Universitair Medisch Centrum, Leiden, 2333ZC, The Netherlands

iScience(2023) 26(3)



*'If you can dream it,
you can do it'*

Walt Disney



Highlights

- Untransformed, non-tumorigenic IPECJ2 and SL-29 cell lines show some aneuploidy.
- Aneuploid chromosomes result in increased gene expression.
- Positive correlation between hypo-methylation, H3K4me1, H3K27ac and gene expression.
- Identified enhancer and promoter regions play an important role as a reference.



Summary

The pig IPEC-J2 and chicken SL-29 cell lines are of interest because of their untransformed nature and wide use in functional studies. Molecular characterization of these cell lines is important to gain insight into possible molecular aberrations. The aim of this paper is to provide a molecular and epigenetic characterization of the IPEC-J2 and SL-29 cell lines, a cell-line reference for the FAANG community, and future biomedical research. Whole genome sequencing, gene expression, DNA methylation, chromatin accessibility, and ChIP-seq of four histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3) and an insulator (CTCF) are used to achieve these aims. Heteroploidy (aneuploidy) of various chromosomes was observed from whole genome sequencing analysis in both cell lines. Furthermore, higher gene expression for genes located on chromosomes with aneuploidy in comparison to diploid chromosomes was observed. Regulatory complexity of gene expression, DNA methylation, and chromatin accessibility was investigated through an integrative approach.

Introduction

The genome of all eukaryotic species is regulated at the chromosome level ¹, where DNA is packaged in a highly organized structure of DNA and histones. Gene expression is regulated through a network of physical interactions of enhancers, promoters, insulators, epigenetic marks, and chromatin binding factors, which is responsible for the chromatin accessibility. Epigenetic marks such as DNA methylation, non-coding RNAs and histone modifications can be investigated to obtain insight into regulation of the epigenome. Some histone modifications are highly informative regarding gene expression and are associated with transcriptional activation, promoters, and enhancers ²⁻⁵. In addition, DNA methylation is important in identifying gene expression and gene silencing as methylation and gene expression generally show an inverse correlation. Together these (epi)genetic marks can be used to annotate the functional genomic elements that determine gene expression.

The Functional Annotation of Animal Genomes (FAANG) consortium is a scientific driven community, with the aim of providing the functional annotation (functional maps) specifically for farm and companion animals ⁶. Earlier projects in human ⁷ and model animals (e.g. Mouse ENCODE) provided strategies for using omics data to obtain insights into the functional genome. This is achieved by performing genome-wide analysis focusing on genome expression, regulatory functions, methylation, chromatin accessibility and modifications providing insights into the functional genome.

Cell lines provide an interesting model to study the genomic architecture and regulatory genome of species of interest. Cell lines directly derived from tissues or organs of an animal

2 Molecular characterization of a chicken and pig cell line



are referred to as a primary cell line. Such cells can then either continue growing indefinitely or die off after a certain number of cell divisions⁸⁻¹⁰. Cell lines that can be grown indefinitely (i.e., that have become immortalized) often show cell aneuploidy or heteroploidy which is most pronounced in cancer cell lines^{11,12}. In this study a pig IPECJ2 and chicken SL-29 (CRL) cell lines were used. The pig IPECJ2 cell line is frequently used in e.g. intestinal transport studies due to the uniqueness of the cell line being neither transformed nor tumorigenic in nature. Chicken SL-29 (CRL) is useful for investigation of the substrate of virus propagation, recombinant protein expression and recombinant virus production¹³. Characterizing commonly used cell lines holds value for the FAANG community, where further comparative and/or combined studies will be performed. Determination of technical variation in data between different labs is important to identify as it will be useful in future comparative analyses in identifying differences due to biological variation. The main aim of the current research was the molecular characterization of the pig IPEC-J2 and chicken SL-29 cell lines using an integrative approach of a variety of omics data (genome sequencing, epigenomic modifications, DNA methylation and RNA-seq).

Results

The IPECJ2 cell line in pigs and the chicken SL-29 cell line are of interest for the animal genomics community because of the untransformed nature and wide use in functional studies in these cells. We have analyzed both the pig and the chicken cell lines with a range of whole genome based assays. We first report the results from the pig IPECJ2 cell line followed by the chicken SL-29 cell line. For each of these cell lines we first focus on chromosome level structural variation followed by an in-depth analysis of the expression, chromatin accessibility, and methylation of the cell line genomes.

Pig IPECJ-2 cell line

Chromosomal abnormalities and variations within the genome

The whole genome sequence data was investigated in different ways to determine the structure of the genome. Chromosomal abnormalities such as aneuploidy and heteroploidy can occur in cell lines which grow indefinitely. We therefore first investigated the chromosomal structures and possible changes in aneuploidy. Aneuploidy events can be detected using WGS data by examining both the read-depth and the ratio of reads that support the alternative allele for heterozygous SNPs. This should be around 0.5 for diploid chromosomes (figure 1d), while e.g. around 0.33 or 0.67 is expected for triploid chromosomes. Figure 1a provides an overview of the read depth and ratio of read support for heterozygous SNPs called from WGS data.

2 Molecular characterization of a chicken and pig cell line

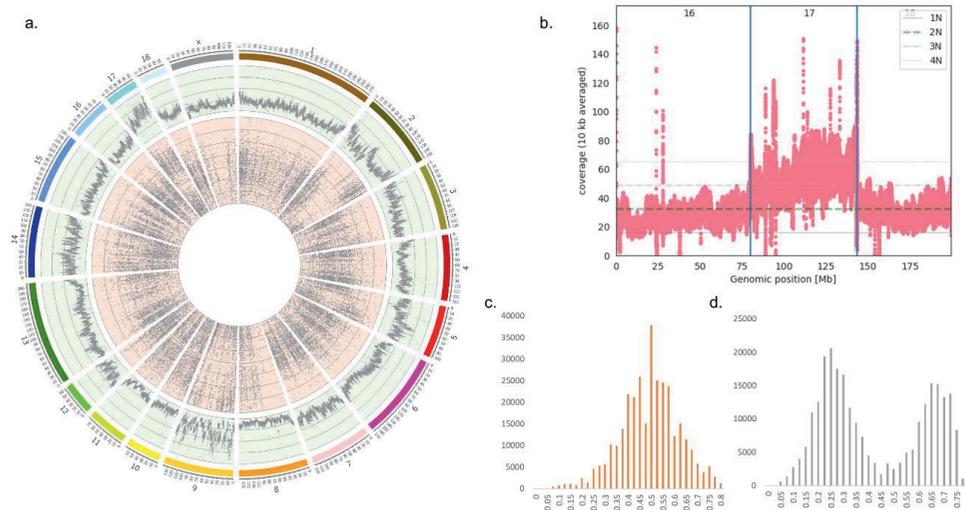


Figure 1. Chromosomal abnormalities in pig IPEC-J2

For a Figure360 author presentation of this figure, see

<https://www.sciencedirect.com/science/article/pii/S2589004223003292?via%3Dihub#mmc4>

(a) Circos plot showing the read-depth per chromosome in bins of 50 kb on the genome-wide level (outer track shown in green) and the allele support for heterozygous SNPs called from the WGS data per chromosome (inside track shown in red). The scale of the read-depth track starts from a minimum of 10 and increases in counts of 20 per line up to a maximum of 90. From the SNP distribution on the inner track of the plot a normal diploid chromosome will result in many heterozygous SNPs where both alleles are supported by 50% of the reads. However, a triploid chromosome would result in heterozygous SNPs supported by 33% or 67% of the reads. (b) Representation of WGS read-depth for chromosomes 16, 17, and 18, indicating triploidy of chromosome 17. (c and d) Histogram of the (ratio of the allele) count of reference allele for heterozygous SNPs of chromosome 16 (c) and chromosome 17 (d).

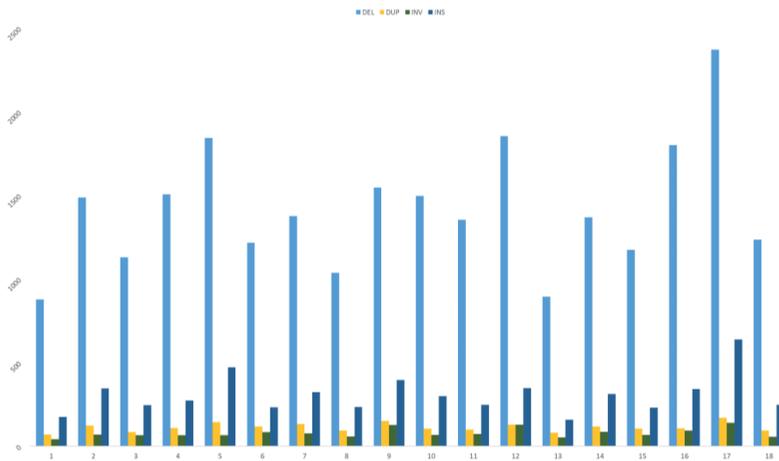
Some chromosomes, e.g. chromosome 17 (figure 1d), show clear evidence of aneuploidy, while other chromosomes contain large structural variations such as chromosome 2. This chromosome shows a diploid allele ratio distribution of 0.5 for the first part of the chromosome and a distribution of the SNPs around 0.75 and 0.25 towards the end of this chromosome, indicating possible triploidy for this segment of the chromosome. This possible triploidy is less clear from the read depth for this part of this chromosome. A low number of heterozygous SNPs is observed for the first half of chromosome 8 (which could indicate partial monosomy),

2 Molecular characterization of a chicken and pig cell line



while the second half of chromosome 8 has SNP support ratios at around 0.25 and 0.75, supporting higher ploidy. However, this observation on chromosome 8 is not very well reflected in the read depth for this chromosome. On chromosomes 9 and 17 the read depth is significantly higher compared to other chromosomes, indicating higher ploidy levels of these chromosomes. This is supported by the allele ratio distributions of these two chromosomes, where most SNPs are observed at a frequency of 0.25-0.35 and 0.65-0.75 indicating possible triploidy. Much variation in read-depth is observed for the individual chromosomes (figure S3), specifically for chromosome 16 (figure 1b) showing a likely deletion between position 9-17 Mb.

a.



b.

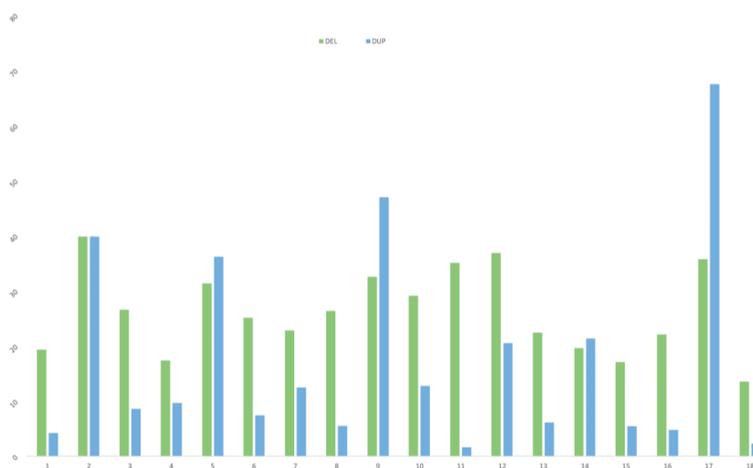


Figure 2. Structural variations observed in pig IPEC-J2 cell line

2 Molecular characterization of a chicken and pig cell line



Relative number of SVs per chromosome. Data normalized for size of chromosomes, with normalized counts shown on the y axis, detected by (a). Manta and (b) CNVnator. DEL, deletion; DUP, duplication; INV, inversion; and INS, insertion.

We further investigated additional structural variants within the genome of this cell line with Manta (Figure 2a) for the detection of small variants and CNVnator for large variants (> 1Mbp) (Figure 2b).

For small variants, deletions are the most abundant type of variant for all chromosomes and relatively more insertions and deletions are observed on chromosome 17. For large variants, the number of deletions and duplications varies between the chromosomes and the chromosomes showing evidence of triploidy (figure 1) have a relatively large number of duplications. We further investigated the effects of the structural variation, i.e. variants potentially affecting genes or regulatory regions using the VEP tool. As expected, most structural variants (small and large variants) are found within intron and intergenic regions. The most prominent effect was observed for large variants with 25% effecting transcription amplification and 17% transcription ablation. Results for both large SV and CNV, variant effect prediction analyses are shown in the supplementary figures S4 and S5.

Gene expression profile

RNA-seq data provides insight into gene expression levels across the genome. This can provide insight into elements that regulate gene expression like promoters, enhancers, and epigenetic marks, as well as chromosomal abnormalities affecting them. Of the 31,907 genes tested, 10,412 were expressed (TPM > 1).

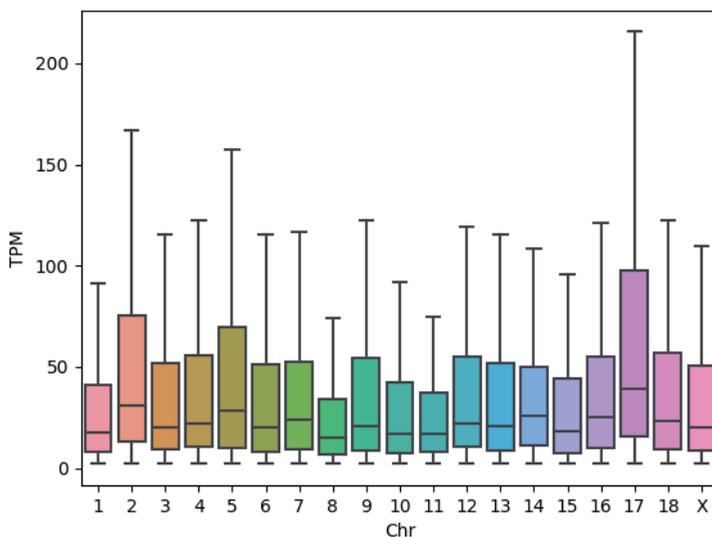


Figure 3. Gene expression profile of the pig IPEC-J2 cell line

Boxplots of the TPM expression values of genes per chromosome. Only genes with TPM>2 were used, which removes genes with very low/no expression.

Interestingly, the expression levels on chromosomes 2, 5 and 17 (all chromosomes with ploidy aberration or large structural changes), were higher compared to the diploid chromosomes. Gene expression levels per chromosome in the IPECJ2 cell line were compared to the gene expression levels for jejunum tissue and jejunum derived organoids and IPECJ2 samples cultured for a longer time (van der Hee *et al.*, 2020 figure S6). The gene expression in IPECJ2 cells seems to be more variable between chromosomes when compared to gene expression in jejunum tissue and the organoids derived of the jejunum tissue. A clear elevated gene expression is detected on chromosome 17 in the IPECJ2 cells compared to tissue and organoids and a notably high gene expression is observed on chromosome 2 for all samples.

Chromatin accessibility and genome architecture

Genome wide profiling of histone modifications provides insight into chromatin structure, as well as the location of regulatory elements such as promoter and enhancer regions. Therefore, ChIP-seq data was generated and analysed for four different histone modifications (table 1) and the insulator CTCF to provide insight into the regulatory genome of the IPECJ2 cell line.

2 Molecular characterization of a chicken and pig cell line



Furthermore, to assess the conformity between ChIP-seq experiments, ChIP-seq for three marks were repeated in another laboratory. The results of peak calling from the two respective experiments are shown in table 1.

Table 1 Number of ChIP-seq reads, peaks identified per mark and overlap between rounds of experiments.

Histone modification	Round1 # Reads	Round2 # Reads	Read coverage correlation of overlapping regions	# Peaks Round1	# Peaks Round2	Overlap	Signal value correlation of overlapping peaks
H3K4me1 ^b	27080156	32321068	0.52	99735	59260	54607	0.1
H3K4me3 ^a	18423318	28552134	0.9	22074	24448	16759	0.62
H3K27ac ^a	25285609	20466802	0.8	52584	48000	45199	0.3
H3K27me3 ^b	26873342	na	na	46470	na	na	na
CTCF ^a	34767672	na	na	7555	na	na	na

^a Narrow peak, ^b Broad peak, na - not analysed

A high number of peaks was shared between the two experiments (70-95%), with histone mark H3K4me3 sharing the lowest number of peaks most likely resulting from the lower number of reads for this mark in the first experiment. We assessed the relationship of the shared peaks and investigated the similarities between the two experiments, by calculating Pearson correlations (table 1). A high correlation is seen for the read coverage of overlapping peak regions for H3K4me3 and H3K27ac. Correlations between the signal values (measure of overall enrichment of the region) of overlapping peak regions for each experiment are low for H3K4me1 and H3K27ac, and moderate for H3K4me3. Histone modifications enriched around the TSS (+/- 1000bp) are generally indicative of promoter activity. Histone marks H3K4me3 and H3K27ac are enriched around the TSS of expressed genes (TPM>1) (figure S7), with histone mark H3K4me3 enriched at approximately ~17,5% of the TSS and histone mark H3K27ac enriched at ~5% of the TSS.

In ChromHMM analyses, the active promoter state is identified by an enrichment of histone marks H3K4me3, H3K27ac and H3K4me1 (Figure 4). The two enhancer states show an enrichment in H3K4me1 and H3K27ac. The histone mark H3K27me3 indicates the gene silencing state. Most of the genome is in a quiescent state (without any of the assayed histone marks; figure 4b right panel) while a small fraction is in the weakly repressive state. Both the

2 Molecular characterization of a chicken and pig cell line

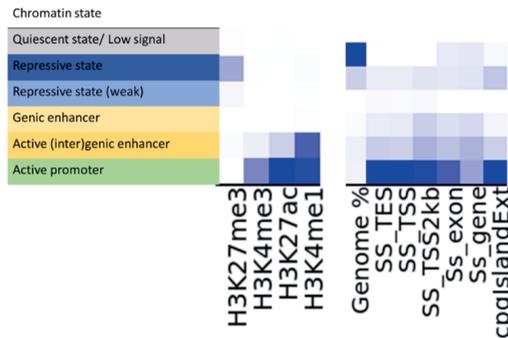


TSS as well as TSS +/- 2kb show enrichment in promoter and enhancer states (Figure 4b, right panel). The CpG islands show strong enrichment for the repressive state as well as the promoter state. The TSS is highly enriched for the active promoter and enhancer states. Figure 4c shows a typical example of the genome distribution of the histone marks. A strong enrichment of H3K27me3, H3K4me1 and H3K27ac can be seen around the *MESD* and *TLNRD1* TSS (see figure S8 for read alignments in this region). Further downstream from these genes, enrichment of the H3K4me3 mark can be observed in a gene poor region (gene desert). Annotation of the peak regions identified for each histone modification provides insight into the types of functional elements close to the histone modifications (figure S9). The annotation of these peak regions shows distribution of histone modification peaks in different genomic regions, which confirm the enrichment of H3K4me3 and H3K27ac around the promoter region.

a.

Mark	Enrichment	Type
H3K4me3	Promoter	Active
H3K4me1	Promoter / Enhancer distal elements	Active
H3K27ac	Enhancer/ Promoter	Active
H3K27me3	Gene silencing	Repressive

b.



c.

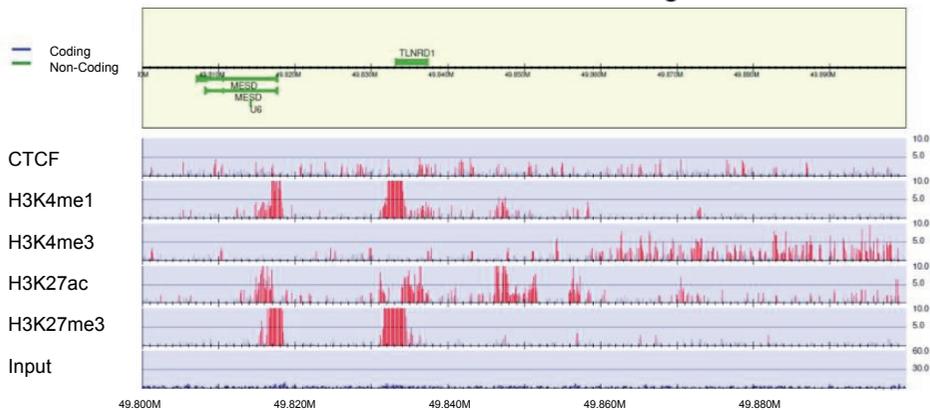




Figure 4. Histone modifications in pig IPEC-J2 cell line

(a) Summary of the informativity of the histone marks. (b) Six chromatin states were defined using 4 histone modifications (H3K4me1, H3K27me3, H3K4me3, and H3K27ac), with the left panel describing the chromatin state annotations, central panel showing the emission coefficients in the ChromHMM model, and the right panel showing the relative enrichment of coverage for the whole genome (genome%), transcription start site (SS_TSS), transcription end site (SS_TES), 2000 bp surrounding the TSS (SS_TSS2kb), exon regions (Ss_exon), genic regions (Ss_gene), and CpG islands. (c) Individual histone modification and CTCF profiles in the IPEC-J2 cell line for the *MESD* (TPM = 110.85) and *TLNRD1* (TPM = 42.63) genes (both involved in mesodermal development) on chromosome 7.

Overrepresented sequences motif analysis for histone marks associated with promoter and enhancer regions identify types of transcriptions factors (TF) related to gene expression. The peaks called for H3K4me3, H3K27ac and H3K4me1 were provided to Homer for identification of enriched TF involved in gene expression in IPECJ2. Table2 shows the three most enriched motifs and their corresponding TF identified for three histone marks associated with promoter and enhancer regions. Other motifs enriched (motifs with p-values <1e-12) are shown in supplementary results (Table S4).

Table 2 Motif enrichment in histone H3K4me3, H3K27ac, H3K4me1 peak regions. Top three enriched known binding motifs identified from consensus peaks.

Histone mark	Motif	Transcription factor	% of target regions	P-value
H3K4me3		PB0075.1_Sp100_1	54.95	1e-316
		IRF2(IRF)	4.45.	1e-287
		SD0001.1_at_AC_acceptor	68.6	1e-220
H3K27ac		E2F	49.69	1e-141
		ZNF449	45.56%	1e-120



		REL	41.29%	1e-111
H3K4me1		Fos(bZIP)	8.24%	1e-305
		TEAD(TEA)	20.48%	1e-188
		HOXC13	50.85%	1e-120

The significant transcription factors identified, were Sp100 and IRF for the peak regions of H3K4me3, E2F and ZNF449 for H3K27ac, and FOS and TEAD for H3K4me1. We identified 36,638 enhancer regions and motif analysis for these enhancers are shown in supplementary table S4. Significant TFs identified within the enhancer regions include Fra1, TEAD3, EWS-ERG fusion, FOXN3, which are responsible for cell growth, tumour suppression, suppression of transcription of transforming growth factor and play a role in several cancers. The CTCF motif generated using both MEME and homer tools (Figure S10), shows high similarity with the human consensus sequence supporting the conserved nature of CTCF binding sites.

DNA Methylation profile of the genome

Gene expression is negatively correlated to DNA methylation, therefore investigating the methylome provides further insight into specific characteristics of the cell line's genome characteristics. The methylome was investigated using both RRBS and WGBS (Table 3).



Table 3 Average methylation levels for cytosine in different base content between RRBS and WGBS for pig IPECJ2 cell line.

Site	Assay	Average methylation level (%)
CpG	RRBS	42.76
CHG		0.6
CHH		0.53
CpG	WGBS	45.6
CHG		0.35
CHH		0.33

Most DNA methylation is observed at CpG sites (42.76 - 45.6%, table 3) while at non-CpG sites little DNA methylation is detected (0.68 -1.13%). Therefore, as expected, the cell line displays a methylation pattern similar to what has been observed in porcine tissue methylation studies (Schachtschneider *et al.*, 2015). The average coverage for the RRBS methylation of chromosomes is 14.2 and most chromosomes are similarly covered (figure S11), except for chromosome 17 which has a higher read coverage of 22.4, confirming the triploid status of this chromosome also supported by the RRBS data. A global view of WGBS CpG methylation levels per chromosome shows methylation levels to fluctuate between 0.3-0.6 (figure S12).

WGBS is referred to as the gold standard for investigating DNA methylation, in particular because it provides more information on a whole genome level compared to RRBS. RRBS is usually perceived as being more cost-effective, as a reduced number of sites are sequenced (usually more focus on CpG regions). We therefore investigated whether all sites identified by RRBS are also covered by WGBS. To do so, we disregarded any sites with a coverage <10 for both types of data as not being sufficiently informative. In total, 23,952 sites covered by RRBS are not covered by WGBS. Further details on possible functional relevance of these RRBS specific sites is done for the chicken cell line (see below).

Integrative insight into epigenome marks

The interactions between regulatory elements and methylation play a critical role in gene expression. We studied these complex interactions and the relation to gene expression using an integrative approach. The individual relationships between the regulation of gene

2 Molecular characterization of a chicken and pig cell line



expression and histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3) and CTCF are shown in figure 5a.

The Histone modifications associated with promoter and enhancer regions (H3K4me3, H3K4me1 and H3K27ac) show patterns as expected with a positively correlation with gene expression. Histone mark H3K27me3, which is associated with gene suppression, shows a negative correlation with gene expression (Figure 5b). As expected, there is a negative correlation for the methylation data at the TSS with the promoter and enhancer regions (H3K4me3 and H3K27ac). The methylation levels within the gene-body show both a weaker negative correlation to CTCF and H3K27me3, and a positive correlation (WGBS_GB) with H3K4me3 and H3K27ac. A positive correlation is also observed between gene expression and the histone marks H3K4me3 and H3K27ac.

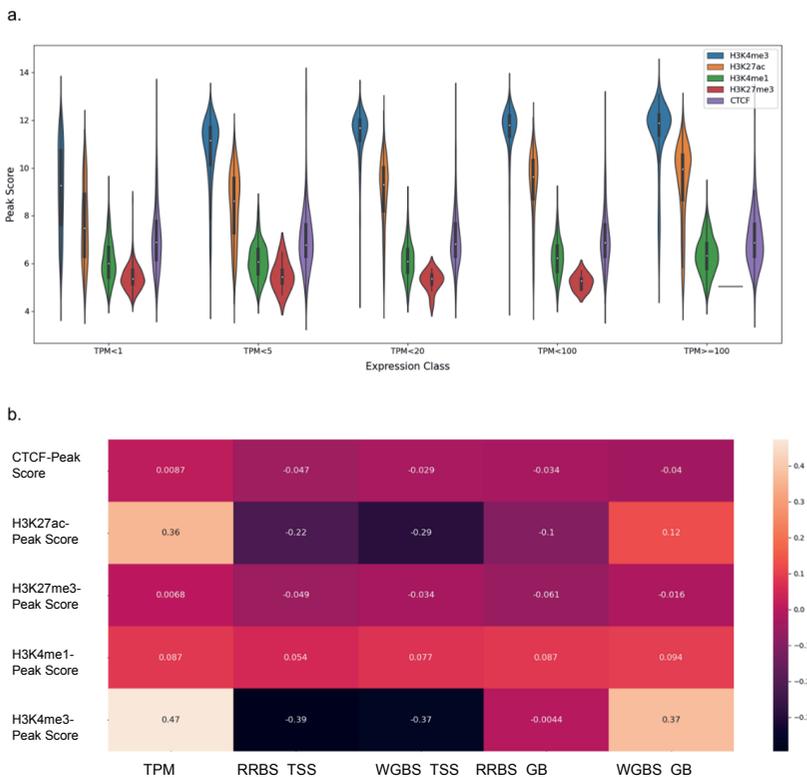


Figure 5. Integrative insight into histone marks and DNA methylation together with expression in pig

IPEC-J2 Integrative analysis of various histone marks and DNA methylation together with gene expression. (a) Violin plots of the peak score for each histone mark relative to the TPM expression values. TPM values are divided into different classes ranging from very lowly

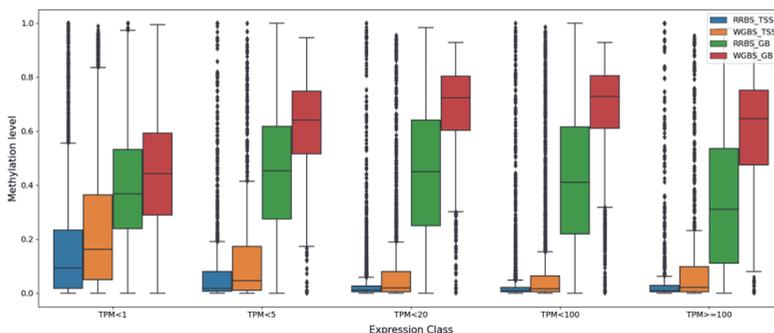
2 Molecular characterization of a chicken and pig cell line



expressed ($TPM > 1$) to very high expression values ($TPM > 100$). (b) Heatmap of the peak score values for each histone mark relative to the TPM expression values as well as methylation levels (RRBS and WGBS) at different positions. RRBS_TSS = RRBS level at the transcription start site (TSS), RRBS_GB = RRBS level within the gene-body (GB), and WGBS level at the same locations TSS and GB. Levels of correlations are shown by colour panel on the right, and the value of each correlation is also shown on the heatmap.

The methylation levels at the TSS and within the gene-body for RRBS and WGBS (figure 6a) follow the expectation that methylation levels are negatively correlated with gene expression (i.e. highly expressed genes show lower methylation levels and vice versa) (figure 6b). This relationship between methylation and expression values is especially evident at the TSS. Low methylation of RRBS within the gene body can likely be explained by a lack of coverage in the gene body compared to WGBS. In addition, for WGBS the methylation level in the gene-body increases at higher gene expression.

a.



b.

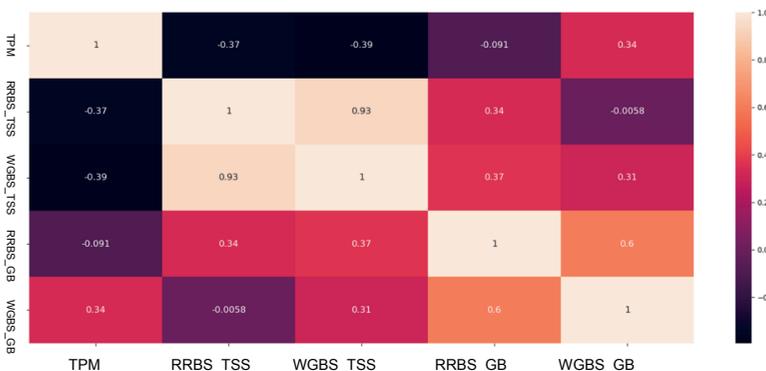




Figure 6. Integrative insight into DNA methylation together with expression in pigIPEC-J2

Box showing integrative analysis plots for methylation levels and expression levels. (a) Average methylation level at various regions (RRBS and WGBS data at the TSS and GB, notations are as above in Figure 5) for different levels of expression. (b) Heatmap of the various groups of methylation levels (RRBS and WGBS at the TSS and GB) together with expression.

Chicken SL-29 cell line

Chromosomal abnormalities and variants in the genome

As with the pig cell line, whole genome sequencing for the chicken cell line provided a comprehensive insight into genetic variation, chromosomal abnormalities, and structural variation at the global genome level. Multiple chromosomal abnormalities such as aneuploidy and heteroploidy are observed for the chicken SL-29 cell line (figure 7).

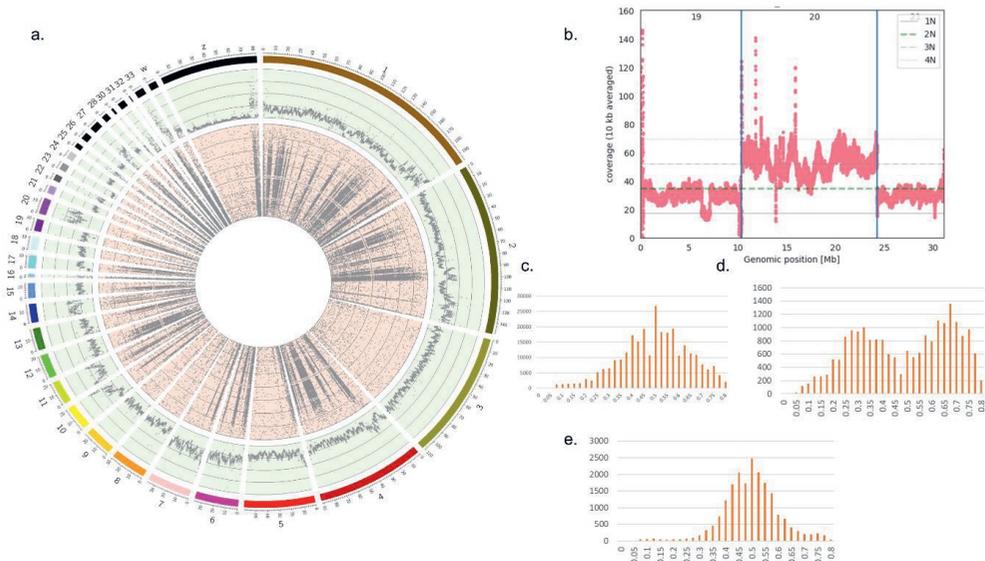


Figure 7. Chromosomal abnormalities in chicken SL-29

(a) Circos plot showing the read-depth per chromosome in bins of 50 kb on the genome-wide level (outer track shown in green) and the inside track (in red) showing the allele support for heterozygous SNPs called from the WGS data per chromosome. (b) Read-depth for individual chromosomes 19, 20, and 21, which show the higher read-depth in a tetraploid chromosome

2 Molecular characterization of a chicken and pig cell line



(chromosome 20) compared with chromosomes 19 and 21. (c–e) Examples of SNP distribution of a diploid. (c) Chromosome 1: it can be observed that many heterozygous SNPs are supported by 50% of the reads. (d) Chromosome 11: heterozygous SNPs supported by 33% or 67% of the reads. (e) Chromosome 20: many heterozygous SNPs are supported by 50% of the reads.

A higher read-depth (figure 7a, first outer line of the read-depth track) is seen for chromosomes 6, 7, 11, 14, 20, 27 and 33 in comparison with other chromosomes in the genome. Figures 7 c-e show the SNP distribution for chromosomes 1, 11 and 20, with chromosome 20 classified as a possible tetraploid. The region at 7-8 Mb on chromosome 19 shows low read depth (figure 7b) suggesting a possible large deletion at that position within this cell line. We assessed the presence of structural variants, in particular deletions and duplications, using Manta. Results show different small structural variants within the genome of this cell line (figure 8a). Several large copy number variants such as deletions and duplications of more than 1Kbp were detected using CNVnator, (figure 8b).

Relative abundance of copy number variants is higher for some of the small micro-chromosomes (i.e. chromosomes 16, 25, 30, 31, 32, 33), while the macro-chromosomes have relatively fewer copy number variants. Intron variants (48% of total variants) and intergenic variants (28% of total variants) are the most abundant consequences from the copy number variants in this cell line. The effects, as determined by VEP, of both copy number variants and large structural variants identified through CNVnator are shown in figures S13 and S14.

2 Molecular characterization of a chicken and pig cell line

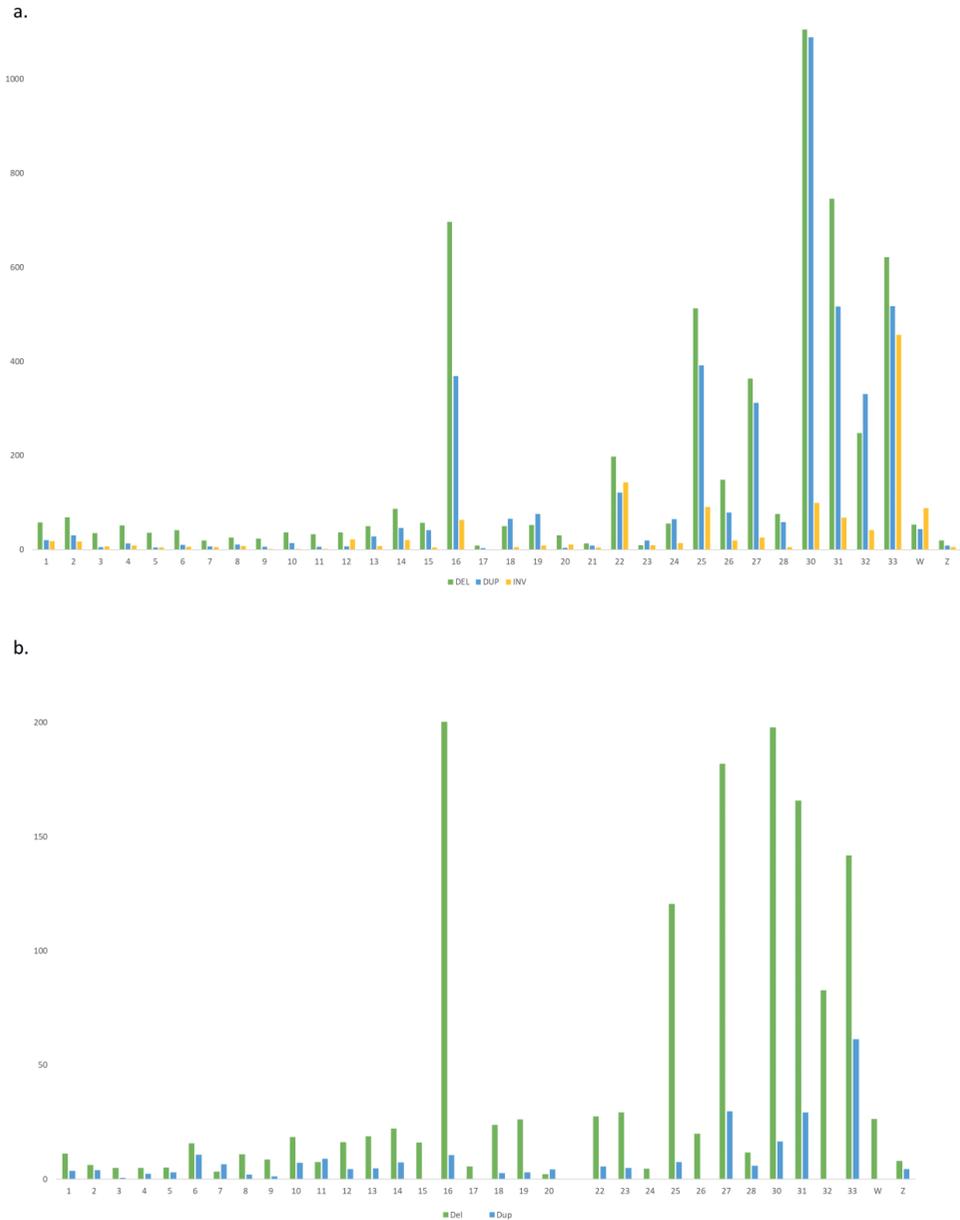


Figure 8. Structural variations observed in chicken SL-29 cell line

Normalized count of SVs per chromosome, with normalized counts on the y axis. (a) Manta for copy-number variants and (b) CNVnator for structural variants >1 kb.



Expression profile of the genome

The expression profile for this cell line was investigated to obtain further insight into the genes expressed, and interaction of regulatory elements, aneuploidy and CNV's affecting gene expression. We tested the expression of 24,356 genes, of which 13,546 were expressed (TPM >1).

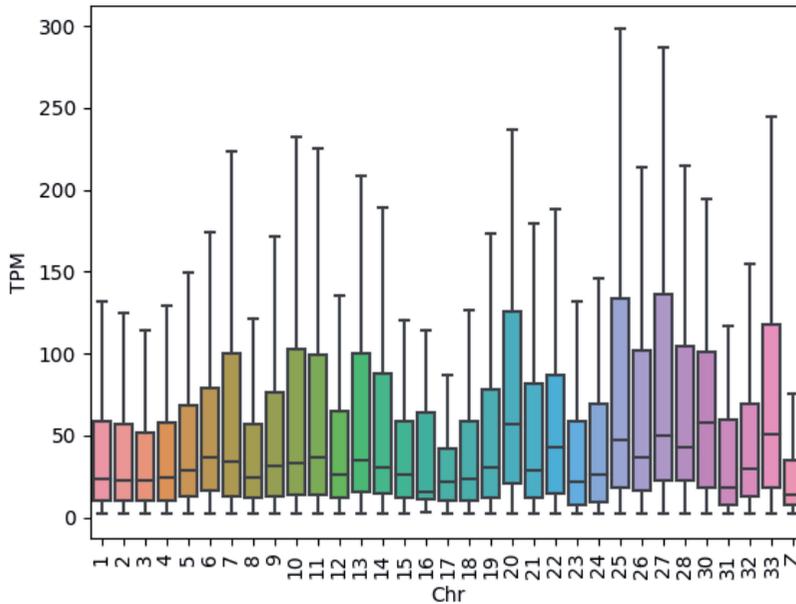


Figure 9. Boxplot of the expression levels of genes per chromosome

Higher levels of gene expression are observed on chromosomes 20, 25, 27, and 33 (Figure 9). This shows the influence of aneuploidy and structural variation on gene expression levels, e.g. with tetraploid chromosome 20 showing a higher expression level compared to the diploid chromosomes (e.g. chr 1).

Chromatin accessibility and genome architecture

As described for pigs, histone modifications are of importance to investigate chromatin accessibility and providing further insight into regulatory elements. The peak calling results for the chicken cell line are shown in table 4.

Table 4 Number of peaks identified per mark for the respective rounds of the experiments, overlap and number of reads for each round of experiment.

2 Molecular characterization of a chicken and pig cell line



Histone modification	Round1 # Reads	Round2 # Reads	Read coverage correlation of overlapping regions	# Peaks Round1	# Peaks Round2	Overlap	Signal value correlation of overlapping peaks
H3K4me1 ^b	36204354	33307742	0.90	46568	58599	42902	0.15
H3K4me3 ^a	51732192	28850772	0.92	17168	16376	12278	0.6
H3K27ac ^a	42046042	20967712	1.00	30157	55507	21974	0.3
H3K27me3 ^b	53134788	na	na	51652	na	na	na
CTCF ^a	86075418	na	na	44922	na	na	na

^a Narrow peak, ^b Broad peak, na - not analysed

The peaks called show much overlap between the replicates for H3K4me1, H3K4me3 and H3K27ac. The two ChIP-seq experiments were compared using Pearson correlations (table 4). High correlations between the read coverage of overlapping regions are observed, together with low correlations between the signal values from overlapping peaks for H3K27ac and H3K4me1. H3K4me3 shows a higher correlation of signal values of overlapping peaks.

2 Molecular characterization of a chicken and pig cell line

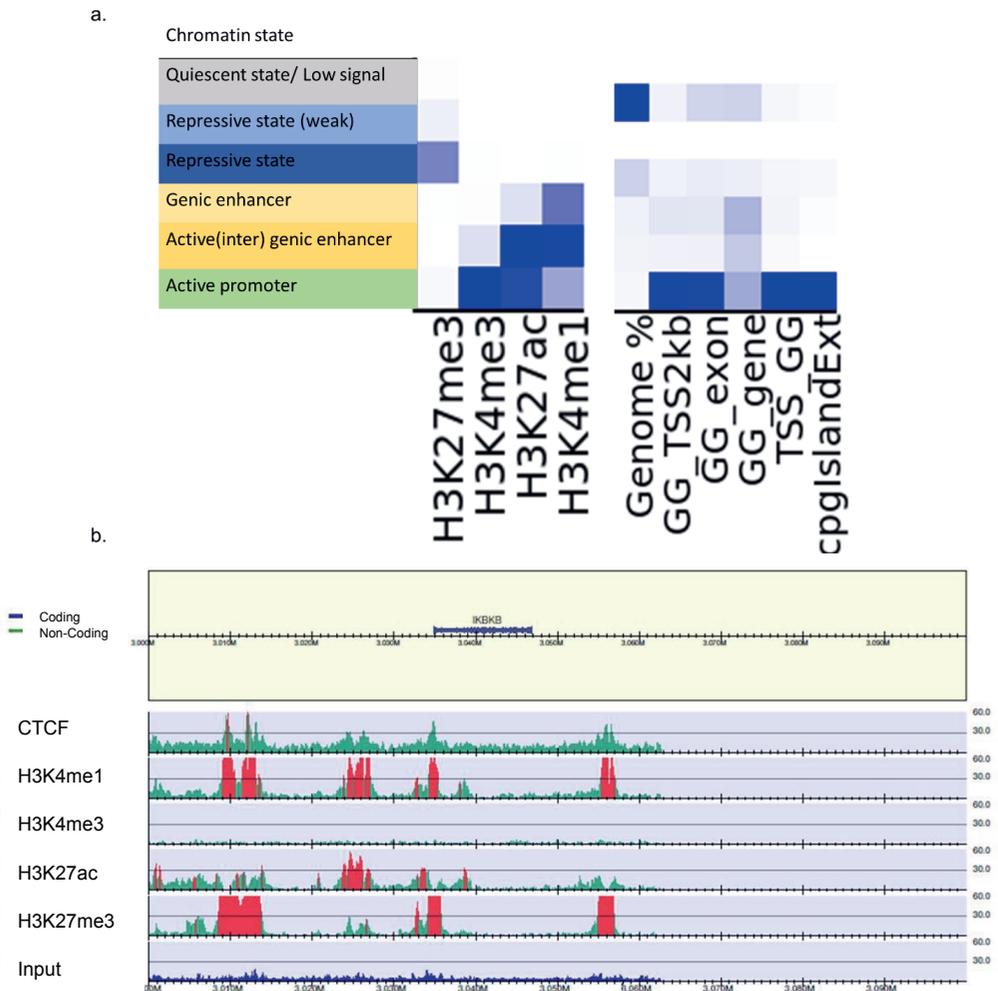


Figure 10. Histone modifications investigated for the chicken SL-29 cell line

(a) Six chromatin states were defined using the 4 histone modifications (H3K4me1, H3K27me3, H3K4me3, and H3K27ac), with the left panel describing the chromatin state annotations, central panel showing the emission coefficients in ChromHMM model, and the right panel showing the relative enrichment of coverage in whole genome (genome %) and in different genomic regions(transcription start site (GG_TSS), transcription end site (GG_TES), 2000bp surrounding the TSS (GG_TSS2kb), exon regions (GG_exon), genic regions (GG_gene), and CpG islands. (b) View of the individual histone modification and CTCF profiles in the SL-29 cell line for MICAL1 (TPM = 0.54), TULP1 (TPM = 0.03), and FKBP5 (TPM = 100.06) on chromosome 26.

2 Molecular characterization of a chicken and pig cell line



Various chromatin states were identified, through identification of presence or absence of histone marks using ChromHMM, which provides insight into interactions between different histone marks. Histone marks H3K4me3 and H3K27ac are enriched around the TSS of expressed genes, with H3K4me3 enriched at approximately 20% of the TSS and H3K27ac enriched at >5% of the TSS (figure S15). The chromatin dynamics (figure 10a) displays states 2 and 3 as repressed states due to the presence of H3K27me3. States 4, 5 and 6 display active dynamics because these are associated with the presence of H3K27ac/H3K4me1, H3K27ac/H3K4me3 and H3K4me1/H3K4me3. The states identified around 2 kb of the TSS are states 4, 5 and 6 with states 5 and 6 showing a very strong enrichment. State 6 is enriched within exons while states 1, 4, 5, and 6 are enriched within genes. State 6 also shows enrichment within CpG islands. A strong enrichment of H3K4me1 and H3K27ac, is seen for the expressed gene *FKBP5* (figure 10b), enrichments of H3K27me3 are seen around the lowly expressed genes (*MICAL1* and *TULP1*).

Annotation of the peak regions to genomic features provides insight into functional elements related to the histone modifications. As for the histone/CTCF marks in the pig cell line, most marks are found in intron and intergenic regions of the genome. Furthermore, a large percentage (16-21%) of the H3K4me3, CTCF and H3K27ac histone modifications are found within promoter regions (figure S16), and 11% of H3K4me3 has been identified in exon regions of the genome.

Table 5 Motif enrichment in histone marks peaks of H3K4me3, H3K27ac, H3K4me1. Top three enriched known binding motifs identified from consensus peaks. Further results shown in table S4 for motifs with p -values $< 1e-12$.

Histone mark	Motif	Transcription factor	% of target regions	P-value
H3K4me3		IRF1(IRF)	3.4	1e-232
		OVOL2	58.27	1e-137
		MSANTD3	28.26	1e-105
H3K27ac		ZNF711(Zf)	9.38	1e-86
		YY2	6.05	1e-78



		AP-2alpha(AP2)	13.78	1e-59
H3K4me1		Atf3(bZIP)	3.93	1e-89
		Pax8(Paired, Homeobox)	0.09	1e-42
		PSE(SNAPc)	0.07	1e-42

Significant consensus sequences for known motifs identified the TFs IRF1, ZNF711, YY2, AP-2alpha, Atf3, and Pax8 for the three histone marks (H3K4me3, H3K4me1 and H3K27ac) (table 5). In the H3K4me1 consensus sequence of the peak region a core promoter factor PSE (SNAPc) is observed. In this cell line a total of 18,516 enhancer regions were identified and the motif analysis is shown in table S5. Significant TFs identified within the enhancer regions include SMAD2:SMAD3, EWS-ERG fusion, TWIST1 and TEAD3 which play important roles in regulation of transcription in transforming growth factors and embryonic development, and which are associated with cancers. The motif sequence for CTCF identified with both homer and MEME (figure S17) is similar to the human consensus sequence, supporting the conservation of the CTCF binding site beyond mammals.

Genome wide chromatin accessibility

Chromatin accessibility was profiled in the chicken SL-29 cell line using ATAC-seq data, from which 86,983 peak regions were identified. To infer the functional significance of accessible regions that were identified, consensus peaks were characterized by genomic localization. Annotation of ATAC-seq showed most accessible (open) chromatin is found in the intron and intergenic regions of the genome, (figure S18), with 12% of accessible chromatin found at promoters (as define by TSS location). To interrogate the potential function of accessible regions (peaks), they were subjected to a consensus motif enrichment analysis.



Table 6 Consensus motif enrichment in predicted open chromatin regions. Top ten enriched known binding motifs identified from the consensus peaks.

Motif	Known transcription factor motif	% of target regions	P-value
	PAX6	1.74	1e-1706
	IRF6	1.78	1e-1220
	PB0201.1_Zfp281_2	1.62	1e-1184
	PRDM1	1.28	1e-1081
	PRDM15(Zf)	2.06	1e-697
	PB0152.1_Nkx3-1_2	6.81	1e-464
	SF1(NR)	7.81	1e-393
	Zac1(Zf)	11.27	1e-376
	ZBTB26	12.4	1e-335
	ZSCAN29	4.53	1e-329

Overall, consensus peaks identified for PAX6 recognition sites as most significant, with about 1.74% of accessible regions harbouring this consensus motif (Table 4). Roles of the TFs identified are as expected related to this cell line, e.g. PRDM1 TF which is involved in immunity, PRDM15 which regulates transcription of WNT and TFs involved in the MAPK-ERK signalling pathway which is related to pluripotency of a cell.

DNA Methylation profile

We also determined the average methylation levels for cytosine for the chicken cell line, calculated from both RRBS and WGBS data (table 7).

2 Molecular characterization of a chicken and pig cell line



Table 7 Average methylation levels for different sites between RRBS and WGBS for chicken SL-29 cell line

Site	Assay	Average methylation level (%)
CpG	RRBS	37.44
CHG		0.55
CHH		0.55
CpG	WGBS	59.66
CHG		0.98
CHH		1.05

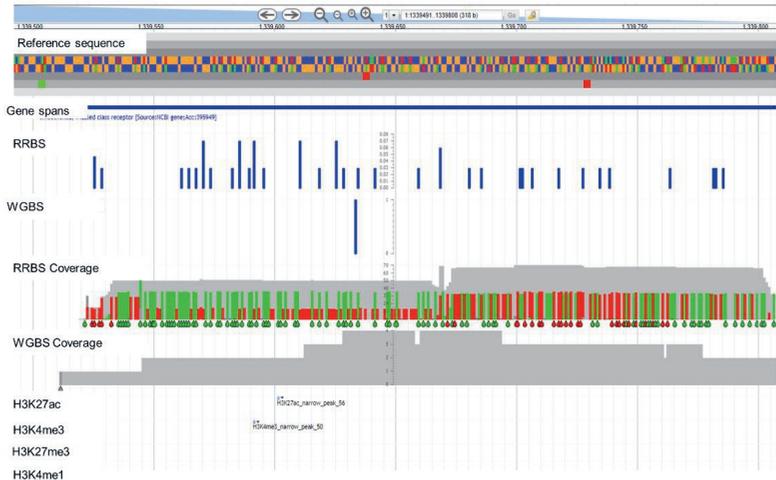
As expected, average methylation (>10 reads) is observed at CpG sites (37-59%) and slightly higher average non-CpG methylation (1-2%) is observed. Average chromosome level methylation levels estimated from WGBS fluctuate around 0.2-0.6 (figure S19). The average read coverage calculated for the WGBS data was 55.1, with chromosome 20 having a very high (~95) read coverage (figure S20).

As stated above for the pig, it is of importance to investigate sites covered by RRBS data while not covered by WGBS (and vice versa), as WGBS is considered as the gold standard for insight into whole genome DNA methylation. 926,495 sites were identified by RRBS and not by WGBS in the chicken cell-line (the total number of WGBS sites is 35,805,306 and for RRBS it is 2,830,991). We further investigated if these sites specifically covered by RRBS overlaid regions within predicted promoters and enhancer regions (from the ChIP-seq data). Examples of such regions (chromosome 1 and 2) covered by only RRBS data including predicted active enhancers/promoters are shown in figure 11.

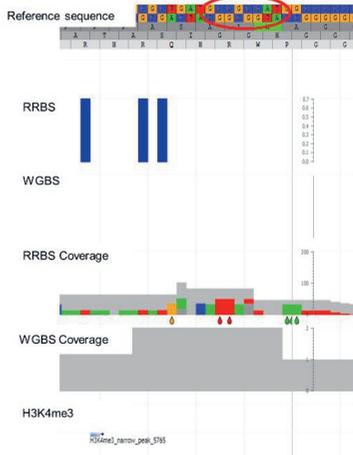
2 Molecular characterization of a chicken and pig cell line



a.



b.



c.



Figure 11. Sites covered by RRBS data while not covered by WGBS

(a) Example of a region on chromosome 1 (1339500-1339800) with high RRBS coverage (703) and lower than 10X WGBS coverage (%43). This region contains part of the *SMO* gene (TPM = 210.09), and the histone marks H3K4me3 and H3K27ac are identified and show a peak within this region, which indicates possible promoter/enhancer regions. The histone marks H3K4me1 and H3K27me3 are also displayed here but no peaks were observed within this region. The difference in RRBS and WGBS data coverage is very evident in this example, together with the presence of promoter and enhancer histone marks within a region that is well

2 Molecular characterization of a chicken and pig cell line



covered by RRBS data and sparsely covered by WGBS data. (b) Region on chromosome 2 (900,400-900425), with RRBS methylated sites (0.7) at a high coverage (833) and WGBS at a low coverage (2). The H3K4me3 mark is identified here together with (c) the most significant (p -value = $1e-55$) motif of identified enhancers (transcription factor YY2).

The transcription factor motif analysis of identified enhancers which overlap with the identified specific RRBS regions is shown in table S6. One of these motifs, for transcription factor YY2, has a strong CpG consensus sequence suggesting that these regions only covered by RRBS may include important regions involved in regulation of gene expression.

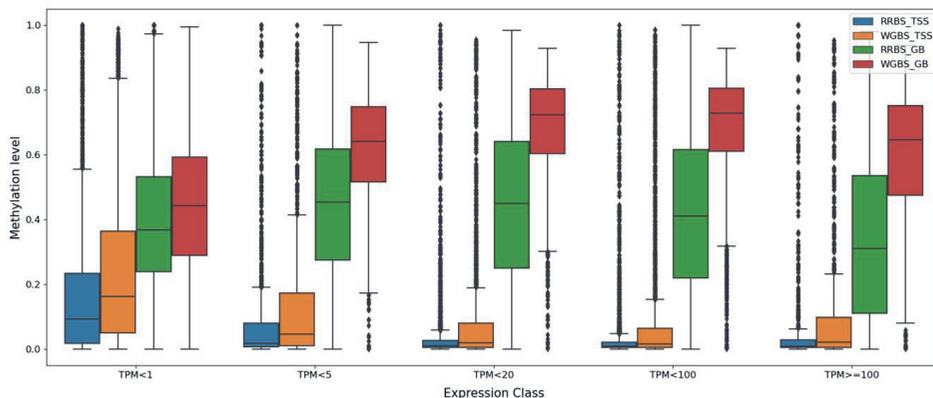
Integrative insight into epigenome marks

An integrative approach was used to gain insight into the dynamics of methylation and histone modifications for regulation of gene expression (figure 12). As expected, methylation levels are negatively correlated with gene expression (i.e. highly expressed genes show lower methylation levels and vice versa) at the TSS. Within the gene body (GB) we observe a slight increase in methylation levels with WGBS followed by a decrease with higher expression levels ($20 > \text{TPM} < 100$). The low methylation seen for RRBS within the gene body can be explained by a lack of coverage in the gene body compared to WGBS. Heatmap correlations reflect the results observed in the boxplots with negative correlations between methylation levels both at the TSS, and within the gene-body and gene expression.

2 Molecular characterization of a chicken and pig cell line



a.



b.

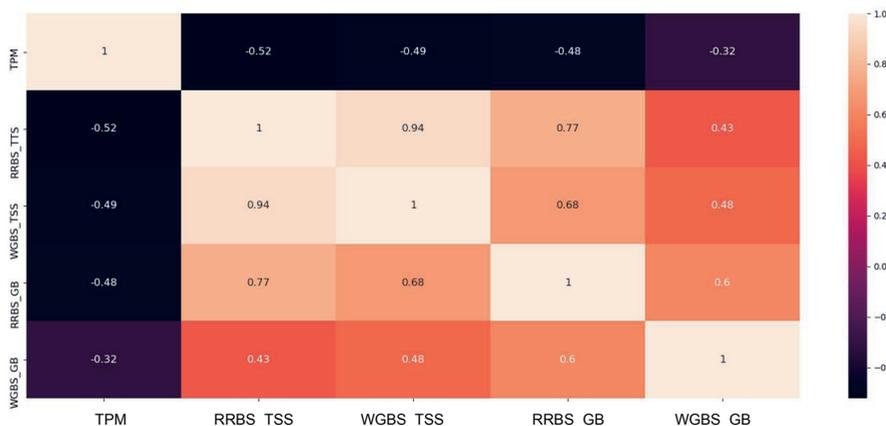


Figure 12. Integrative approach for investigation into regulation of gene expression by epigenomic marks in chicken SL-29

(a) Boxplots of the methylation levels at TSS and GB for RRBS and WGBS data across 5 classes of gene expression levels. (b) Heatmap of the correlations between methylation levels and TPM expression values.

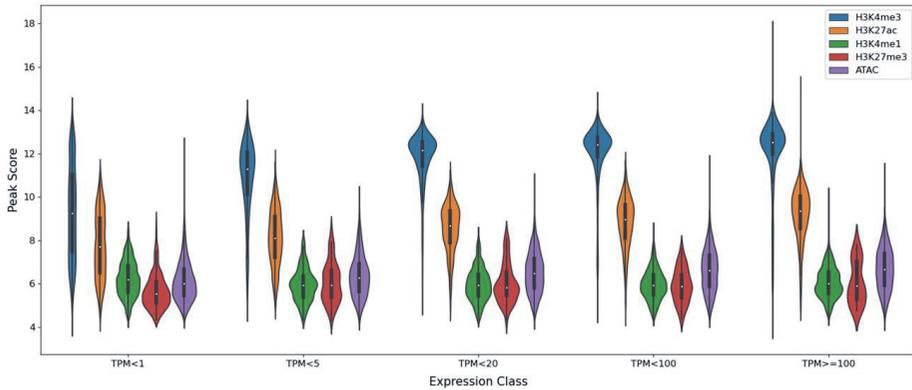
The correlations between the 4 histone marks (H3K4me1, H3K4me3, H3K27ac and H3K27me3), ATAC-seq and distinct classes of gene expression levels are visualised in figure 13. Enhancer histone marks H3K27ac and H3K4me3, together with ATAC-seq show an increase in peak score for genes with a higher gene expression level. H3K4me1 and H3K27me3 show little variation in peak scores across different classes of gene expression.

2 Molecular characterization of a chicken and pig cell line



For H3K27me3, slightly lower peak scores were observed for very highly expressed genes while higher peak score were seen for very lowly expressed genes.

a.



b.

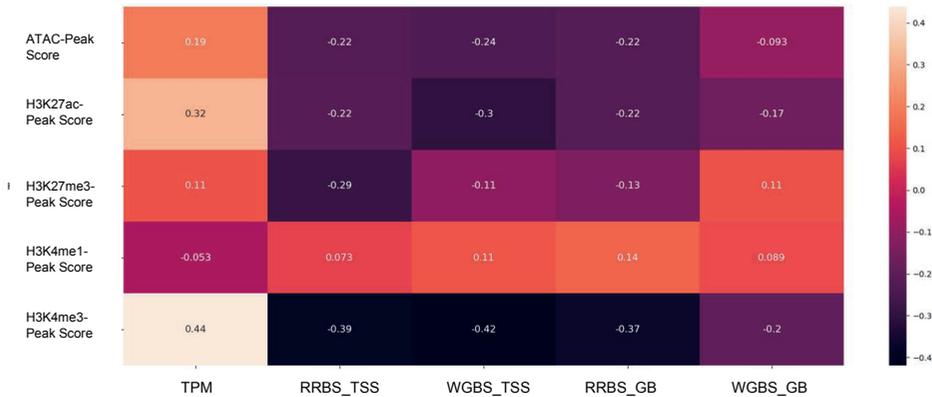


Figure 13. Integrative analysis of histone marks (H3K4me1, H3K4me3, H3K27ac and H3K27me3) and ATAC-seq with gene expression in chicken SL-29

(a) Violin plots of the relationship between the 4 histone marks, ATAC-seq, and 5 classes of gene expression levels (TPM). (b) Heatmap showing the correlations between the 4 histone marks, ATAC-seq, gene expression, and methylation levels of both RRBS and WGBS data at the TSS, as well as gene-body.

The positive correlation observed for H3K4me3 and H3K27ac with the gene expression (figure 13b) is higher compared with the methylation and gene expression results presented in figure 12b. Negative correlations are observed between methylation levels at both the TSS and gene-body for RRBS and WGBS with promoter and enhancer marks (H3K4me3, H3K27ac), as well



as with ATAC-seq. Low correlations between the histone marks H3K4me1 and H3K27me3 with methylation are observed. An example of a region with genes and all epigenomic modifications is shown in figure S21.

Discussion

Cell lines provide an ethical approach for research in animal production, and thus molecular characterisations are necessary for functionally relevant cell lines. The pig IPECJ2 and chicken SL-29 cell lines have never been characterized using an integrative genomics approach. These two cell lines were chosen specifically for their usability and application to research in animal production as well as biomedical research. Using different omics data (WGS, RNA-seq, ChIP-seq, ATAC-seq, RRBS, WGBS), the genome structure, transcriptome, methylome, and chromatin accessibility were investigated and characterized for these cell lines. This provides a reference of the genome architecture of these cell lines for future functional studies using these cell lines as well as for farm animal research community.

Results for both pig and chicken cell lines show that aneuploidy is common in both cell lines as we observed various chromosomes that were either (partly) monoploid, triploid, and even tetraploid. More aneuploidy as well as more structural variants were observed for the chicken SL-29 cell line in comparison to the pig IPECJ2 cell line. For SV this could be due to the additional filtering of common structural variants from the pig cell line. Previous studies have shown that more chromosomal abnormalities, as well as structural variants occur when cell lines are maintained in culture for a longer time (more passages) ^{14,15} emphasizing the importance of limiting the number of cell passages for cell line experiments. It has also been suggested that the culture conditions can influence the chromosomal stability ¹⁶. Conditions such as techniques for cell detachment and disaggregation, and oxygen concentration during culture can also affect the chromosomal stability and genomic integrity over a longer period of culture ¹⁷. To ensure that conditions do not influence the genomic integrity precisely defined protocols for cell culture should be followed as much as possible.

The increase in ploidy leads to an increase in expression of the genes on these chromosomes, likely affecting functional relevant aspects of these cell lines. A comparison of the results from the pig cell line to similar tissue type and organoid samples showed pronounced higher rates of expression of the genes on the triploid chromosome 17 of the pig cell line. From this comparison it is evident that these cell lines show higher rates of gene expression on all chromosomes, followed by organoids, and with tissue showing the lowest rates of gene expression. This is in agreement with studies showing that organoids resemble gene expression levels and physiology of tissue more closely than cell lines ^{18,19}. It has also been

2 Molecular characterization of a chicken and pig cell line



shown that aneuploidies and structural variants can influence gene expression level, specifically structural variants can cause changes in cis-regulatory elements, promoters, and enhancers²⁰⁻²². These expression observations in cell lines provide a useful resource for studies where potential genes of interest can be identified and investigated for increased expression levels.

The results from the WGS and RNA-seq show the potential of using these assays to detect chromosomal abnormalities, in addition to investigation of variation in the genome^{23,24} and gene expression respectively. Traditional methods such as karyotyping and staining (multifluor-FISH), are limited to detection of chromosomal abnormalities specifically chromosomal rearrangements such as translocations and provide little insight into genome variation²⁵. WGS as a tool for detection of pre- and post-natal anomalies is investigated and implemented more regularly^{26,27}. An example of using NGS (Next Generation Sequencing) for detection of pre-natal anomalies is discussed in Guseh, 2020, where trisomy 21 is detected when a higher proportion of DNA fragments are mapped to chromosome 21 in comparison to a reference²⁸. This shows the potential of WGS for detection of chromosomal abnormalities.

Chromatin accessibility and histone modifications were investigated to gain further insight into the genome architecture of both cell lines using ChIP-seq data for CTCF and four histone modifications: H3K4me1, H3K4me3, H3K27ac, H3K27me3. The importance of using a standardized protocol for comparative ChIP-seq studies was explored using two technical replicates of the ChIP-seq experiments for three histone marks, performed in two different laboratories and using different inputs as a background (IgG and DNA, respectively). This provides insight into the reproducibility of results between different laboratories and the use of DNA or IgG as an input. We conclude there is a high consensus between overlapping peaks of the experiments and secondly the read coverage between experiments showed medium to high consensus between experiments. However, signal values show little correlation between experiments suggesting that the confidence related to the high number of overlapping peaks is limited and should therefore be used with care if the signal values are used for comparative analysis. A reason for this low correlation of signal values could be the differences in background signal in the two experiments. ChIP-seq guidelines and practices from (mod)ENCODE have found that an IgG control mimics a ChIP experiment more closely than a DNA input control. Cases of strong sonication bias are rarely observed, but this can potentially affect peak calling²⁹. Thus, for comparative studies utilizing ChIP-sequencing similar protocol should be followed.

Peak regions identify possible binding sites of proteins associated with DNA (protein-chromatin interactions) and provide insight into regulatory regions and elements. The number of histone

2 Molecular characterization of a chicken and pig cell line



marks identified are different between tissues, stages of development and number of reads sequenced. ENCODE standards have shown that the number of peaks that can be identified ranges from thousands to tens of thousands of peaks^{30,31}. Coincidentally the number of peak regions called for narrow peaks (H3K4me3 and H3K27ac) for both pig and chicken cell lines were similar to previous studies in monogastric species^{32,33}. We confirmed the quality of the data by investigating the occurrences of the respective marks around 2 kb of the TSS. We observed an elevation of the marks H3K4me3 and H3K27ac around the TSS. It has been observed that H3K4me3 is most often found at TSS (regardless of H3K27ac) and indicative of a promoter, as seen in this study. H3K4me3 is often co-occupied by H3K27ac in the genome^{34–36}, however H3K27ac is not always found to be co-occupied by H3K4me3 and is also observed further from the TSS site. H3K4me1 and H2K27ac are indicative for enhancer regions³⁷. The number of broad peak regions of H3K27me3 for both cell lines is similar to that observed in other vertebrate species, with variations between tissues³⁸. This mark is associated with gene silencing, as shown in chromatin state analyses^{39,40}. Combinations of histone modifications result in variable chromatin structures leading to different levels of transcription, which is also reflected in the integrative analysis.

To evaluate the quality of the experiments and the success thereof we compared the consensus motif for the CTCF sites identified for pig and chicken to the human CTCF consensus sequence. There was good similarity between the consensus motifs for both pig and chicken to human. Furthermore, this motif was also identified in the human K562 cell line as a CTCF mark. CTCF is a highly conserved protein in mammals (between pig and mouse) as well as in vertebrates^{41–43}. This is indeed confirmed by our CTCF results, supporting the good quality of our results for CTCF.

Further investigation into motifs identified for the histone marks in both cell lines, together with comparison of similarities thereof to known motifs, provides insight into possible TFs. Firstly, interferon receptor factor (IRF) was identified in both cell lines (pig IRF2 and chicken IRF1). IRF is part of a TF family that is found in humans as super enhancer TFs and which is highly conserved within species⁴⁴. This TF plays a role in immunity, cell growth, differentiation and anti-tumor defences in vertebrates^{45–47}. TFs such as YY2, TEAD, E2F, which regulate cell growth and proliferation, as well as development were also identified^{48,49}.

Open chromatin was investigated using ATAC-seq data for the chicken cell line with a number of regions identified, slightly lower than expected when following ENCODE recommendations which suggests >100,000 peaks. However, this is similar to the number of enriched regions identified by other studies in animals^{50,51}. Most of the accessible chromatin was identified

2 Molecular characterization of a chicken and pig cell line



within intronic, intergenic, and promoter regions, which is in line with previous research in multiple species and suggests similar patterns between cell-lines and tissues ^{52,53}.

Further insight into gene expression and the molecular characteristics of both cell lines was obtained through characterization of the methylome, which ensures comprehensive characterization of the functional genome. DNA methylation is an epigenetic mark that is found in most species and has been found to be inherited and influenced by environmental factors (and often used for comparative analysis and fundamental research into e.g. adaptation ^{54,55}). In both cell lines a higher coverage in the methylation data (RRBS and WGBS) was observed on aneuploid chromosomes, relating to the change in copy number which is reflected in the methylome as well as on a whole genome level. Methylated cytosines occur primarily at CpG sites in most cell types while non-CpG methylation (CHG, CHH) occurs only in specific cell types such as brain, oocytes, and stem cells ⁵⁶. This is in agreement with our observation in our cell lines, where CpG methylation is the primary type of methylation. Additionally, the non-CpG methylation levels in the chicken cell line are slightly higher than observed in tissues of birds, excluding brain tissue ^{57,58}. This occurrence could be due to the chromosomal aberrations and higher ploidy observed and the cell lines behaving in a tumorigenic manner. Ploidy effects on DNA methylation (epigenome) has been theorized in studies in plants ^{59,60} and tumours ^{61,62}. We observed regions within the genome covered by RRBS that were not identified by WGBS or had a low coverage (<10 x) by WGBS. Some of these regions are highly relevant as they are located at the promoter of specific genes (close to the start site of the gene) or overlapping with enhancer elements. It is noteworthy that the coverage of the WGBS data is high (>50x) and therefore should theoretically cover all of the genome especially at informative sites ⁶³. This is a relevant observation as WGBS is often seen as the golden standard for investigating the methylome as it is supposed to cover almost all sites in contrast to RRBS which is seen as a cost-effective alternative method. RRBS usually shows reduced coverage of methylated sites in intergenic and distal regulatory regions, especially in comparison to WGBS ⁶⁴ but our study suggests that RRBS is complementary to WGBS and to obtain the most comprehensive genome wide estimation of DNA methylation the two methods should be combined.

Finally, we attempted to integrate the various epigenetic marks together with gene expression to show how the functional genome regulates expression levels. Correlations between the expression data and methylation, indicate that promoter methylation has a reverse relationship with gene expression, with, as expected, methylated sites inhibiting gene expression. Interestingly the methylation level within the gene-body in the pig cell line increases slightly with higher expression levels, whereas the methylation in the gene-body of the chicken cell

2 Molecular characterization of a chicken and pig cell line



line decreases slightly with increased expression levels. This is similar to what has been observed within the gene-body in earlier studies of the methylome in pig and avian species^{65,66}. This phenomenon has not been studied extensively, however Derks *et al*⁶⁶ suggested possible explanations such as methylation suppressing the transposable elements (TE) and preventing TE insertions which can be interruptive in the genome. A reason for this could be the higher number of TE in mammals compared to birds, which require higher methylation levels in gene bodies. As expected, the histone modification H3K27me3 which is associated with gene silencing shows a strong positive correlation to lowly expressed genes and a negative correlation to highly expressed genes. Histone modifications associated with promoter and enhancer regions all show a positive correlation to highly expressed genes and a negative correlation to lowly expressed genes. This confirms previous studies regarding these epigenetic marks regulating gene expression levels.

Conclusion

This paper is the first to describe the molecular characteristics (structure) of the pig IPECJ2 and chicken SL-29 cell lines. The genomic approaches provided an insight into the different levels of the epigenome influencing gene expression in these cell lines, as well as provided a description of the architecture of the epigenome. Chromosomal abnormalities, copy number variations and aneuploidy, typical for a cell line, were identified for several chromosomes for both cell lines. These cell lines are referred to as non-tumorigenic and non-transformed, however as these cells go through many passages aneuploidy events do occur. Future researchers should note the characteristics of these cell lines and proceed with caution for interpretation of results. Epigenetic marks such as histone modifications, chromatin accessibility and DNA methylation were integrated with expression data for both cell lines. This provided insight into the interactions between the epigenetic marks and gene expression. The characteristics as described in this paper for these cell lines will be similar for cells cultured using the same protocol and cells grown for the same number of passages. Deviations from these guidelines/methodologies are expected to result in different genomic and epigenomic characteristics. Understanding these cell lines and the (epi)genetic make-up thereof can provide a better understanding of the limitations of these cell lines as a model for *in vivo* research. We propose these cell lines as a reference for future functional and comparative studies in animals whereby knowledge of ploidy, expression profile, chromatin landscape and methylome provide the backbone for the comparison.



Limitations of the study

The ATAC-seq data was unavailable for the pig IPECJ-2 cell line. Further investigation using traditional methods for confirmation chromosomal abnormalities e.g. karyotypes staining would be beneficial.

Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2023.106252>.

Acknowledgements

The author J de Vos is funded by the GENE-SWitCH project, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under the grant agreement no 817998.

Authors' contributions

MG, RC, OM set up the experimental design for this manuscript. RC, BD, SK performed lab experiments. JdV performed analysis under supervision of OM, MG and MD contributed to analysis. MG, OM, JdV discussed and interpreted the results. JdV wrote the first version of the manuscript. All authors provided feedback and approved the analysis and manuscript for submission.

Declaration of interests

The authors declare that they have no competing interests.

STAR Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ole Madsen (ole.madsen@wur.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

2 Molecular characterization of a chicken and pig cell line



- All data types (WGS, RNA-seq, WGBS, RRBS, ChIP-seq – H3K4me1, H3K4me3, H3K27ac, H3K27me3, CTCF and ATAC-seq) from pig IPECJ2 and chicken SL-29 cell lines have been deposited at ENA / FAANG DCC and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Zenodo (DOI10.5281|zenodo.7274310) and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The intestinal epithelial pig IPECJ2 (ACC-701) cell line was obtained from the cell repository at DSMZ ⁶⁷, which is an intestinal columnar epithelial cell line derived from the mid-jejunum of a neonatal unsuckled piglet (piglets less than 12 hours old). These cells were originally isolated in 1989 by Helen Berschneider at the University of North Carolina ⁶⁸. For chicken the SL-29 cell line (ATCC CRL-1590), was derived from embryonic fibroblast cells obtained from the cell repository at ATCC ⁶⁹. Cells were cultured at 37°C and 5% CO₂ in Dulbecco's MEM with 5% FBS, Pen/Strep and Glutamax using a standard FAANG operating procedure. The media was refreshed twice a week and progressing to the next passage mostly 1/20 of the cells were transferred to a new flask. These cells were cultured for 4 passages in chicken SL-29 and 67 passages in pig, before harvesting.

METHOD DETAILS

Sequencing and assays

These cells were then used for whole genome sequencing (WGS), RNA sequencing, reduced representation bisulfite sequencing (RRBS), whole genome bisulfite sequencing (WGBS), ChIP-seq and ATAC-seq. DNA and RNA were isolated from the cell lines using the All Prep DNA/RNA Mini Kit (Qiagen) following manufacturer's instructions. WGS libraries of ~300-400 bp fragments were prepared using Illumina paired-end kits (Illumina, San Diego, CA) and 150 bp paired-end sequenced with Illumina HiSeqX. RNA-seq library preparation and sequencing was done as described in van der Hee et al (2020) using TruSeq RNA sample preparation kit (Illumina), incorporated within the Novogene manufacturer's protocol. Thereafter, samples were sequenced with Illumina Hi-Seq 4000 producing raw data with 150 bp paired-end reads. RRBS was done as described in Corbett et al (2020). In brief, DNA was

2 Molecular characterization of a chicken and pig cell line



fragmented using the *MSP*I restriction enzyme followed by a 20-250 bp fragment size selection and library preparation using the Ovation RRBS library (NuGEN). Samples were pooled and sequenced with the TruSeq SBS sequencing kit version 4 on the HiSeq 2500 (Illumina). A biological replicate was also sequenced following the same procedure for both pig IPECJ2 cell line. For WGBS genomic DNA was spiked with lambda DNA, fragmented by sonication to 200-400 bp with Covaris S220 (Covaris, Inc., Woburn, MA, USA), followed by end repair and A-ligation. Cytosine-methylated barcoded adapters were ligated to the sonicated DNA. The DNA bisulfite conversion was performed using the EZ DNA Methylation Gold Kit (Zymo Research, Irvine, USA). DNA fragments were size selected and amplified using the KAPA HiFi HotStart ReadyMix (2X) (Kapa Biosystems, Wilmington, USA). Library concentration was quantified using a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, USA) and qPCR (iCycler, BioRadLaboratories, Hercules, USA). Libraries were sequenced using the HiSeqX S4 flow cell with PE150 strategy.

ChIP-seq for both cell lines was performed for the insulator Anti-CTCF (polyclonal antibody lot # 2887267; Millipore) and histone marks H3K4me1 (polyclonal antibody ChIP grade ab8895; Abcam), H3K4me3 (polyclonal antibody lot # A1052D; Diagenode), H3K27ac (polyclonal antibody ChIP grade 4729; Abcam) and H3K27me3 (polyclonal antibody, lot # a1811-001P; Diagenode). These histone marks were chosen as they provide insight into transcriptional activation and the location of enhancers and promoters. ChIP-seq data sets were generated in two different laboratories (two replicates; experiment 1 and 2) for each cell-line. As input control in the first experiment, an IgG “mock” control was used, whereas in experiment two, an “input” DNA control was used for the ChIP-seq studies. In the second experiment only three histone marks (H3K4me1, H3K4me3 and H3K27ac) were assayed. The same ChIP-seq protocol was applied in both laboratories. Chromatin preparation was performed where cells (cultured in petri-dishes) were cross-linked with 1.1% formaldehyde for 10 min, stopped by adding 1/10 vol of 1.25 M Glycine for 2 min and washed with cold PBS. Cells were harvested by scraping, incubated with different buffers and finally resuspended in an incubation buffer with PIC with a final concentration of 15 million cells/ml. Shearing of the cells was performed in 300 μ l cell suspension /tube with 10 cycles 30 seconds on and 30 seconds off at 4°C using the Bioruptor Pico sonicator (Diagenode). Lastly the sonicated material was divided into aliquots and stored at -80 °C. The overnight immunoprecipitation step with the different antibodies at 4°C was performed on the chromatin using 4.5 million cells as input per antibody. Immunoprecipitated chromatin was incubated overnight with a 50:50 mix of PureProteome Protein A and G magnetic beads (Millipore). The beads were washed (6 washes with 4 wash buffers), rotated and de-cross-linked. The de-cross-linked DNA was finally purified (MinElute

2 Molecular characterization of a chicken and pig cell line



PCR Purification columns (Qiagen)), and DNA quantities were determined with Qubit fluorometric quantification (ThermoFisher Scientific). A qPCR analysis of ChIP DNA was performed with iQ SYBR Green Supermix (Bio-Rad) on a CFX96 Real-Time System C1000 Thermal Cycler (Bio-Rad). Library prep was performed using the Kapa Hyper Prep Kit for Illumina sequencing using the manufacturers protocol, with the following adjustments. DNA was used as an input together with NextFlex adapters (Bioo Scientific), followed by PCR amplification. Post-amplification cleanup was performed using QIAquick MinElute columns (Qiagen) and library size selection (300-bp fragments) was performed using the E-gel iBase (Invirtogen). Thereafter the quality and quantity of the library was examined using a High Sensitivity DNA Chip on a Bioanalyzer 2100 system (Agilent). Finally, the libraries were paired-end sequenced using Illumina high-throughput sequencing protocol on a HiSeq2000 (Illumina). For the second experiment sequencing was performed on a HiSeq4000 (Illumina).

Lastly ATAC-sequencing was completed following the Fast-ATAC-sequencing protocol as described in Corces *et al.*, 2016, with the following exceptions we used 25k cells as input and the standard Illumina Nextera primers for library amplification. These libraries were sequenced on a HiSeq4000 (Illumina), paired-end with 150 bp.

Data analysis

Pig (*Sus Scrofa* 11.1) and chicken (*Gallus GRCg6a*) reference genomes, together with ENSEMBL annotations (*Sus Scrofa* 11.1 - release 103 & *Gallus GRCg6a* - release 103) were utilized for all data analyses of our study. Default settings were used unless otherwise stated, and a brief overview of the data analyses is shown in figure S1. Genome indexes were built using the required reference genomes (and annotations where required) with the tools described below. Quality of all datasets was evaluated, and the statistics thereof is shown in tables S1 and S2.

Whole genome sequence analysis

Whole genome sequences were trimmed using Sickle v1.33⁷² in paired-end mode, where a sliding window approach was used for trimming adapters. Alignment of the trimmed reads together with removal of duplicates was completed using bwa mem (v0.7.15)⁷³ together with Sambalster (v0.1.26)⁷⁴. The aligned reads were further processed using samtools (v1.9)⁷⁵ to fill in mate coordinates, as well as add requirements from mate related flags. Mapping quality was evaluated using Qualimap (v2.2.1)⁷⁶ to ensure correct and accurate mapping.

Read-depth, genome structure, and possible large structural variants were evaluated using the tinycov package (v0.3.0)⁷⁷ and SNV calling was done using FreeBayes (v1.3.1)⁷⁸. Read

2 Molecular characterization of a chicken and pig cell line



support (ratio) was evaluated for heterozygous variants within the VCF file using a custom unix script (de Vos, 2021) and the results were plotted as histograms. The read-depth and SNV's were then plotted using Circos (v-0.69-9) ⁸⁰ for the visualization at a whole genome level, as well as for specific regions of interest.

Structural variant analysis of both cell lines was completed using Manta (v1.4.0) ⁸¹. For the pig cell line these variants were filtered in the following way: structural variants from healthy pig tissue samples with similar high read depth (in house samples: two muscle and one liver) were identified using Manta and overlapping structural variants between cell-lines and tissues were filtered out from the cell line structural variants. This strategy was used to exclude naturally occurring structural variants not unique to the cell line. For chicken Manta analysis this strategy was not possible due to lack of WGS chicken data with sufficient high coverage. Large SVs (deletions and duplications) were investigated using CNVnator (v0.3.3) ⁸², and results verified using a genome browser ⁸³. Variant Effect Predictor ⁸⁴ was used to determine the consequences of all copy number and structural variants on the genomes.

RNA-sequencing analysis

Stranded RNA-seq data sets for both cell lines (pig IPECJ2 and chicken SL-29) were trimmed for adapters, quality and minimum length using TrimGalore v0.6.4 a wrapper for Cutadapt v1.18 ⁸⁵ and the sequence data quality was evaluated using FastQC v0.11.9 ⁸⁶. The trimmed reads were used for alignment and gene quantification using RSEM ⁸⁷, with STAR v2.7.3a as aligner ⁸⁸. Further analyses were completed using custom shell scripts for basic statistics and average gene expression level per chromosome was calculated and plotted using a custom python script with the Seaborn package (de Vos, 2021). Various minimum transcript per kilobase million (TPM) thresholds were implemented for different analyses to reduce noise of uninformative genes that are very low expressed.

Additional raw RNA-sequencing data was downloaded from ENA from the PRJNA610529 project: two pig jejunum organoid samples grown for different time periods (3 weeks (SAMN14300031) and 12 weeks (SAMN14300021), a 5 week old pig jejunum tissue sample (SAMN14300018), cell lines IPECJ87, an IPECJ2 cell line grown for 87 passages (SAMN14300016), and IPECJ91, an IPECJ2 cell line grown for 91 passages (SAMN14299997). We trimmed, aligned, and completed gene quantification of this data following the same procedure as the above procedure used for the IPECJ2 cell line used in this study. These samples were used to compare the average gene expression level per chromosome in jejunum tissue, organoid and IPECJ2 cell line.

ChIP-sequencing analysis

2 Molecular characterization of a chicken and pig cell line



Raw reads were trimmed with TrimGalore v0.6.4 a wrapper for Cutadapt v 1.18⁹⁰ for adapter sequences, length, and quality. Reads for the different marks were aligned using bowtie2 (v2.3.2)⁹¹. Secondly, reads with red label (very low) on “per tile sequence quality” metric of FastQC were scanned with FilterbyTile (v.38.20) from BMAP package⁹². FilterbyTile increases the quality of Illumina reads, which are dependent on location in a flow cell. Moreover, the reads of the second experiment were truncated to match the read length of the first experiment (36 bp) allowing better comparison of the two. Samtools was used on the aligned reads for conversion of the alignment into BAM format, sorting, removing PCR duplicates, and keeping only paired-end reads, as well as uniquely aligned reads.

Peak calling for the respective marks, was completed using MACS2 (v2.7.1)^{93,94}, with the input (IgG and DNA respectively) used as the negative control. Visualization of the marks around the transcription start site (TSS) of expressed genes (TPM>1) was achieved using plot enrichment from deeptools (v3.1.3)⁹⁵. ChromHMM (v1.22)^{96,97} was used for the identification of different chromosome states based on interactions between marks, and the interaction around the TSS. Motif-based sequence analysis with MEME suite specifically MEME-ChIP was used, which is suitable for ChIP-seq data (v5.2.0)^{98,99}, for the CTCF mark to determine consensus motifs at the CTCF peak regions. A 500 bp region around the mid-position of CTCF called peaks is used for identification of motifs with MEME-ChIP (-norand, -ccut 0, -meme-nmotifs 30, -meme-minw 8 -meme-maxw 13) and MEME (-nmotifs 10 -minw 8 -maxw 12). Homer software (v4.1.0)¹⁰⁰ was used for gene, promoter, and transcription factor binding site (TFBS) discovery (-size 300 -len 8,10,13 -mset vertebrates), as well as annotating peak regions for the histone and CTCF marks. Regions showing gene silencing, promoters and enhancers were visualized using DROMPA¹⁰¹. Lastly, enhancers were identified as H3K27ac peaks which are not within 1000 bp of H3K4me3 peaks¹⁰². Read coverage and signal value of peaks, for respective histone marks H3K4me3, H3K4me1, H3K27ac of each experiment are compared using bedtools (v2.30.0)¹⁰³ and plot correlation.

ATAC-sequencing analysis

Trimming and alignment of the ATAC-seq reads were completed as described above for ChIP-seq reads. PCR duplicates were removed using picard (v2.23.9), and only unique, paired-end reads were kept for further analysis¹⁰⁴. Further filtering included removing the mitochondrial (MT) data, as a method of reducing bias in the results. Reads were shifted + 4 bp and - 5 bp for positive and negative strands respectively using an in-house unix script (de Vos, 2021) and this was done to account for the 9 bp duplication that occurs due to DNA repair of Tn5 transposase nick¹⁰⁵. This shift ensures accurate regions of the chromatin state for TF-footprint

2 Molecular characterization of a chicken and pig cell line



and motif related analysis. MACS2 (v2.7.1)^{106,107} was used for peak calling using default parameters. Peak annotation (homer v4.1.0) and motif analysis¹⁰⁸ for the peak regions were obtained, and identified motifs were scanned for known motifs such as TFBS and TATA-box using the Homer tool¹⁰⁹.

Methylation analysis

RRBS and WGBS raw reads were trimmed as described above for ChIP-seq reads, with an additional `--rrbs` parameter for RRBS data. Genome index was built using BSseeker2 v 2.1.8¹¹⁰, with bowtie2 as aligner, for RRBS and WGBS genome, with additional parameters for RRBS: `-r -l 10 -u 280`. Thereafter the reads were aligned using BSseeker2 (v 2.1.8), with additional parameters for RRBS: `--rrbs, -c MspI, -L 10 -U 280 -m 4` and for WGBS: `-l 0 -X 1000 -m 4`. BSseeker2 was used for the alignment as this tool is tailored for RRBS as it 'builds' a custom reference based on the restriction enzymes cutting sites. It is also more suitable to align gapped-reads than other tools commonly used for methylation analysis¹¹¹. We decided to keep methylation analysis standard across assays and thus implemented BSseeker2 for WGBS data as well. The aligned reads of the biological replicates of the pig cell line were merged for further analysis, after a Pearson correlation¹¹² showed a high correlation of 0.96 between the two RRBS technical replicate samples (figure S2). CGmaptools (v0.1.2)¹¹³ was implemented for the methylation calling from the aligned reads. Further statistical analysis of the methylation data was completed using CGmaptools (v0.1.2) and MethylKit (v1.16.1)¹¹⁴. Correlations and clustering between the biological replicates were analysed using MethylKit.

Finally, the number of sites identified by RRBS data and not by WGBS data was investigated using the following approach. Firstly, the methylation calls were filtered for only CG sites and for a coverage of more than 10. Thereafter bedtools (v2.30.0)¹¹⁵ was used to identify the sites unique to RRBS data, and for merging these regions. Functional importance of these regions was investigated by overlaying regions with enhancer and promoter regions detected from the ChIP-seq analysis, visual examination of the sites using Jbrowse¹¹⁶, and motif discovery using homer.

Integrative analysis

An integrative approach was used to investigate the relationship between WGS, expression data (RNA-seq), methylation status and ChIP-seq marks. Output files from the homer annotate called peaks for the respective marks, gene expression values from RSEM, and methylation calls from CGmaptools were used. For this investigation, correlations, scatterplots, and boxplots were created from these files using an in-house python script (de Vos, 2021).



QUANTIFICATION AND STATISTICAL ANALYSIS

Pearson correlations between the read coverage and signal value of peaks for histone marks H3K4me1, H3K4me3 and H3K27ac of each experiment was done using bedtools (v2.30.0)¹⁰³, statistical function (scipy.stats) and seaborn package for visualisation in python. A pearson correlation was performed between two technical replicates of the pig IPECJ2 cell line, we used the tool MethylKit (v1.16.1)¹¹⁴ for this quantification. A high correlation of 0.96 (figure S2) confirmed that the two samples could be merged for further downstream analysis. Integrative analysis used heatmaps, Pearson correlations, scatterplots and boxplots to investigate the relationships between gene expression, various histone modifications indicating promoters, enhancers and gene silencing (H3K4me1, H3K4me3, H3K27ac, H3K27me3), TF (CTCF) and open chromatin (ATAC-seq). This was done using an in-house python script which is available (de Vos, 2021).

References

1. Bell, O., Tiwari, V.K., Thomä, N.H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics* 2011 12:8 12, 554–564. 10.1038/nrg3017.
2. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129, 823–837. 10.1016/J.CELL.2007.05.009.
3. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics* 2008 40:7 40, 897–903. 10.1038/ng.154.
4. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009 459:7243 459, 108–112. 10.1038/nature07829.
5. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* 107, 21931–21936. 10.1073/PNAS.1016071107.
6. Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* 16, 1–6. 10.1186/S13059-015-0622-4/METRICS.

2 Molecular characterization of a chicken and pig cell line



7. Feingold, E.A., Good, P.J., Guyer, M.S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F.S., Gingeras, T.R., Kampa, D., Sekinger, E.A., et al. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* *306*, 636–640. 10.1126/SCIENCE.1105136.
8. Bols, N.C., Kawano, A., and Lee, L.E.J. (2011). CELLULAR, MOLECULAR, GENOMICS, AND BIOMEDICAL APPROACHES | Culture of Fish Cell Lines. *Encyclopedia of Fish Physiology* *3*, 1965–1970. 10.1016/B978-0-12-374553-8.00253-7.
9. Verma, A. (2014). Animal Tissue Culture: Principles and Applications. *Animal Biotechnology: Models in Discovery and Translation*, 211–231. 10.1016/B978-0-12-416002-6.00012-2.
10. Verma, A., Verma, M., and Singh, A. (2020). Animal tissue culture principles and applications. *Anim Biotechnol*, 269. 10.1016/B978-0-12-811710-1.00012-4.
11. Molina, O., Abad, M.A., Solé, F., and Menéndez, P. (2021). Aneuploidy in Cancer: Lessons from Acute Lymphoblastic Leukemia. *Trends Cancer* *7*, 37–47. 10.1016/J.TRECAN.2020.08.008.
12. Verma, A., Verma, M., and Singh, A. (2020). Animal tissue culture principles and applications. *Anim Biotechnol*, 269. 10.1016/B978-0-12-811710-1.00012-4.
13. RAMIREZ, A.D., ROCHA, E.M.M., and KRETTLI, A.U. (1995). Antisporozoite Antibodies with Protective and Nonprotective Activities: In Vitro and In Vivo Correlations Using *Plasmodium gallinaceum*, an Avian Model. *Journal of Eukaryotic Microbiology* *42*, 705–708. 10.1111/J.1550-7408.1995.TB01620.X.
14. Närvä, E., Autio, R., Rahkonen, N., Kong, L., Harrison, N., Kitsberg, D., Borghese, L., Itskovitz-Eldor, J., Rasool, O., Dvorak, P., et al. (2010). High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat Biotechnol* *28*, 371–377. 10.1038/NBT.1615.
15. Vcelar, S., Jadhav, V., Melcher, M., Auer, N., Hrdina, A., Sagmeister, R., Heffner, K., Puklowski, A., Betenbaugh, M., Wenger, T., et al. (2018). Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting. *Biotechnol Bioeng* *115*, 165–173. 10.1002/BIT.26453.
16. Hynds, R.E., Vladimirov, E., and Janes, S.M. (2018). The secret lives of cancer cell lines. *Dis Model Mech* *11*. 10.1242/DMM.037366.
17. Rebuzzini, P., Zuccotti, M., Redi, C.A., and Garagna, S. (2015). Chromosomal Abnormalities in Embryonic and Somatic Stem Cells. *Cytogenet Genome Res* *147*, 1–9. 10.1159/000441645.
18. Clevers, H. (2016). Modeling Development and Disease with Organoids. *Cell* *165*, 1586–1597. 10.1016/J.CELL.2016.05.082.

2 Molecular characterization of a chicken and pig cell line



19. van der Hee, B., Madsen, O., Vervoort, J., Smidt, H., and Wells, J.M. (2020). Congruence of Transcription Programs in Adult Stem Cell-Derived Jejunum Organoids and Original Tissue During Long-Term Culture. *Front Cell Dev Biol* 8, 375. 10.3389/FCELL.2020.00375/BIBTEX.
20. Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R.F., Barany, F., et al. (2006). Relationship of Gene Expression and Chromosomal Abnormalities in Colorectal Cancer. *Cancer Res* 66, 2129–2166. 10.1158/0008-5472.CAN-05-2569.
21. Harewood, L., and Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Hum Mol Genet* 23. 10.1093/HMG/DDU278.
22. Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., and Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Med Genet* 20, 1–14. 10.1186/S12881-019-0909-5/FIGURES/6.
23. Lappalainen, T., Scott, A.J., Brandt, M., and Hall, I.M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. *Cell* 177, 70–84. 10.1016/J.CELL.2019.02.032.
24. Bařinka, J., Hu, Z., Wang, L., Wheeler, D.A., Rahbarinia, D., McLeod, C., Gu, Z., and Mullighan, C.G. (2022). RNAseqCNV: analysis of large-scale copy number variations from RNA-seq data. *Leukemia* 2022 36:6 36, 1492–1498. 10.1038/s41375-022-01547-8.
25. Hochstenbach, R., van Binsbergen, E., Schuring-Blom, H., Buijs, A., and Ploos van Amstel, H.K. (2019). A survey of undetected, clinically relevant chromosome abnormalities when replacing postnatal karyotyping by Whole Genome Sequencing. *Eur J Med Genet* 62, 103543. 10.1016/J.EJMG.2018.09.010.
26. Zhou, J., Yang, Z., Sun, J., Liu, L., Zhou, X., Liu, F., Xing, Y., Cui, S., Xiong, S., Liu, X., et al. (2021). Whole Genome Sequencing in the Evaluation of Fetal Structural Anomalies: A Parallel Test with Chromosomal Microarray Plus Whole Exome Sequencing. *Genes* 2021, Vol. 12, Page 376 12, 376. 10.3390/GENES12030376.
27. Nurchis, M.C., Riccardi, M.T., Radio, F.C., Chillemi, G., Bertini, E.S., Tartaglia, M., Cicchetti, A., Dallapiccola, B., and Damiani, G. (2022). Incremental net benefit of whole genome sequencing for newborns and children with suspected genetic disorders: Systematic review and meta-analysis of cost-effectiveness evidence. *Health Policy (New York)* 126, 337–345. 10.1016/J.HEALTHPOL.2022.03.001.
28. Guseh, S.H. (2020). Noninvasive prenatal testing: from aneuploidy to single genes. *Hum Genet* 139, 1141–1148. 10.1007/s00439-019-02061-1.
29. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 22, 1813–1831. 10.1101/GR.136184.111.

2 Molecular characterization of a chicken and pig cell line



30. Prowse-Wilkins, C.P., Wang, J., Xiang, R., Garner, J.B., Goddard, M.E., and Chamberlain, A.J. (2021). Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. *Front Genet* 12, 1027. 10.3389/FGENE.2021.664379/BIBTEX.
31. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 22, 1813–1831. 10.1101/GR.136184.111.
32. Gao, Y., Hyttel, P., and Hall, V.J. (2010). Regulation of H3K27me3 and H3K4me3 during early porcine embryonic development. *Mol Reprod Dev* 77, 540–549. 10.1002/MRD.21180.
33. Han, K., Ren, R., Cao, J., Zhao, S., and Yu, M. (2019). Genome-wide identification of histone modifications involved in placental development in pigs. *Front Genet* 10, 277. 10.3389/FGENE.2019.00277/BIBTEX.
34. Kimura, H. (2013). Histone modifications for human epigenome analysis. *Journal of Human Genetics* 2013 58:7 58, 439–445. 10.1038/jhg.2013.66.
35. Prakash, K., and Fournier, D. (2018). Evidence for the implication of the histone code in building the genome structure. *Biosystems* 164, 49–59. 10.1016/J.BIOSYSTEMS.2017.11.005.
36. Jiang, S., and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. *Brief Funct Genomics* 17, 104–115. 10.1093/BFGP/ELY002.
37. Kimura, H. (2013). Histone modifications for human epigenome analysis. *Journal of Human Genetics* 2013 58:7 58, 439–445. 10.1038/jhg.2013.66.
38. Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G., Saelao, P., Waters, S., Xiang, R., Chamberlain, A., et al. (2021). Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nature Communications* 2021 12:1 12, 1–11. 10.1038/s41467-021-22100-8.
39. Prakash, K., and Fournier, D. (2018). Evidence for the implication of the histone code in building the genome structure. *Biosystems* 164, 49–59. 10.1016/J.BIOSYSTEMS.2017.11.005.
40. Jiang, S., and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. *Brief Funct Genomics* 17, 104–115. 10.1093/BFGP/ELY002.
41. Martin, D., Pantoja, C., Fernández-Miñán, A., Valdes-Quezada, C., Moltó, E., Matesanz, F., Bogdanović, O., De La Calle-Mustienes, E., Domínguez, O., Taher, L., et al. (2011). Genome-wide CTCF distribution in vertebrates defines equivalent sites that can aid in the identification of disease-associated genes. *Nat Struct Mol Biol* 18, 708. 10.1038/NSMB.2059.
42. Moore, J.M., Rabaia, N.A., Smith, L.E., Fagerlie, S., Gurley, K., Loukinov, D., Disteche, C.M., Collins, S.J., Kemp, C.J., Lobanenko, V. V., et al. (2012). Loss of Maternal CTCF Is Associated with

2 Molecular characterization of a chicken and pig cell line



Peri-Implantation Lethality of Cctf Null Embryos. *PLoS One* 7, e34915. 10.1371/JOURNAL.PONE.0034915.

43. Li, F., Wang, D., Song, R., Cao, C., Zhang, Z., Wang, Y., Li, X., Huang, J., Liu, Q., Hou, N., et al. (2020). The asynchronous establishment of chromatin 3D architecture between in vitro fertilized and uniparental preimplantation pig embryos. *Genome Biol* 21, 1–21. 10.1186/S13059-020-02095-Z/FIGURES/5.

44. Luan, Y., Zhang, L., Hu, M., Xu, Y., Hou, Y., Li, X., Zhao, S., Zhao, Y., and Li, C. (2019). Identification and Conservation Analysis of Cis-Regulatory Elements in Pig Liver. *Genes* 2019, Vol. 10, Page 348 10, 348. 10.3390/GENES10050348.

45. Liu, Y., Xu, J., Fu, W., Wang, W., Niu, X., Liu, J., Ding, X., and Zhang, Q. (2012). Molecular cloning, expression, mapping of interferon regulatory factor 2 (IRF2) gene and its association with immune traits in pigs. *Livest Sci* 148, 201–207. 10.1016/J.LIVSCI.2012.06.001.

46. Du, K., Zhong, Z., Fang, C., Dai, W., Shen, Y., Gan, X., and He, S. (2018). Ancient duplications and functional divergence in the interferon regulatory factors of vertebrates provide insights into the evolution of vertebrate immune systems. *Dev Comp Immunol* 81, 324–333. 10.1016/J.DCI.2017.12.016.

47. Liu, B. qin, Jin, J., and Li, Y. yuan (2020). Ubiquitination modification: critical regulation of IRF family stability and activity. *Science China Life Sciences* 2020 64:6 64, 957–965. 10.1007/S11427-020-1796-0.

48. Wu, X.N., Shi, T.T., He, Y.H., Wang, F.F., Sang, R., Ding, J.C., Zhang, W.J., Shu, X.Y., Shen, H.F., Yi, J., et al. (2017). Methylation of transcription factor YY2 regulates its transcriptional activity and cell proliferation. *Cell Discovery* 2017 3:1 3, 1–22. 10.1038/celldisc.2017.35.

49. To, B., and Andrechek, E.R. (2018). Transcription factor compensation during mammary gland development in E2F knockout mice. *PLoS One* 13. 10.1371/JOURNAL.PONE.0194937.

50. Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., EsquerréEsquerr, D., Zytnicki, M., Derrien, T., Bardou, P., et al. (2019). Transcriptome and chromatin structure annotation of liver, CD4+ and CD8+ T cells from four livestock species. *bioRxiv*, 316091. 10.1101/316091.

51. Halstead, M.M., Kern, C., Saelao, P., Chanthavixay, G., Wang, Y., Delany, M.E., Zhou, H., and Ross, P.J. (2020). Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Scientific Reports* 2020 10:1 10, 1–12. 10.1038/s41598-020-61678-9.

52. Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., EsquerréEsquerr, D., Zytnicki, M., Derrien, T., Bardou, P., et al. (2019). Transcriptome and chromatin structure annotation of liver, CD4+ and CD8+ T cells from four livestock species. *bioRxiv*, 316091. 10.1101/316091.

2 Molecular characterization of a chicken and pig cell line



53. Halstead, M.M., Kern, C., Saelao, P., Chanthavixay, G., Wang, Y., Delany, M.E., Zhou, H., and Ross, P.J. (2020). Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Scientific Reports* 2020 10:1 10, 1–12. 10.1038/s41598-020-61678-9.
54. Yong, W.S., Hsu, F.M., and Chen, P.Y. (2016). Profiling genome-wide DNA methylation. *Epigenetics & Chromatin* 2016 9:1 9, 1–16. 10.1186/S13072-016-0075-3.
55. Lacial, I., and Ventura, R. (2018). Epigenetic Inheritance: Concepts, Mechanisms and Perspectives. *Front Mol Neurosci* 11. 10.3389/FNMOL.2018.00292.
56. Patil, V., Ward, R.L., and Hesson, L.B. (2014). The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics* 9, 823. 10.4161/EPI.28741.
57. Derks, M.F.L., Schachtschneider, K.M., Madsen, O., Schijlen, E., Verhoeven, K.J.F., and van Oers, K. (2016). Gene and transposable element methylation in great tit (*Parus major*) brain and blood. *BMC Genomics* 17. 10.1186/S12864-016-2653-Y.
58. Zhang, M., Yan, F. Bin, Li, F., Jiang, K.R., Li, D.H., Han, R.L., Li, Z.J., Jiang, R.R., Liu, X.J., Kang, X.T., et al. (2017). Genome-wide DNA methylation profiles reveal novel candidate genes associated with meat quality at different age stages in hens. *Sci Rep* 7, 45564–45564. 10.1038/SREP45564.
59. Zhang, H., Ali, A., Hou, F., Wu, T., Guo, D., Zeng, X., Wang, F., Zhao, H., Chen, X., Xu, P., et al. (2018). Effects of ploidy variation on promoter DNA methylation and gene expression in rice (*Oryza sativa* L.). *BMC Plant Biol* 18, 1–12. 10.1186/S12870-018-1553-5/TABLES/4.
60. Liu, J., Zhou, F., Cui, S., Yang, Y., Sun, Q., Guan, Q., Li, D.L., Zhang, S., and Wang, R. (2022). Effects of ploidy variation on DNA methylation and gene expression in Pear (*Pyrus communis* L.). *Sci Hortic* 293, 110713. 10.1016/J.SCIENTA.2021.110713.
61. Martin-Trujillo, A., Vidal, E., Monteagudo-Sánchez, A., Sanchez-Delgado, M., Moran, S., Hernandez Mora, J.R., Heyn, H., Guitart, M., Esteller, M., and Monk, D. (2017). Copy number rather than epigenetic alterations are the major dictator of imprinted methylation in tumors. *Nat Commun* 8. 10.1038/S41467-017-00639-9.
62. Cadieux, E.L., Mensah, N.E., Castignani, C., Tanić, M., Wilson, G.A., Dietzen, M., Dhami, P., Vaikkinen, H., Verfaillie, A., Martin, C.C., et al. Copy number-aware deconvolution of tumor-normal DNA methylation profiles. 10.1101/2020.11.03.366252.
63. Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols* 2011 6:4 6, 468–481. 10.1038/nprot.2010.190.

2 Molecular characterization of a chicken and pig cell line



64. Yong, W.S., Hsu, F.M., and Chen, P.Y. (2016). Profiling genome-wide DNA methylation. *Epigenetics & Chromatin* 2016 9:1 9, 1–16. 10.1186/S13072-016-0075-3.
65. Schachtschneider, K.M., Madsen, O., Park, C., Rund, L.A., Groenen, M.A.M., and Schook, L.B. (2015). Adult porcine genome-wide DNA methylation patterns support pigs as a biomedical model. *BMC Genomics* 16. 10.1186/S12864-015-1938-X.
66. Derks, M.F.L., Schachtschneider, K.M., Madsen, O., Schijlen, E., Verhoeven, K.J.F., and van Oers, K. (2016). Gene and transposable element methylation in great tit (*Parus major*) brain and blood. *BMC Genomics* 17. 10.1186/S12864-016-2653-Y.
67. German Collection of Microorganisms and Cell Cultures GmbH: Details.
68. Berschneider, H.M. (1989). Development of normal cultured small small intestinal epithelial cell lines which transport Na and Cl. *Gastroenterology* 96, A41.
69. SL-29 | ATCC.
70. van der Hee, B., Madsen, O., Vervoort, J., Smidt, H., and Wells, J.M. (2020). Congruence of Transcription Programs in Adult Stem Cell-Derived Jejunum Organoids and Original Tissue During Long-Term Culture. *Front Cell Dev Biol* 8, 375. 10.3389/FCELL.2020.00375/BIBTEX.
71. Corbett, R.J., te Pas, M.F.W., van den Brand, H., Groenen, M.A.M., Crooijmans, R.P.M.A., Ernst, C.W., and Madsen, O. (2020). Genome-Wide Assessment of DNA Methylation in Chicken Cardiac Tissue Exposed to Different Incubation Temperatures and CO₂ Levels. *Front Genet* 11, 1310. 10.3389/FGENE.2020.558189/BIBTEX.
72. Joshi NA, F.J. (2011). GitHub - najoshi/sickle: Windowed Adaptive Trimming for fastq files using quality.
73. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
74. Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 30, 2503–2505. 10.1093/BIOINFORMATICS/BTU314.
75. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. 10.1093/BIOINFORMATICS/BTP352.
76. García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., and Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28, 2678–2679. 10.1093/BIOINFORMATICS/BTS503.
77. tinycov · PyPI.
78. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.

2 Molecular characterization of a chicken and pig cell line



79. Vos, J. de GitHub - Jani-94/scripts.
80. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* 19, 1639–1645. 10.1101/GR.092759.109.
81. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222. 10.1093/BIOINFORMATICS/BTV710.
82. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21, 974–984. 10.1101/GR.114876.110.
83. Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L., et al. (2016). JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol* 17, 1–12. 10.1186/S13059-016-0924-1/TABLES/3.
84. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol* 17, 1–14. 10.1186/S13059-016-0974-4/TABLES/8.
85. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17, 10–12. 10.14806/EJ.17.1.200.
86. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data – ScienceOpen.
87. Li, B., and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 1–16. 10.1186/1471-2105-12-323/TABLES/6.
88. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. 10.1093/BIOINFORMATICS/BTS635.
89. Vos, J. de GitHub - Jani-94/scripts.
90. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17, 10–12. 10.14806/EJ.17.1.200.
91. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012 9:4 9, 357–359. 10.1038/nmeth.1923.
92. Bushnell, B. BMAP download | SourceForge.net. 2014.

2 Molecular characterization of a chicken and pig cell line



93. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, 1–9. 10.1186/GB-2008-9-9-R137/FIGURES/3.
94. Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols* 2012 7:9 7, 1728–1740. 10.1038/nprot.2012.101.
95. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44, W160–W165. 10.1093/NAR/GKW257.
96. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 2012 9:3 9, 215–216. 10.1038/nmeth.1906.
97. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 12, 2478–2492. 10.1038/NPROT.2017.124.
98. Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697. 10.1093/BIOINFORMATICS/BTR189.
99. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. *Nucleic Acids Res* 43, W39–W49. 10.1093/NAR/GKV416.
100. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38, 576–589. 10.1016/J.MOLCEL.2010.05.004.
101. Nakato, R., and Shirahige, K. (2018). Statistical Analysis and Quality Assessment of ChIP-seq Data with DROMPA. *Methods Mol Biol* 1672, 631–643. 10.1007/978-1-4939-7306-4_41.
102. Klein, J.C., Keith, A., Agarwal, V., Durham, T., and Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. *Genome Biol* 19, 1–13. 10.1186/S13059-018-1473-6/FIGURES/5.
103. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. 10.1093/BIOINFORMATICS/BTQ033.
104. Picard Tools - By Broad Institute.
105. Yan, F., Powell, D.R., Curtis, D.J., and Wong, N.C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* 2020 21:1 21, 1–16. 10.1186/S13059-020-1929-3.
106. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, 1–9. 10.1186/GB-2008-9-9-R137/FIGURES/3.

2 Molecular characterization of a chicken and pig cell line



107. Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols* 2012 7:9 7, 1728–1740. 10.1038/nprot.2012.101.
108. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. *Nucleic Acids Res* 43, W39–W49. 10.1093/NAR/GKV416.
109. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38, 576–589. 10.1016/J.MOLCEL.2010.05.004.
110. Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.Y., and Pellegrini, M. (2013). BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* 14, 1–8. 10.1186/1471-2164-14-774/TABLES/4.
111. Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.Y., and Pellegrini, M. (2013). BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* 14, 1–8. 10.1186/1471-2164-14-774/TABLES/4.
112. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 13, 1–9. 10.1186/GB-2012-13-10-R87/FIGURES/6.
113. Guo, W., Zhu, P., Pellegrini, M., Zhang, M.Q., Wang, X., and Ni, Z. (2018). CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* 34, 381–387. 10.1093/BIOINFORMATICS/BTX595.
114. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 13, 1–9. 10.1186/GB-2012-13-10-R87/FIGURES/6.
115. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. 10.1093/BIOINFORMATICS/BTQ033.
116. Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L., et al. (2016). JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol* 17, 1–12. 10.1186/S13059-016-0924-1/TABLES/3.

Chapter 3

GSM-pipeline: GENE-SWitCH pipeline for comprehensive bisulfite sequencing analysis

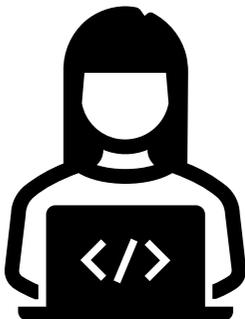
Jani de Vos^{1*}, Martijn F.L. Derks¹, Cyril Kurylo², Martien A.M. Groenen¹, Ole Madsen¹

¹ Animal Breeding and Genetics Group, Wageningen University & Research, Wageningen, The Netherlands

² GenPhySE, Université de Toulouse, INRAE,INPT, ENVT, Castanet Tolosan, France

Available on ResearchSquare

<https://doi.org/10.21203/rs.3.rs-2984574/v1>



'Success is the sum of small efforts, repeated day in and day out'

Robert Collier

Abstract

Background

DNA methylation is a process by which methyl groups are added to the DNA molecule. Methylation of promoters influence gene expression, as the addition of a methyl group can prevent the binding of certain transcription factors and repress the transcription of the associated gene. Bioinformatics pipelines are a series of computational steps or workflows designed to analyse biological data such as sequencing data. Typically several steps, such as data pre-processing, quality control, alignment, and downstream analysis are involved. These pipelines can be tailored to a specific research question and can be customized with various software and tools. GENE-SWitCH is an EU H2020 project with the aim of providing the functional genome annotation in pig and chicken, with many different assays specifically DNA methylation.

Results

We present a pipeline for processing bisulfite sequencing data, which uses nf-core methylseq as a foundation. This extension pipeline includes useful steps such as analysing a bam file, methylation calling, methylation visualisation on a genome-wide level, and methylation statistics.

Conclusion

Our pipeline is useful for the analysis of bisulfite sequencing data, ensuring reproducible results, and stable tool versions. It is easy to use and implement for any given data set and we executed it for analysis of > 80 RRBS and WGBS data sets.

Keywords

DNA methylation, alignment, pipeline, methylation calling

Background

DNA methylation is an epigenetic mechanism whereby a methyl group is added to the 5th position of cytosine. DNA methylation changes the activity of a DNA segment and is generally associated with repression of gene expression when located in a regulatory sequence [1, 2]. This epigenetic modification has become more widely studied and there are different methods to assess the methylation status of DNA positions. One method is bisulfite sequencing, whereby the unmethylated cytosines are converted to a uracil and methylated C's stay unchanged after bisulfite treatment, followed by sequencing. This gives a resolution at the

3 GSM-pipeline

single nucleotide level over the entire genome, WGBS (whole genome bisulfite sequencing) or on a reduced level, by which approximately 2% of the genome is sequenced (Reduced representation bisulfite sequencing – RRBS) [3]. RRBS provides insight into a subset of the genome especially areas of the genome with a high CpG content (promotor and other regulatory regions). Costs for RRBS sequencing are lower in comparison with WGBS, which is a major advantage of using this method over full genome bisulfite sequencing [4].

Given the strong association with regulation of gene expression, DNA methylation has become a research topic of interest and many datasets are produced [1]. Hence, one of the important aspects of this research is the reproducibility of results. Workflow management systems provide an important resource that contributes to reproducibility of results. Tools such as Nextflow [5] and Snakemake [6] contribute to data and code management [7] together with container platforms such as Singularity and Docker [8, 9]. For Nextflow, nf-core is a scientific community driven effort to create bioinformatics pipelines for genomic data using Docker for containerization of the tools [10].

A range of tools are currently available for the analysis of bisulfite sequencing data, including quality control, methylation calling, and downstream comparative analyses. More specifically, steps for analysing bisulfite sequence data commonly include quality control, trimming, alignment/mapping, deduplication, methylation calling, and visualization. Further downstream analyses are specific to the research questions and can include differential methylation analyses, and investigations into the associations with expression data. Tools commonly used for alignment of methylation data include Bismark, BSseeker2, and bwa-meth, which implement different alignment strategies [11, 12].

The GENE-SWitCH project is an EU H2020 project, with the aim of obtaining the functional genome of two monogastric species (pig and chicken) during three developmental stages in seven different tissues representative of germ layers. Different assays are used to investigate the functional genome within the project (including WGBS and RRBS datasets) [13]. Determination of the changes in gene expression, epigenetic states and activity of functional regulatory elements, provide insight into development and the underlying mechanisms regulating the biological process of development. The development and identification of suitable pipelines for the analysis of the various data types produced within this research is a strategy implemented within the project to ensure reproducible results.

In the current study we aimed to compile the various steps for methylation data processing into an easy to use bioinformatics pipeline to ensure reproducibility of results. Here we present a GENE-SWitCH DNA methylation analysis pipeline (GSM), specifically for the analysis of

bisulfite sequencing data. This pipeline was initially developed using the nf-core methylseq pipeline as a foundation including additional features for visualization and integration of gene expression data.

Implementation

The GSM-pipeline is written using nextflow as the workflow manager, and a docker container for all required tools. The pipeline includes different tools to analyse bisulfite sequencing data from raw reads (fastq) to methylation calling, and provides detailed general methylation statistics. The tools included in this pipeline have been extensively validated individually and we have combined the best practice tools into this pipeline. **Fig 1** shows the flow of the pipeline together with the various tools incorporated to achieve basic analysis of bisulfite sequencing data.

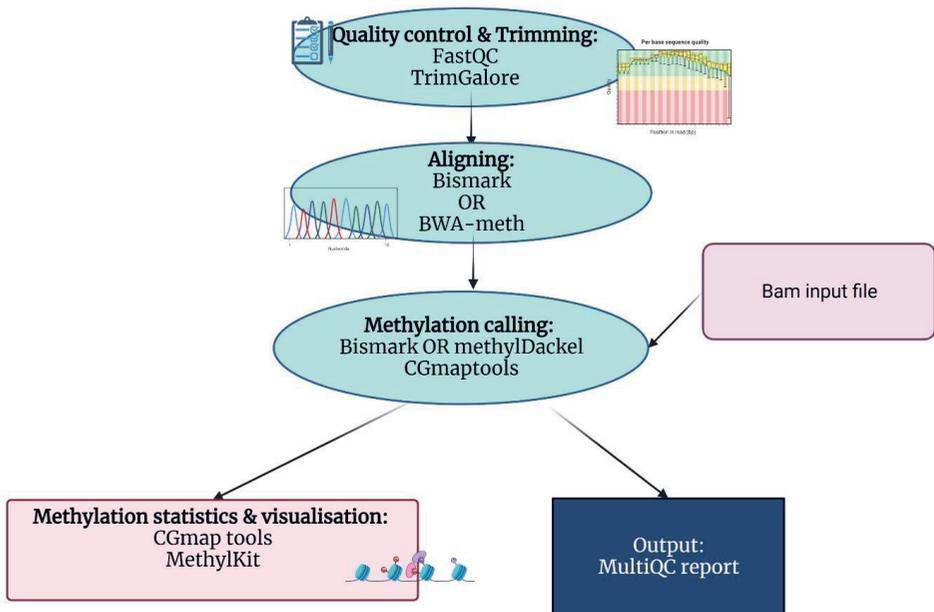


Figure 1 Overview of the workflow of the GSM-pipeline.

Quality control & Mapping

The pipeline starts with processing raw reads and performing quality control using FastQC [14], and adapter trimming with TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Users of the pipeline have the flexibility to decide between two different aligners, either Bismark [15] which relies on

3 GSM-pipeline

Bowtie2 or HISAT2 as well as Samtools for alignment, or BWA-meth [16]. Deduplication for the removal of PCR duplicates is dependent on the aligner used. Bismark includes a deduplication process and Picard MarkDuplicates [17] is executed when using BWA-meth as aligner. In addition, alignment files (BAM files) can be used as input for downstream analysis. Lastly the alignment statistics are determined with Qualimap [18] and all results from QC, trimming and alignment are compiled into a single report file using MultiQC [19].

Methylation calling & statistics

Methylation calling is performed using Bismark to generate a comprehensive methylation report, which is necessary for downstream analysis and MethylDackel [20] which is only used in conjunction with the BWA-meth aligner. In addition to the above processes provided by the original nf-core pipeline, our pipeline includes CGmaptools and MethylKit. CGmaptools [21] is used as an additional tool for methylation calling, providing CGmap and ATCGmap output files which include comprehensive information (e.g. non-cytosine context) regarding the methylome provided for downstream analysis. CGmaptools provides insight into genome-wide and chromosomal methylation levels, and coverage of the methylome can be investigated in detail. Additional descriptive statistics on DNA methylation profiles, such as frequency of methylation per CpG and mean CpG coverage are visualised using the MethylKit package [22].

Results

Datasets for development and testing

Results shown in this section are from a WGBS pig IPECJ2 cell line aiming to demonstrate the usability of the pipeline. This dataset contained 1.13 billion raw reads, with 939 million aligned paired end reads and an alignment rate of 85%.

Performance

Performance reports are provided from the pipeline which displays CPU usage, job run time, and memory usage for each individual process in the pipeline. We observed a total of 437.7 CPU hours for completion of the pipeline. Alignment is the single process which can be done using multi-threading (we used 8 CPUs) and uses the most memory (>30G) and takes most runtime. Trimming and methylation calling processes use 4 CPU's (~4) and indexing as well as Qualimap needs about 20-25G of memory for a single WGBS dataset. This is as expected when evaluating the performance of these tools and processes separately. Run time, CPU and memory use vary depending on the type of data (WGBS vs. RRBS).

Methylation and quality statistics

Results from CGmapprools provide information on read coverage per chromosome (**Fig 2b**), methylation statistics (**Fig 2c**), and methylation levels per chromosome (**Fig 2a**).

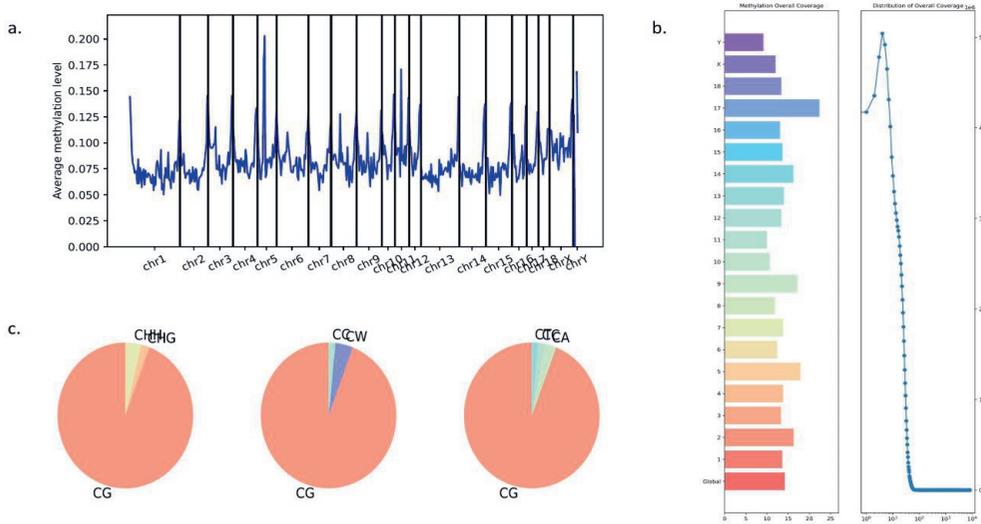


Figure 2 Results for the WGBS from pig IPECJ2 cell line showing a.) average CG methylation level per chromosome b.) Methylation effective coverage (MEC) plot, providing the global read coverage as well as coverage per chromosome (left panel). The (MEC) is calculated as the average read coverage only for cytosines. The right panel shows the distribution of the overall coverage calculated for each chromosome, which is calculated as the average read coverage on all nucleotides on both strands. c.) Pie charts describing the percentage of CpG, CHH and CHG sites across the genome.

The methylation contexts are different for plants and animals. For plants, the contexts for DNA methylations are known as CG, CHG and CHH, where H= {A, C, T}. For animals, the situation is different. Guo *et al.*, (2014) [23] showed that it is unnecessary to separate CHG methylations and CHH methylations in human. In 2016 Guo [24], designed the MiDD method and de novo predicted that the main separated contexts for non-CG (CH) methylation should be CW {W = A, T} and CC, and that mCW is cell-type specific and conserved between human and mice.

The average CG methylation levels for the WGBS from the pig IPECJ2 cell line are between 0.5-0.7 with peaks on chromosomes 5, 8 and 12 (**Fig 2a**). Most methylation is CpG methylation and non-CpG methylation is relatively rare (**Fig 2c**). Lastly the average read coverage for this sample is ~14.6x and most chromosomes are well covered (coinciding with the average), except for chromosome 17 which has a very high read coverage (25).

3 GSM-pipeline

Histogram of the frequency of the methylation levels provides valuable insight into the quality of the data, and a bimodal distribution is expected. Differences between RRBS and WGBS data types as well as between species can be observed in research. This metric is useful in determining the significant difference in methylation % parameter which is implemented in downstream differential analysis.

Conclusion

We present a comprehensive pipeline for the analysis of bisulfite sequencing data. We benchmarked the pipeline using a WGBS dataset of pig IPECJ2 cell line. In addition, the pipeline is extensively tested on a variety of datasets from the GENE-SWitCH consortium (7 different tissues in 2 vertebrate species, chicken and pig), about 80 RRBS and WGBS. The processed data from the pipeline can easily be imported into tools for visualization of methylation levels across genomic features. Additionally, this pipeline is currently in DSL1 format, however a newer version of DSL (DSL2) was released. There will be updates to the code in future to accommodate with the continuous release of various tools and programming languages. We believe this pipeline will be of great value for scientists working with bisulfite sequencing data as it is easy to use and implement for any given dataset. This pipeline contributes to attaining reproducible results for bisulfite sequencing analysis to investigate methylation.

Availability and requirements

Project name: GSM pipeline (GENE-SWitCH Methylation pipeline)

Project home page: <https://github.com/FAANG/GSM-pipeline>

Operating system(s): Platform independent

Programming language: Nextflow

Other requirements: Nextflow, Singularity or Docker

License: MIT license

Any restrictions to use by non-academics: No

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The pipeline is available in github: <https://github.com/FAANG/GSM-pipeline>. The data set used in the current study is available on ENA with accession number (ERR10825172-ERR10825181).

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank other members of the GENE-SWitCH data analysis group. This work is supported by the GENE-SWitCH programme which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998.

Funding

GENE-SWitCH has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998.

Authors contributions

JdV developed the extensions to the pipeline, created container and drafted the manuscript. CK assisted with containerization. OM, MG, MD contributed ideas and to the manuscript. All authors have read and approved the manuscript.

References

1. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* 2013 38:1. 2012;38:23–38.
2. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* 2019 20:10. 2019;20:590–607.
3. Beck D, Maamar B, Skinner MK. Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons. 2021. <https://doi.org/10.1080/15592294.2021.1924970>.
4. Paun O, Verhoeven KJF, Richards CL. Opportunities and limitations of reduced representation bisulfite sequencing in plant ecological epigenomics. *New Phytologist*. 2019;221:738–42.

3 GSM-pipeline

5. DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nature Biotechnology* 2017 35:4. 2017;35:316–9.
6. Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, et al. Sustainable data analysis with Snakemake. *F1000Res*. 2021;10.
7. Larssonneur E, Mercier J, Wiart N, Le Floch E, Delhomme O, Meyer V. Evaluating Workflow Management Systems: A Bioinformatics Use Case Genomic annotation curation using non-classic logic View project MicroScope View project Evaluating Workflow Management Systems: A Bioinformatics Use Case. 2018. <https://doi.org/10.1109/BIBM.2018.8621141>.
8. Abhishek M, Kumar Abhishek M. Containerization for shipping Scientific Workloads in Cloud. Article in *International Journal of Advanced Trends in Computer Science and Engineering*. 2020;9:5327–31.
9. Hunt GJ, Gagnon-Bartsch JA. A Review of Containerization for Interactive and Reproducible Analysis. 2021.
10. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*. 2020;38.
11. Rauluseviciute I, Drabløs F, Rye MB. DNA methylation data by sequencing: Experimental approaches and recommendations for tools and pipelines for data analysis. *Clin Epigenetics*. 2019;11:1–13.
12. Chatterjee A, Rodger E, Morison IM, Eccles M, Rodger EJ, Eccles MR, et al. Tools and Strategies for Analysis of Genome-Wide and Gene-Specific DNA Methylation Patterns The genes of life and death: a role for placental-specific genes in cancer? View project Haemoglobinopathies and Thalassaemias View project Chapter 15 Tools and Strategies for Analysis of Genome-Wide and Gene-Specific DNA Methylation Patterns. *Methods in Molecular Biology*. 2017;1537.
13. Acloque H, Harrison PW, Lakhali W, Martin F, Archibald AL, Beinat M, et al. Extensive functional genomics information from early developmental time points for pig and chicken. *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*. 2022;:2281–4.
14. Andrews S. FastQC: a quality control tool for high throughput sequence data – *ScienceOpen*. 2010.
15. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011;27:1571–2.
16. Pedersen BS, Eyring K, De S, Yang I V, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. 2014.
17. Picard Tools - By Broad Institute. <https://broadinstitute.github.io/picard/>. Accessed 28 Jul 2022.
18. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*. 2012;28:2678–9.

19. Ewels P, Magnusson M, Lundin S, Källér M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32:3047–8.
20. GitHub - dpryan79/MethylDackel: A (mostly) universal methylation extractor for BS-seq experiments. <https://github.com/dpryan79/MethylDackel>. Accessed 28 Jul 2022.
21. Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z. CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics*. 2018;34:381–7.
22. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13:1–9.
23. Guo W, Chung WY, Qian M, Pellegrini M, Zhang MQ. Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. *Nucleic Acids Res*. 2014;42:3009–16.
24. Guo W, Zhang MQ, Wu H. Mammalian non-CG methylations are conserved and cell-type specific and may have been involved in the evolution of transposon elements. *Scientific Reports* 2016 6:1. 2016;6:1–14.

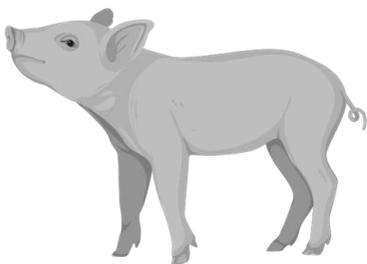
Chapter 4

Methylome dynamics regulating fetal development of seven tissues in pig

J de Vos¹, M F.L. Derks¹, H. Acloque², S. Djebali³, S. Foissac⁴, C. Guyomar⁴, C. Kurylo⁴, E. Giuffra², M A.M. Groenen¹, O Madsen¹

¹Animal Breeding and Genomics, Wageningen University & Research, The Netherlands; ²Paris-Saclay University, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France.; ³IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, 31024 Toulouse, France; France; ⁴GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, 31326, France

(In preparation)



*'If I have seen further than others, it is
by standing on the shoulders of giants'*

Isaac Newton

Abstract

DNA methylation is an epigenetic modification which plays a crucial role in mammalian development. However, there is still limited understanding of the overarching dynamics of DNA methylation in the developing fetus. In this study we used both whole genome- and reduced representation bisulfite sequencing data to evaluate the methylome dynamics in seven tissues (liver, kidney, brain, muscle, skin, small intestine and lung) during three developmental stages: 30 days post fertilization (30 dpf), 70 days post fertilization (70 dpf), and newborn in the pig. We further assessed differentially methylated expressed genes by combining the DNA methylation with the transcriptome of the same seven tissues. Developmental transitions were investigated using a two-fold approach: 1) based on whole genome methylation data dividing the methylome into unmethylated regions (UMR), indicative of possible promoters, lowly methylated regions (LMR) indicating possible enhancers, and fully methylated regions (FMR) and 2) based on both whole genome and reduced genome methylation performing differential methylation analyses. The number of UMR across developmental stages within the various tissues ranged from 10,822 to 14,680 and from 35,073 to 75,477 for LMR. The defined methylation states (UMR, LMR and FMR) were used to determine the dynamic changes of the methylome and associated cis-regulatory elements during fetal development. The most striking was the shift of methylation states, from hypo to hyper methylation, in liver 70 dpf stage to newborn stage. Differentially methylated expressed genes are involved in biological pathways related with general growth, e.g. regulation of developmental processes, during early stages of development (30 dpf to 70 dpf). This indicates a lack of complete tissue differentiation at 30 dpf, which is established at 70 dpf. This pattern was observed for most tissues, except in liver and brain, which showed a distinct methylome profile during early developmental stages.

Background

Epigenetics is the study of phenotypic changes that do not involve alteration in the DNA sequence. The epigenome is most often defined as many interdependent layers of epigenetic modifications, such as DNA methylation, histone modifications and chromatin accessibility, that affect the regulation of gene expression [1]. Variations in the genome and epigenome within regulatory genomic elements, e.g. promoters and enhancers, determine the specific level of gene expression in cells and tissues. Consequently, understanding variation in the functional regulatory genome will help explaining the differences between tissues and species, and thus improve the functional annotation of the genomes of species studied.

Regulation of gene expression is determined by a variety of epigenomic marks which work together to facilitate correct gene expression during a lifetime. One of the most important marks to regulate gene expression is DNA methylation, defined as the transfer of a methyl group onto cytosine at the 5th position forming 5-methylcytosine [2,3]. DNA methylation plays an important role in mammalian development, with critical roles in gene silencing, genomic imprinting [4], and X-chromosome inactivation [5,6]. In vertebrates methylation is mostly restricted to the CpG context, with four enzymes (DNMT1, DNMT3a, DNMT3B, TET) playing important roles in the deposition, maintenance, development, and removal of DNA methylation.

Numerous assays have been developed to investigate genome-wide DNA methylation. Bisulfite sequencing, which is considered as the current golden standard, is a method whereby unmethylated cytosine residues are converted to uracils and methylated cytosines remain unchanged. Site specific DNA methylation changes are detected through this methodology, however applying this method on a whole genome scale (whole genome bisulfite sequencing; WGBS) is costly. Reduced representation bisulfite sequencing (RRBS) applies genome-wide DNA methylation analysis with reduced sequencing of about 2% of the genome (primarily targeting CpG rich regions often present in e.g. promoter regions), which reduces the costs for investigating DNA methylation at large scale [7].

GENE-SWitCH is a project under the umbrella of the FAANG consortium, aimed at providing functional annotation of farm animal genomes. This project investigates the changes occurring in the functional regulatory genome specifically during embryonic development in two monogastric species, namely chicken and pig. Most studies thus far have predominantly been investigating adult tissue in livestock species, advocating for characterising the epigenome during development. During embryonic development three primary germ layers, the endoderm,

4 DNA methylome in the developing pig

ectoderm, and mesoderm, are formed, giving rise to specific tissues and organs. The germ layers develop in an intricate fashion, with the liver, pancreas, lining of the respiratory and digestive tracts developing from the endoderm layer, the mesoderm layer giving rise to tissues such as muscle, bones, and kidney, and from the ectoderm layer the nervous system, skin and sensory organs are developed.

The importance of dynamics of DNA methylation during mammalian development is apparent as the genome undergoes two waves of CpG methylation pattern reprogramming [6]. The first wave, which involves a large scale resetting of DNA methylation, occurs after fertilization, whereas the second wave occurs after germline cell specification [8]. During the second wave the reestablishment of DNA methylation on specific regions of the genome occurs. In this research we aimed to identify the dynamic changes in the methylome during development in pig which regulate gene expression during tissue formation (development). We specifically assess the changes in the methylome from 30 days post fertilization (dpf; early organogenesis) to 70 dpf (late organogenesis) and from 70 dpf to newborn.

Results

To study the methylome changes during development we analysed seven different tissues, representing the three germ layers, at two fetal stages of development (30 and 70 dpf) and in newborn piglets. We generated 63 RRBS datasets (3 samples * 3 timepoints * 7 tissues) and 21 WGBS datasets (1 sample * 3 timepoints * 7 tissues).

Pig methylome during development

Global cytosine methylation levels ranged from 8–10% for RRBS and 3–4% for WGBS. CpG methylation levels ranged between 60–80% for both RRBS and WGBS data sets and non-CpG methylation levels were between 0.38–0.65% and 0.95–1.5% for WGBS and RRBS, respectively (**Table S1**). Sample correlations within WGBS datasets exhibit moderate to high levels of correlation (0.6–1.00; **Fig. S1**), particularly among tissues originating from the same germ layer and developmental stage.

In the following section, we will primarily focus on genome-wide results using WGBS data, while results from RRBS data are mentioned in notable instances. The lowest CpG methylation levels are seen in liver at 30 dpf and 70 dpf (**Fig. 1a**), with an increase in liver methylation levels at the newborn stage but still lower than for other tissues at newborn stage. Similar patterns are observed in the distribution of methylation levels in liver, showing a lower percentage of fully methylated sites at 30 dpf and 70 dpf in comparison to liver newborn and other tissues (**Fig. S2**). The brain, skin, and muscle show a decrease in methylation levels at

4 DNA methylome in the developing pig

70 dpf followed by an increase at the newborn stage. Lastly, the kidney, lung, and small intestine show relatively stable overall methylation levels throughout development.

All samples show a clear decrease in methylation levels at the transcription start site (TSS) followed by an increase in the gene body (**Fig. 1b**). This trend is similar for all seven tissues at the various development stages, except for liver tissue at 30 dpf and 70 dpf showing the lowest methylation level at the TSS, and more pronounced low methylation within the gene body.

From the hierarchical clustering we observed a distinct cluster at the first developmental stage (30 dpf) for muscle, skin, kidney, lung, and small intestine (**Fig. 1c**). Interestingly, lung and small intestine, originating from the endoderm germ layer, consistently cluster together at each stage of development. Liver and brain, on the other hand, showed another pattern, with tissue specific clustering across all three stages of development, and forming separate branches from other tissues (**Fig. 1c**). Notably for RRBS data, the liver clusters on a separate branch for samples at 30 dpf and 70 dpf, while liver newborn clusters on a different branch (**Fig. S3a**).

PCA clustering analysis (**Fig. 1d**) revealed distinct patterns in tissue clustering at different developmental stages. Notably, brain tissue at 70 dpf and newborn tissues exhibited a close clustering pattern, while liver at 30 dpf and 70 dpf showed a similar clustering proximity. However, liver at the newborn stage was found further away from this cluster. These results are similar to PCA clustering results from RRBS (**Fig. S3b**).

Hierarchical and PCA clustering for RRBS data are shown in **Fig. S3** and we observed that in both WGBS and RRBS data, the majority of individual samples cluster based on tissue type at the two later developmental stages.

4 DNA methylome in the developing pig

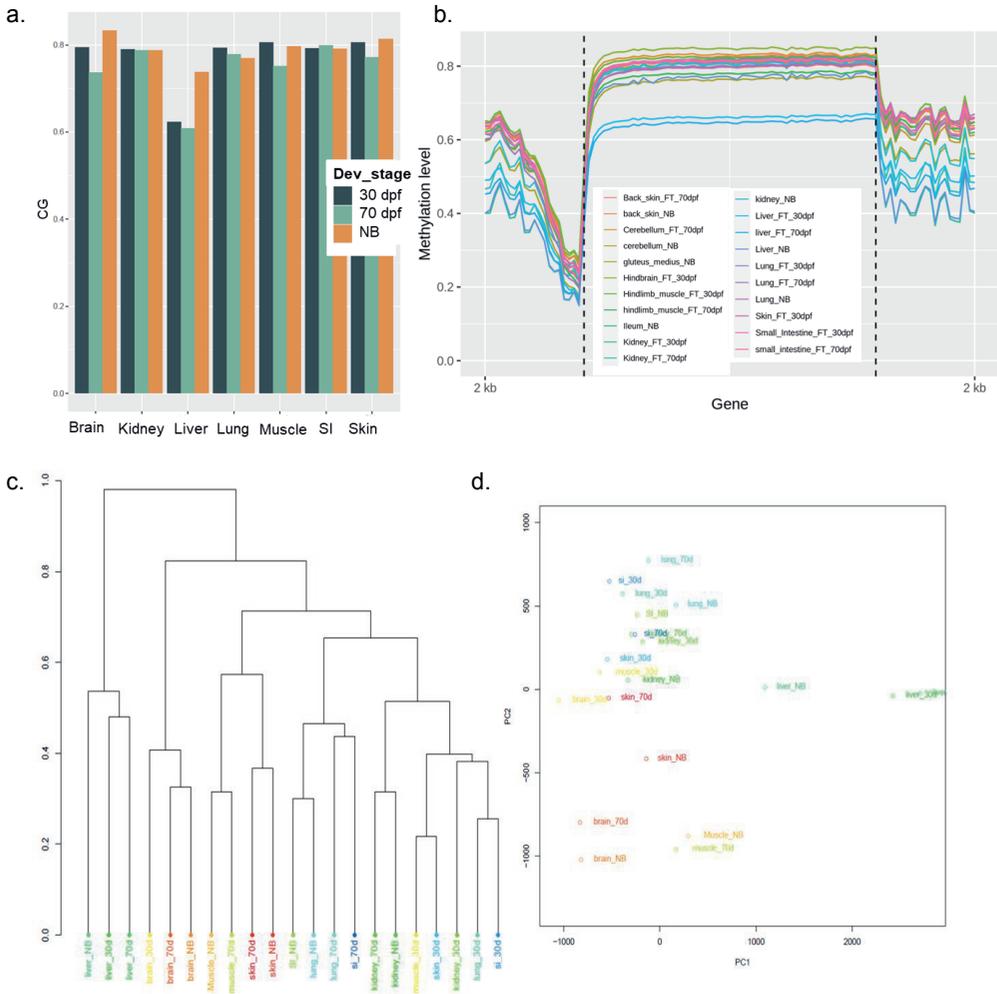


Fig. 1 Methylation patterns and distribution of the tissues at different developmental stages based on WGBS data **a.** CpG methylation levels for each tissue and timepoint. CpG methylation levels on the y-axis are actual methylation values ranging from 0-1, with 0 being unmethylated and 1 being fully methylated. SI=small intestine. **b.** Methylation levels across genes for each tissue per developmental stage from 2000 bp upstream, of the TSS, to 2000 bp downstream of the gene. **c** & **d.** Clustering of all tissue samples at 3 developmental stages. **c.** Hierarchical cluster using Ward.D2, with distances being measured from percent methylation per base. **d.** PCA clustering showing PC1 and PC2 with highest explained variance.

4 DNA methylome in the developing pig

Dynamics of DNA methylation in regulatory and repeat elements

We examined the methylation levels across different chromatin states as defined by Pan et al. 2021 [9] (**Fig. 2a, Data S1**). We observed lowest methylation levels in active promoters (E1), and low methylation levels in TSS-proximal transcribed regions (E3-E4), E2 (Flanking active TSS without ATAC (assay for transposase-accessible chromatin)) and E5 (Transcribed region without ATAC) showed moderate methylation levels. Among the enhancer states (E6-E10), the strongly active (E6) and medium active enhancers with ATAC (E7) exhibited the lowest methylation levels. Conversely, the enhancer states categorized as weakly active, poised, and active without ATAC (E8-E10) demonstrated higher methylation levels.

Furthermore, we evaluated the methylation patterns in different repeat types and families (**Fig. 2b, c, d, Data S2 and S3**). In general repeats are highly methylated (**Fig. 2b**). Only the LTR/ERVK and LTR/ERV1 repeat elements have somewhat lower methylation, especially in small intestine, lung, and muscle (**Data S2**). Methylation levels of LTR/ERVK elements show increased methylation levels at the 5' and 3' end of the repeat element and lower methylation patterns in between (**Fig. 2b**), however, for other repeats this pattern is less pronounced (**Data S2**). ERVK and ERV1 repeats are known to be associated with activation or repression of gene expression leading to somewhat lower methylation levels at these repeat families [10,11]. L1, L2 and ERV1 elements showed a peak in methylation patterns just upstream and downstream of the repeat elements (**Fig. 2c,d, Data S2**), while DNA transposons showed stable high methylation levels across the elements. Furthermore, liver showed low methylation levels in L1 and ERV1 in comparison to other tissues, indicating a systematic lower methylome profile for the liver in earlier developmental stages (**Fig 2c,d**).

4 DNA methylome in the developing pig

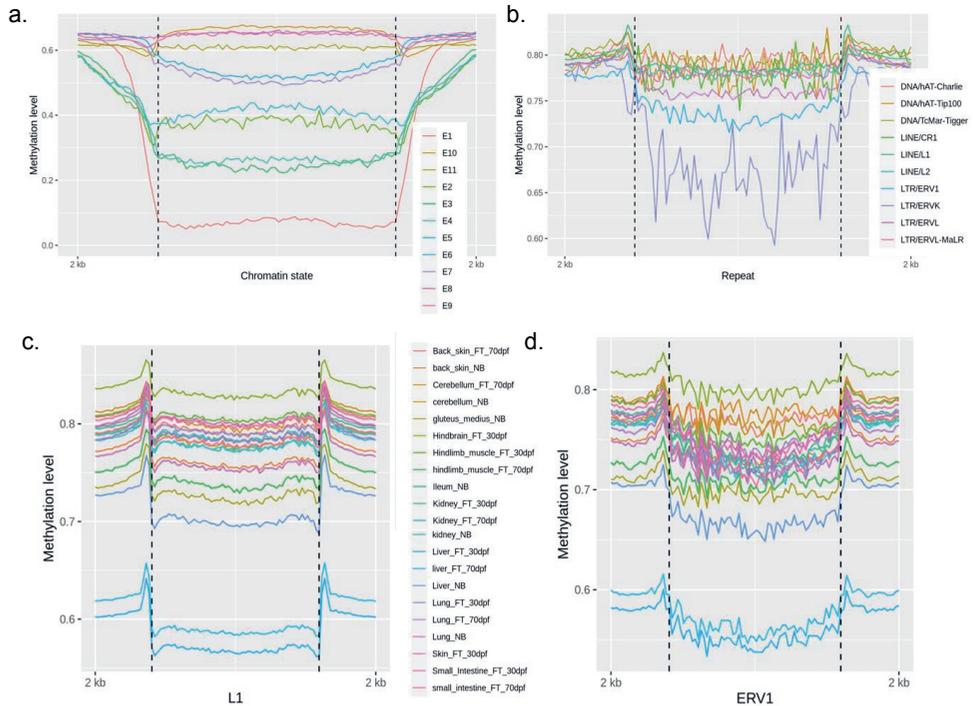


Fig. 2 Visualisations of the relationship between methylation levels, regulatory elements and repeats. **a.** Distribution of methylation levels (y-axis) within and surrounding different chromatin states in liver 30 dpf. Chromatin states: E1: Strongly active promoter, E2: Flanking active TSS without ATAC, E3: Transcribed gene (TSS-proximal transcribed regions), E4: Weak transcribed gene, E5: Transcribed region without ATAC. E6: Strong active enhancer, E7: Medium enhancer with ATAC, E8: Weak active enhancer, E9: Active enhancer without ATAC, E10: Poised enhancer, E11: ATAC Island. (Results for other tissues are available in **Data S1**). **b.** Distribution of methylation levels within and surrounding different repeat types and families in small intestine at the newborn developmental stage. Methylation levels ranging from 0 (unmethylated) to 1 (fully methylated) are shown on the y-axis. (Results for other tissues at developmental stages are available in **Data S2**). **c.** Distribution of methylation levels within and surrounding L1 **d.** Distribution of methylation levels within and surrounding ERV1 repeats (Results for other repeats are available in **Data S3**). Colours represent different samples.

4 DNA methylome in the developing pig

Methylation levels across gene expression classes during development

In order to investigate how DNA methylation changes relate to gene expression, we used transcriptome data collected from the same tissues and timepoints. As expected, we observe a negative correlation between TSS methylation levels and gene expression (**Fig. 3**, brain at newborn stage as an example). At higher levels of gene expression (TPM > 10) the methylation levels fall below 20% around the TSS. Unexpressed or very lowly expressed genes (0-1 TPM) showed a much higher methylation level around the TSS (~60%). We also observe lower methylation levels upstream and downstream (up to 2000 bp) of the gene body for highly expressed genes (TPM >= 75) with levels fluctuating around 20-30%. We observe high levels of methylation within the gene-body for all categories of gene expression.

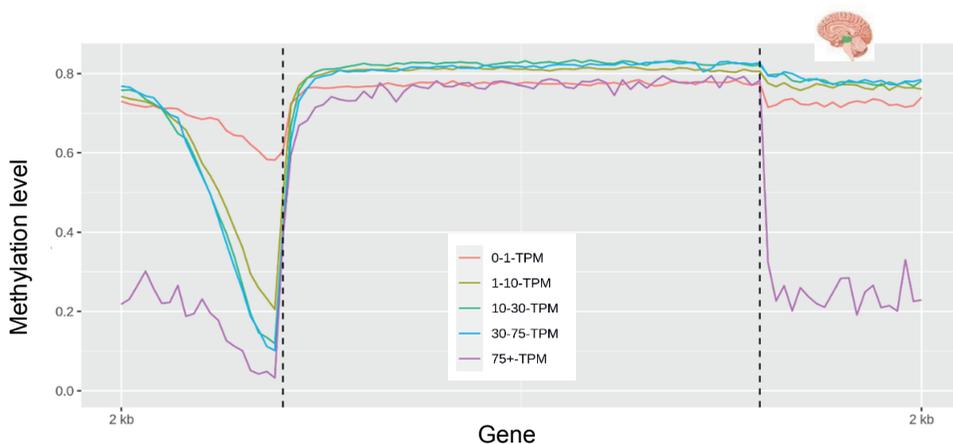


Fig. 3 Distribution of methylation levels upstream, downstream, and within the gene-body. Methylation levels ranging from 0 (unmethylated) to 1 (fully methylated) are shown on the y-axis. Categories of gene expression levels ranging from unexpressed and lowly expressed or very highly expressed genes are plotted in different colours. Y-axis shows the TSS as the first horizontal dotted line from the left, followed by the gene-body and downstream of the gene.

Dynamics of DNA methylation during developmental and maturation phases in seven pig tissues

Regulatory landscape during embryonic development

We used the detection of unmethylated and lowly-methylated regions (UMR/LMR) in WGBS data to identify regulatory elements like promoters and enhancers. As described in the

4 DNA methylome in the developing pig

methods section; a UMR has a methylation level between 0-50% and is CpG-rich (indicative for promoter regions) while an LMR is defined as regions with 10-50% methylation levels and being CpG-poor indicative for enhancer regions [13]. Given these parameters we identified between 10,822-14,680 UMR and 35,073-75,477 LMR in the various tissues at different developmental stages (**Fig. 4b, c**). General statistics of the length per region across tissues and time-points, as well as the number of promoter regions associated per region are shown in **Table S2**. Tables of regions and associated genes can be found in **Data S4** (UMR) and **S5** (LMR).

Close proximity of UMR and LMR to TSSs suggests overlap with potential cis-regulatory elements (**Fig. 4a**). UMR are found to be in closer proximity to TSSs compared to LMR which are more enriched further upstream of TSSs. This supports UMR and LMR as indicative for promoters and enhancers, respectively. We investigated the dynamic nature of the methylome, specifically changes in UMR and LMR during development using brain and liver as illustrative examples with interesting patterns (**Fig. 5**). Results for the remaining tissues are presented in **Fig. S4**.

In brain (**Fig. 5a**) a large proportion of fully methylated regions (FMR) transition to LMR, while a smaller proportion transition to UMR. Additionally, a small proportion of the LMR transition to FMR and UMR, respectively. During the transition from 70 dpf to newborn, a substantial proportion of LMR remains as LMR, while a small proportion transitions to FMR and UMR. Only a limited number of UMR transition to FMR and LMR, with the majority remaining in an UMR state. Notably, a large proportion of FMR transitions to both UMR and LMR.

4 DNA methylome in the developing pig 

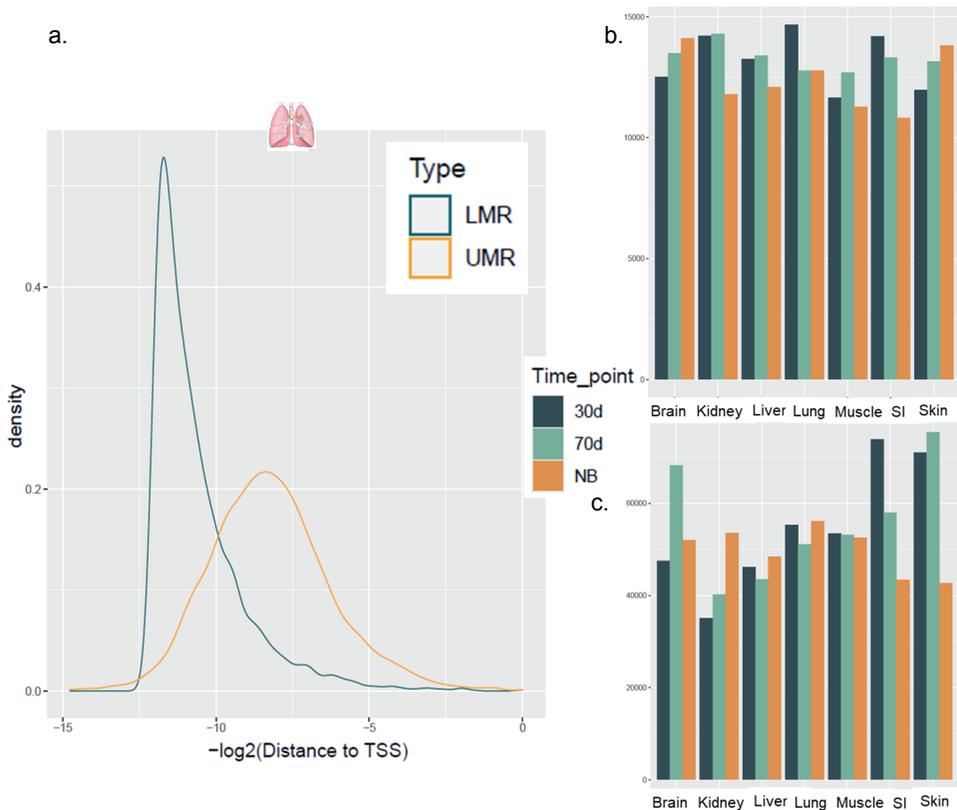


Fig. 4 a. Descriptions of UMR and LMR from WGBS data of lung at 30 dpf. Density plot showing the distribution of the UMR (yellow) and LMR (blue-green) from TSSs. Y-axis shows the densities of the regions (UMR and LMR) and x-axis is the log distance of the regions to the TSS. **b.** Number of UMR per tissue and developmental stage with small intestine abbreviated as SI **c.** Number of LMR per tissue at the developmental stages.

In contrast, in liver (**Fig. 5b**), the number of LMR is smaller compared to brain, with a significant proportion of LMR remaining unchanged during the first transition. FMR in the liver tend to transition to UMR and LMR over the course of development. **Table S3** provides information on the number of transitions and the associated promoter regions for each tissue and tables of all regions transitioning together with associated genes are shown in the **Data S6**.

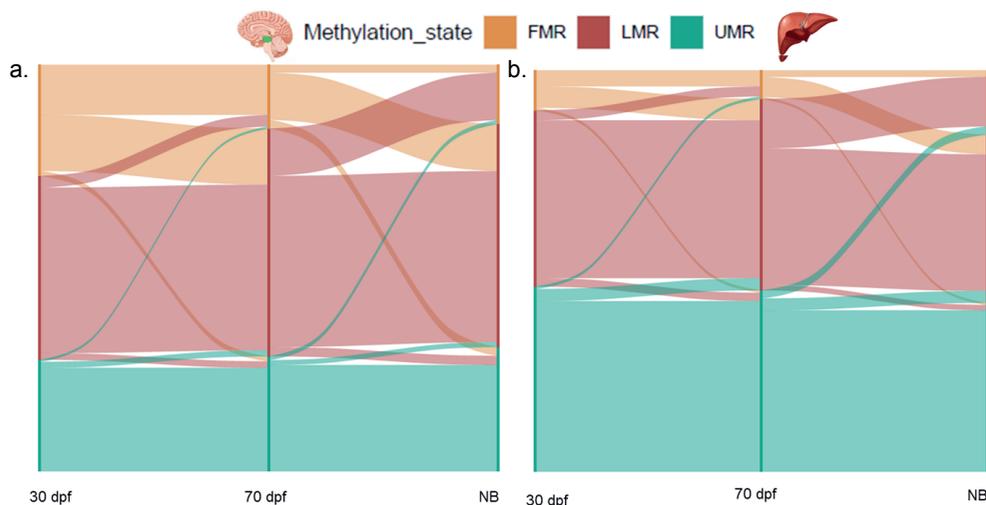


Fig. 5 Time-point specific changes occurring in **a.** Brain and **b.** Liver. Developmental stages are annotated in different colours, with orange showing FMR, red showing LMR and green showing UMR. For clarity, it should be noted that any FMR that did not exhibit changes across all developmental stages were excluded from this plot, focusing solely on conservation and dynamic changes of UMR and LMR.

Differential methylation analysis during developmental and maturation phases

An additional approach that sheds light on the time-point specific changes that occur during development, is differential methylation (DM) analysis. For DM analyses the transition from 30 dpf to 70 dpf is classified as the developmental transition, while the transition from 70 dpf to the newborn stage is considered the maturation transition. We performed DM analyses on a per site level (DMS) using RRBS data (3 samples per tissue per time-point) and on a region level (DMR) using WGBS data (1 sample per tissue per time-point).

For all tissues, except liver, more DMS are observed in the developmental transition compared to the maturation transition (**Table 1**). Likewise, the number of DMS in promoter regions is also higher for all tissues, except liver, in the development transition compared to the maturation transition. Both DMS and DMR analyses revealed intriguing patterns in the liver, characterized by a remarkable increase (~6 fold for DMS and ~40 fold for DMR) in the number of regions. At

4 DNA methylome in the developing pig

the developmental transition, the liver exhibited the lowest number of both DMS and DMR, but this changed to the highest number of DMS and DMR at the maturation transition. Conversely, during the developmental transition, muscle displayed the largest number of DMR, which notably decreased in the maturation transition, mirroring the trends observed in DMS. Tables of all DMR and DMS together with associated promoter of genes are shown in **Data S7** (DMS) and **S8** (DMR).

Table 1 Counts of DM analyses from both RRBS and WGBS data of pig tissues with 25% difference and q-value <0.01. DM was done in a linear fashion, with developmental transition indicating differences between 30 dpf to 70 dpf and maturation transition indicating differences between 70 dpf to newborn developmental stage, respectively.

Tissue	RRBS (DMS)				WGBS (DMR)			
	Developmental transition (Dev)	Maturation transition (Mat)	Promoter of genes		Developmental transition (Dev)	Maturation transition (Mat)	Promoter of genes	
			Dev	Mat			Dev	Mat
Skin	20,042	6,680	868	262	246,864	214,496	3,215	2,624
Brain	32,334	19,843	1,022	608	207,490	214,798	2,488	2,376
Muscle	30,975	6,848	1,330	245	335,293	105,979	4,343	1,350
Kidney	22,711	2,039	838	79	93,401	111,248	1,317	1,333
Liver	8,023	50,303	428	1,543	23,088	975,926	383	9,880
Lung	16,697	2,880	710	125	106,805	119,430	1,600	1,455
Small intestine	16,112	8,709	814	379	129,262	155,656	1,875	2,070

For further understanding of the methylation dynamics during development, we defined hypo-methylated regions as DMR with a decrease in methylation levels (<25%), and hyper-methylated regions as DMR with an increase in methylation levels (>25%). **Fig. 6a** shows the distribution of hyper- and hypo- DMR for the tissues during development, with liver showing the highest number of hyper-methylated regions at the maturation stage. During the developmental transition, muscle tissue exhibits a notable abundance of hyper-methylated regions, particularly in intergenic regions and within the gene-body (**Fig. 6b**). Brain tissue shows more hypo-methylated regions in comparison to hyper-methylated regions during both developmental transitions, which is a common trend observed in most tissues.

DMS provides valuable insights into the promoter specifically as RRBS is enriched in CpG regions, whereas DMR predominantly covers regions within the gene body and intergenic regions. Liver shows a remarkable number of hyper-methylated DMS at the maturation transition (**Fig. S5**), which is similar to DMR patterns in the liver (**Fig. 6a**). Furthermore, the

4 DNA methylome in the developing pig

brain shows a large number of hyper-methylated DMS at the developmental transition, which changes to a larger number of hypo-methylated DMS at the maturation transition, with only a small proportion of hyper-methylated DMS (**Fig. S5**).

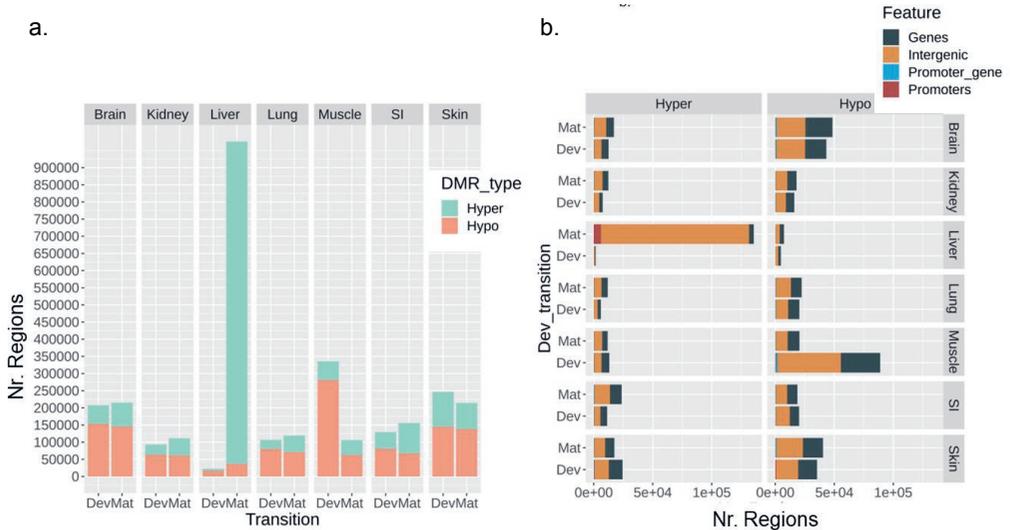


Fig. 6 Distribution of DMR identified using WGBS data. SI=small intestine. **a.** Number (y-axis) of hypo and hyper methylated regions per tissue (top x-axis), for each developmental transition. Dev is defined as developmental transition (30 dpf to 70 dpf) and Mat is defined as maturation transition (70 dpf to newborn). Hyper and hypo methylation are shown in green and red as stacked bars. **b.** Annotation of the hyper and hypo methylated regions. Dev and Mat are as defined in a. Features are defined as follows: Intergenic is 1000 bp upstream of TSS until next TSS. Genes are defined as the gene body. Core promoter is 500 bp upstream of the TSS. Promoter_gene are regions which overlap both core promoter and gene body.

Tissue specific DMR, DMS & genes

We identified both tissue specific DMR and DMS, and subsequently combined them with gene promoter regions. **Table 2** shows the number of DMS and DMR with the corresponding promoter region for each tissue. We observe the highest number of tissue-specific DMR and DMS for liver while the lowest number of unique DMR is observed for kidney. No significant enrichment of GO terms were identified for the DMR tissue specific genes, while GO terms enriched for DMS tissue specific genes are shown in **Table S6**. This showed an enrichment of the biological process supramolecular fiber organization in skin specific DMS and extracellular space in liver specific DMS.

4 DNA methylome in the developing pig

Table 2 Tissue specific differentially methylated regions and sites, with the corresponding genes

Tissue	DMS	DMS Gene	DMR	DMR gene
Kidney	211	158	834	617
Liver	710	528	13,304	7,427
Lung	161	147	1,120	802
Brain	400	289	1,838	1,198
Muscle	392	262	3,149	2,191
Skin	227	189	2,785	1,928
Small intestine	304	208	1,261	845

We conducted a gene-level investigation of selected genes known for their roles in embryonic development, tissue differentiation, and organogenesis, to gain further insights into the methylome dynamics at a tissue specific level. A few genes identified for tissue specific DMS include the *FGF21* gene in muscle, *CDX4*, *HOXD12*, *HOXB5* in small intestine, *HOXA5* in skin, and *HOXC8*, *HOXD8*, *HOXA11*, *HOXC5* in muscle.

Furthermore, for tissue specific DMR we identified *PAX3*, *HOXC5* in muscle, *HOXC11* in kidney, *HOXD4*, *HOX2A*, *HOXC10*, and *HOXC9* in the small intestine, *FGF1* in brain, and *CDX1*, *CDX4* in liver. Tables of tissue specific DMR and DMS together with associated gene promoters are shown in the **Data S9** (DMS) and **S10** (DMR).

DM-associated differential expressed genes (DMEG) involved in functional enrichments related to development and tissue specificity

Lastly, the regulatory effects of DNA methylation were investigated by integrating differentially expressed genes (DEG) with both DMR and DMS, respectively. Approximately 2-20% of DEG identified in the various tissues at the developmental transition and 5-50% of DEG identified at the maturation transition were associated with corresponding DMR at the promoter sites. Additionally, approximately 1-8% of DEG identified for the various tissues were associated with the corresponding DMS (**Data S11**, **Table S7**). In this section, we focus primarily on the DMR integrated with differentially expressed genes, referred to as DMEG (**Fig. 7a**).

We observed the largest number of DMEG at the developmental transition in muscle and skin and the lowest number in liver tissue (135) and kidney (389) (**Fig. 7a**). For most tissues the number of DMEG either decreased slightly in the maturation transition, like in skin and muscle or changed very slightly like in brain, small intestine, lung and kidney. For liver, a large increase in DMEG was observed in the maturation transition with more than 3,000 observed DMEG.

4 DNA methylome in the developing pig

We distinguished between up-regulated and down-regulated genes (based on expression fold change) with hypo or hyper-methylated (DMEG) associated promoters in brain tissue during the maturation transition (**Fig. 7b**). For up and down-regulated genes with hypo or hyper-methylation the results showed the largest number of DMEG which are hypo-methylated and up-regulated (hypo-up), a smaller number of hyper-down and hypo-down genes, and the smallest number of genes in the hyper-up category (**Fig. 7b**).

Subsequently, we first investigated the identified genes using functional enrichment analysis within the up-regulated (combined hyper- and hypo-methylated) and down-regulated (combined hyper- and hypo-methylated) categories. The overrepresentation gene ontology (GO) analyses of up- and down-regulated categories of DMEG (**Fig. 7c, d**) provides insight into the biological systems involved during development at the two transitions and within the different tissues. Tables of DMEG are available in **Data S11**, together with the heatmaps of promoter regions of DMEG (**Data S12**). Below we present the most interesting global findings of the GO analyses for genes upregulated (switched on) and downregulated (switched off) during the two transitions for DMEG.

Up-regulated genes (switched 'on')

Most notable is the variation of significantly enriched terms per tissue, with tissues like liver and lung showing no significantly enriched terms at the developmental transition while several terms are observed at the maturation transition for biological processes (**Fig. S6a**). Regulatory processes are observed as up regulated during the developmental process in kidney, muscle, skin and small intestine (**Fig. S6a**). Defence, immune and response to stimuli processes are enriched for kidney, liver, lung, skin, and small intestine at the developmental as well as maturation transition (**Fig. 7c, Fig. S6a**). Epithelial migration is notable during developmental transition in skin, however it is not enriched in small intestine which is not as expected. Tissue specific cellular component terms such as contractile fiber, myofibril, myosin complex, sarcomere, Z disc and I band are observed during developmental transition in muscle, with mitochondrial terms and muscle specific terms observed during the maturation transition in muscle (**Fig. 7d**).

On an individual gene level, we investigated genes that are known to play crucial roles in embryonic development, organogenesis, and germ layer formation. *WNT16*, *WNT5b*, and *WNT9B* genes are up regulated in lung, small intestine, and skin, respectively, during the developmental transition. *MSTN (myostatin)*, is up-regulated in brain, muscle and small intestine during the developmental transition. *IGF1* is up-regulated in the brain during the developmental transition. A few examples of HOX genes identified include *HOXB5* in skin, *HOXD13*, *HOXD11*, in small intestine and *HOXA11*, *HOXD11* in kidney are up-regulated

4 DNA methylome in the developing pig

during the developmental transition. Similar results are observed for DMS associated with differential expressed genes (**Data S11**).

Down-regulated genes (switched 'off')

Similarly, for down-regulated genes there are many biological process terms enriched for liver at the maturation transition and none at developmental transition. Liver is also the only tissue with significant GO enrichment for biological process of down regulated genes during the maturation transition (**Fig. S6b**). Interestingly, in many tissues significant biological process terms are general developmental terms such as limb development and embryonic morphogenesis. Lastly, down-regulated enrichment analyses provide insight into processes which are 'switched off' at various stages of development. Most notable are biological processes such as vasculature development, tube morphogenesis, tissue development, regulation of developmental processes, immune system processes, embryonic morphogenesis, embryo development, angiogenesis, and circulatory system development enriched in muscle, skin, and small intestine (**Fig. S6a**). Kidney shows one notable term enriched in biological process and cellular component ontologies namely A-band (**Fig. S6 c, d**).

Furthermore, we investigated the presence of a few key genes involved in embryonic development, tissue differentiation, and growth. Noteworthy observations include the down-regulation of *PAX3*, in both muscle and brain at the developmental and maturation transitions respectively. The small intestine shows down regulation during the developmental transition for *WNT5a* gene. Additionally, *MYOD* and *MSTN* (*myostatin*), show down regulation in the liver and small intestine, respectively, at the developmental transition. During the developmental transition in brain tissue, the *CDX1* gene is down-regulated. This succinctly highlights a subset of identified genes, and the results for DMEG and DMS associated with differentially expressed genes can be found in **Data S11**.

4 DNA methylome in the developing pig

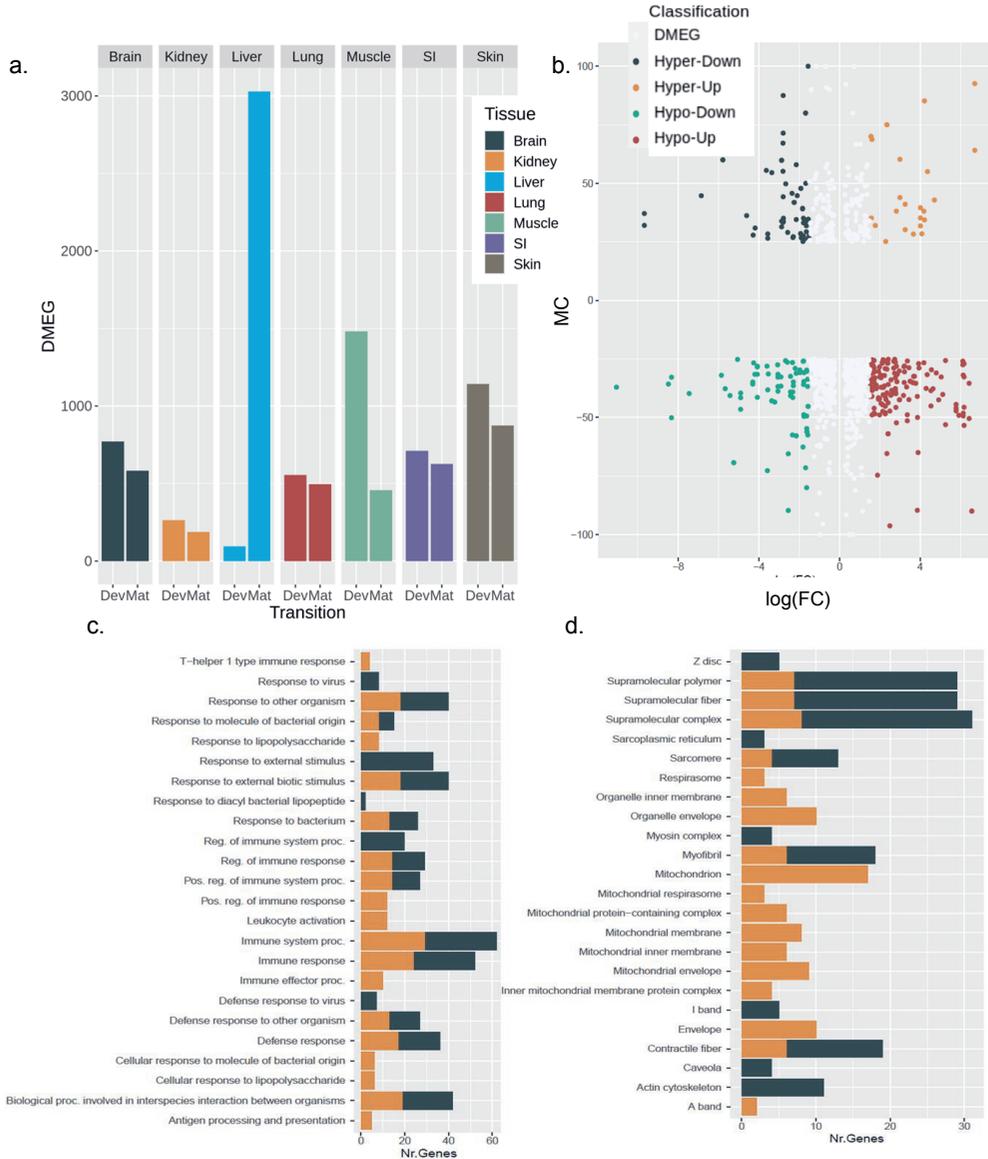


Fig. 7 Integration of DMEG on a temporal level for the various tissues. **a.** Number of differentially methylated expressed genes (DMEG) for the developmental and maturation transition (x-axis) with the number of genes shown on the y-axis. Tissues are shown in different colours. **b.** Scatter plot of the methylation fold change (y-axis) and the log expression fold change (x-axis) for brain tissue at the maturation transition. Each point represents a DMR-DEG pair. Genes classified as Hyper-Up, Hypo-Up, Hyper-Down and Hypo-Down have an expression log(FC) with a cut-off at 1.5. The remaining genes are shown in white, with a log(FC) between 1.5 and -1.5. All methylated regions have a minimum difference of 25%. **c.**

4 DNA methylome in the developing pig

Biological processes gene ontology (GO) terms for small intestine up-regulated genes which are both hypo and hyper methylated. **d.** Cellular component GO terms for muscle up-regulated genes. All terms are filtered for an $FDR < 0.05$. Genes are filtered as follows for classifications: up-regulated $\log(FC) > 1.5$, and down-regulated $\log(FC) < -1.5$. All methylated regions have a minimum difference of 25%. GO enrichment analyses of other tissues can be found in **Fig.S6**.

Discussion

Development is a highly dynamic process with many changes in gene expression and the epigenome. DNA methylation is an epigenetic mechanism which has a well-established link to development and differentiation of tissues [6,14–16]. Here we investigated the dynamic nature of the DNA methylome during development of seven tissues in pigs. The nature of this data enabled the identification of developmental specific changes occurring within the methylome of different tissues. Although the DNA methylome has been studied in model organisms such as mouse, this is the first study investigating development of various tissues representative of the three germ layers in a farm animal. Research on human fetal development is limited. Hence, research into DNA methylation regulating fetal development in a biomedical model species such as the pig holds potential for understanding fetal diseases, as well as fetal development in mammals. Our work provides the first DNA methylome map in seven key tissues of the pig during fetal development from 30 dpf until newborn piglets.

Global methylation patterns across tissues at the different developmental phases provides insight into general methylation patterns occurring across the genome. Liver has lower global methylation levels, especially during the first two stages of fetal development, in comparison with the other tissues investigated in this study. Thus, during earlier stages of development accessed in this study (30-70 dpf) there are less fully methylated sites in liver, and during the transition from early development to newborn there are considerable methylation events occurring. Patterns of lower global methylation and a bias towards demethylation events in the liver is also observed in mouse and human (in-vitro) during development [17–19]. Higher global methylation levels across tissues such as fetal hindbrain, kidney, lung, and intestine during development are also observed in the mouse [19,20]. In a study conducted by Yang *et al.*, [21] investigating DNA methylation patterns during skeletal muscle development in mouse from embryo day 30 until 180 days post-natal, similar findings were observed regarding global methylation levels in muscle tissue.

4 DNA methylome in the developing pig

Clustering at different time-points based on similar global methylation patterns (WGBS) for muscle, skin, small intestine, lung, and kidney, show similar patterns of DNA methylation at 30 dpf. Tissue specific segregation within these tissues start to occur at 70 dpf and newborn. The lung and small intestine are from the endoderm germ layer [22] and these tissues show similar patterns of regulation by DNA methylation. Brain and liver already show distinct tissue specific DNA methylation patterns from the first developmental time-point and both tissues cluster separately from the remaining tissues. This pattern is not supported by the RRBS clustering, suggesting that CpG rich regions (RRBS) may have distinct changes in methylation compared to global methylation (WGBS) during development of these two tissues.

We inferred potential cis-regulatory elements based on the methylation level and density of CpG sites and found that these regions varied across tissues and time-points. In the brain and skin an increase in number of UMR from early development to newborn indicates an increase in the number of active promoters. This shows a de-methylation occurring in these tissues from early development to newborn. Typically, methylation is observed from early development to newborn in the tissues such as kidney, liver, lung, muscle, and small intestine. This is the expected mechanism occurring during development where many de-methylation and re-methylation events are observed, with many regions being re-methylated within newborn and adult tissues [21]. By integrating methylation classes (UMR, LMR, FMR) and differential methylation on a spatio-temporal scale, we demonstrated the dynamic nature of the methylome, as well as the dynamic repurposing of promoter and enhancer elements. The inference of developmental enhancers by using the DNA methylome has also been implemented in zebrafish [23]. In the liver most regions become re-methylated from fetal development to the newborn stage, which has a possible link to enhancer and promoter elements, as most sites are found within intergenic and promoter regions. Muscle shows a high degree of re-methylation during the developmental transition, with most of these changes occurring in intergenic and gene-body regions. This is in agreement with earlier findings in porcine skeletal muscle during fetal development where most DMR are found in cis-regulatory elements such as enhancers [21].

Liver shows quite different patterns of methylation in comparison to other tissues. Research examining DNA methylation and gene expression changes in human liver during the transition from fetal to adult stages has revealed an increase in methylation levels. Surprisingly, these changes do not appear to play a regulatory role for gene expression nor show the expected decrease in gene expression, pointing towards distinct DNA methylation dynamics during fetal and adult liver development [24]. In contrast, recent research in mice [19] have shown a noteworthy pattern in the liver, where methylation levels are significantly higher (~ 0.75) during early developmental stages, then decline during early and late organogenesis (~ 0.6), and

4 DNA methylome in the developing pig

subsequently rise as the organism progresses to the newborn (~0.7) and adult stages (~0.8) of development. These patterns align with our study's findings. Notably, while this pronounced shift is not reflected in the number of differentially expressed genes, this suggests an alternative regulatory role such as alternative splicing, in the fetal liver.

During early development in the pig, basic primordial organs such as the liver, spleen, lung buds, and primordial intestine can be distinguished at 18-20 dpf [25]. By 35 dpf, the liver completes its main organogenesis, making it a key tissue in the embryo. This is most likely linked to the process of haematopoiesis, which starts in the yolk sac, but becomes non-functional at 24-27 dpf [25,26]. Subsequently, hematopoietic activity transitions to the fetal liver from 20 dpf and eventually progresses to the bone marrow. These observations suggest that the liver plays an important functional role in haematopoiesis from 30 dpf, a process which gradually decreases until birth. Notably, in mice, the liver's distinct methylation profile coincides with fetal liver haematopoiesis [25,26].

Integration of transcriptomic and methylation data provides insight into their complex relationship, especially since DNA methylation in general is considered an epigenetic repressor of gene expression. Typically, hypo-methylation in the promoter and enhancer regions is associated with upregulation of genes while hyper-methylation is associated with downregulation. The association of promoter hypo-methylation with gene expression is also supported in this study where we observe global decreased promoter methylation related to gene expression (**Fig. 6**). However, we also found associations between hyper-methylation and gene up-regulation as well as hypo-methylation and gene down-regulation in all tissues across development. Thus, our results show that the association between gene expression and DNA methylation is variable and complex, similar to recent observations in Anastasiadi *et al.*, [27].

The GO enrichment analyses provided some interesting observations with muscle showing many muscle related functions and components which are present at the early developmental stage, while at the later stage of development, elements related to energy (mitochondria) are observed. The muscle fiber is a complex design of myofibrils, mitochondria, and nuclei, encapsulated by the sarcolemma. The sarcomere is the most basic unit of the myofibrils which are composed of actin and myosin (thick and thin filaments). These components are arranged as follows: I-band contains thin actin filaments, the A-band contains mainly myosin thick filaments and actin thin filaments on the edges, and the Z-line indicates the border of each sarcomere. These components of the muscle fibre work together in a complex fashion to contract and relax [28–30]. We identified the basic components of the muscle fiber such as sarcomere, contractile fiber, Z-disc and I-band during earlier developmental transition, thus

4 DNA methylome in the developing pig

from 30 dpf to 70 dpf, followed by development of components such as mitochondria, A-band and sarcoplasm at 70 dpf to newborn. Similar observations have been made for skeletal muscle development in pig and other mammals such as human, cattle, and sheep, where embryonic morphogenesis occurs first, followed by muscle tissue development, and thereafter glucose metabolism [21,29,31–34].

Kidney and muscle develop from the mesoderm layer [35], and the observed down-regulation of the A-band pathway (muscle fibre component) in kidney during developmental transition in our results is notable. It shows the germ layer specificity present within this tissue at 30 dpf which becomes tissue specific from 70 dpf onwards. Similarly, many general developmental processes are ‘switched off’ from 30 dpf to 70 dpf in muscle, skin, and small intestine. Lastly, the gastro-intestinal tract, which includes the small intestine, plays a role in providing passive immunity to newborn mammals from colostrum ingested shortly after birth (within the first 24 hours of birth). This process is crucial for long term survivability and establishment of intestinal microbiota, which is important for maturation of the intestinal immune system [36,37]. Interestingly we observe immunity GO terms enriched at the maturation transition for up-regulated genes in the small intestine, which confirms the development of the intestinal microbiota and intestinal immune system in the pig.

Genes (e.g. *HOX*, *PAX*, *WNT*) previously identified for their role in embryonic development, tissue differentiation, and growth are identified across multiple tissue types. This shows the importance of these genes in shaping the developmental trajectory of different tissues within our dataset. Interestingly, the *MSTN* (myostatin) gene, plays a crucial regulatory role in muscle development and is highly conserved in mammals [38,39]. Genetic variations within the *MSTN* gene have been linked to the double muscling phenotype observed in several species, including cattle, sheep, dogs, and humans [40].

This is the first research to investigate the dynamics of DNA methylation in fetal development in pig in different tissues representing the three embryonic germ layers. The pig is used frequently as a biomedical model for humans [41–45], and in this research area there have been very few studies performed in human. Implications of a better understanding of fetal development can be utilised in future mammalian research, pharmaceutical assessments during pregnancy, and studies of human fetal diseases such as e.g. intrauterine growth retardation and polycystic ovary syndrome and preeclampsia which naturally occurs in both pigs and humans [45,46].

Conclusion

We investigated the development of seven tissues in the pig, from early fetal development until newborn piglets. Through a variety of analyses the complex dynamics of the DNA-methylome

4 DNA methylome in the developing pig

at a spatio-temporal level was assessed. Differentially methylated sites and regions were identified, together with categories of methylation, which were combined for insight into the developmental methylation dynamics. Differentially expressed genes were integrated with the methylation status, which added an additional layer of insight into the complex relationship between this regulatory mechanism and gene expression. We identified tissue specific changes, and we conclude that liver shows remarkably different differentiating patterns in comparison to other tissues. Lung, muscle, kidney, skin, and small intestine have germ layer specific methylation characteristics at early development (30 dpf), which then transitions to tissue specific methylation from 70 dpf until newborn stages.

Material and Methods

Experimental model and tissue collection

This study investigated seven different tissues namely brain (hindbrain at 30 dpf , cerebellum in 70 dpf and newborn), liver, kidney, skin, muscle (hindlimb muscle at 30 dpf and 70 dpf, *Gluteus Medius* in newborn), lung, and small intestine (ileum in newborn) in pig. The samples of these tissues were collected at 30 dpf, 70 dpf and at newborn piglet stage. Pooling and sampling strategies of samples at 30 dpf are described in more detail on the FAANG DCC. In brief, after dissection the fetuses were sexed using a standard PCR procedure. Pooling was performed to obtain two pools of females and two pools of males, exact combinations of fetus in each pool are described in the standard operating protocol (SOP). Once pools were established, the tissues corresponding to each sampling pool were cut up and distributed into four tubes and stored at -80 °C. Samples collected at 70 dpf are described in full detail in the SOP available on the FAANG DCC. In summary, animal dissection was performed after the uterus of the slaughtered sow was removed and the fetuses were removed as quickly as possible. Tissues were removed from the fetuses in a predetermined order and each tissue was placed into the plate which has the tissue name and animal number. Tissues were then processed and cut into small cubes and stored in a cryotube. Thereafter samples were snap frozen in liquid nitrogen and after transportation (dry ice) to the laboratory, were cryofrozen. These samples were also sexed using the afore mentioned PCR procedure. Lastly the newborn samples were collected by stunning and slaughtering the piglet, and at dissection the tissues were removed. Each tissue was placed on a petri dish which were pre-labelled with the animal number and tissue name. Each tissue sample was then cut into cubes, washed in PBS and stored in an empty cryotube (each tube contains six aliquots). Freezing and conservation process of the samples were as described at 70 dpf. The full description of the procedure is described in the SOP on FAANG DCC.

4 DNA methylome in the developing pig

Sample processing and bisulfite sequencing (RRBS and WGBS)

DNA extraction was performed on each tissue sample at each developmental stage as described on the FAANG DCC [SOP](#). Firstly, a preparation step was needed where each tissue was weighed and excised. Thereafter each piece of tissue was placed into a centrifuge tube to complete a homogenization step. Genomic DNA was purified using the Qiagen Allprep kit, and quantification was performed to assess the quality of the DNA and thereafter stored. Sexes selected for methylation sequencing were as follows, RRBS: 30 dpf – pooled sample, 70 dpf - male, newborn – male; and WGBS: 30 dpf – pooled, 70 dpf - male, newborn - female.

[RRBS](#) and [WGBS](#) library preparation is described in the SOP available on the FAANG DCC. In brief the RRBS library was constructed with the Ovation RRBS Methyl-Seq kit from Tecan and sequenced on the Illumina NovaSeq 6000. The isolated DNA was cut with MspI enzyme and fragments were selected at 50-500 bp. After sequencing the bcl2fastq v2.17.1.14 conversion software was used to generate and demultiplex fastq files from pooled samples. WGBS library construction and sequencing was performed using the Illumina NovaSeq 6000 at Novogene, with the bisulfite conversion performed using the EZ DNA Methylation Gold Kit (Zymo Research) and fragments selected and quantified using the KAPA HiFi HotStart ReadyMix (Kapa Biosystems). Library was then sequenced with a PE150 strategy.

Data processing & clustering analysis

Raw RRBS FASTQ files were first trimmed using the NuGEN Technologies' [diversity trimming](#) scripts, for removal of any additional sequences added by the diversity adaptors. Subsequently reads were filtered for only reads which begin with the expected YGG trinucleotide sequence. On average 38 million reads were produced for each library of RRBS (21,7 – 46 million) and an average of 426 million reads for WGBS (range of 390 – 475 million). Coverage of all reads was investigated and an average of 19X for RRBS and 25X for WGBS data was calculated. Quality statistics of both data sets, per tissue-time-point is shown in **Tables S4** (WGBS) and **S5** (RRBS). We determined the bisulfite conversion rate (~99.5%) for samples by calculating the ratio of unmethylated to total number of reads for covered cytosines on the mitochondrial genome for RRBS data sets. This is used as a quality metric to determine the efficiency of the bisulfite conversion. Diversity trimmed reads of RRBS samples (single end) and raw WGBS FASTQ files (paired end) were processed using the [GSM-pipeline](#) [47], v1.0 used for RRBS and v2.0 for WGBS, with [Sus Scrofa 11.1](#) reference genome assembly. Reads were first trimmed with TrimGalore tool [48] using default options. Read alignment was performed within the pipeline using Bismark aligner with default options, and for RRBS the parameter --rrbs was specified. WGBS data was first processed per sample, time-point per lane. Thereafter the bam

4 DNA methylome in the developing pig

files were merged to have data per tissue time-point and then processed using the pipeline with the `--bam` parameter. Binary alignment files were converted to CGmap files using CGmapTools within the pipeline, and reports of methylation rates for all covered CpGs are reported in the CGmap file, as well as in an extensive bismark methylation report. The pipeline reports on global methylation rates per sample. The CGmap file is converted to CpG report text files and processed using methylKit within the pipeline, which provides methylation distributions per sample. Correlations and clustering were calculated between various tissues at the three developmental stages, as well as correlations between tissues from the same germ layer, and developmental stage. The results were achieved by implementing the generated CpG text report files into MethylKit v1.18.0 for PCA clustering, correlation, and hierarchical clustering (Ward.D2 cluster method). Clustering of the three RRBS biological samples (**Fig. S3**), indicates the consistency of the biological sampling

Methylome segmentation

First, we filtered the CGmap files from WGBS to include only CpGs with a coverage of at least 10 reads within a sample. This data was then utilized to identify unmethylated and lowly methylated regions (UMR and LMR) using the MethylSeekR v 1.32.0 package in R. The following criteria were specified for segmentation of the methylome as described in [13]: (1) FDR < 5% for regions, (2) DNA methylation levels of 0 – 50 % for UMR and 10 -50 % for LMR, and (3) >5 CGs per region. An additional distinction between UMR and LMR are the density of CG sites within the region, with UMR having very dense number of CGs (>30 CGs) close together within a region and LMR having lower densities of CGs (<30 CGs) more distributed throughout a region. Genomic relevance of cis-regulatory elements was investigated by calculating the distance of the UMR and LMR to the TSS of genes and thereafter plotting the inverse log₂ of the distance to the TSS in R. We further identified FMRs using custom scripts, with the following requirements: (1) CG sites with methylation levels >0.75, (2) regions of 1000 bp with > 10 CG sites, (3) remove redundant regions repeated in UMR/LMR. The dynamics of the methylome were analysed using the 'intersect' function in Bedtools v2.30.0, ensuring that only regions with a minimum of 10% overlap were considered. FMR that remain unchanged throughout all stages of development were purposefully excluded, as the goal of this analysis was to highlight the dynamic nature of low and unmethylated regions. The resulting regions are plotted in R and the associated promoters of genes identified. This information is available in **Data S6**.

Differential methylation analysis

Both the RRBS and WGBS CGmap report text files generated from the pipeline were filtered to remove any CpG sites with less than 10 reads in a sample, and bases/regions having a higher coverage than the 99.5 percentile are also filtered. Thereafter we normalized read coverage distributions between samples (normalizeCoverage) using MethylKit (R) package v1.18.0 (Akalin *et al.*, 2012). Finally reads on both strands of a CpG nucleotide are merged, which provides better counts of the cytosines at a specific CpG site. Regions in WGBS were calculated using a sliding window approach with a window size of 500 bp and a step size of 250 bp. A pairwise comparison using an F-test together with basic correction for overdispersion (MN) [49,50] was implemented to calculate the methylation differences between development stages for each respective tissue. Pairs were defined as follows: 30 dpf to 70 dpf is the developmental transition, and 70 dpf to newborn is the maturation transition. Methylation differences are calculated on a per site level for RRBS data, whereas for WGBS data this is calculated on a region level. CpG sites/regions were considered as significantly differentially methylated if there was >25% difference in methylation rate for each transition (e.g. 30 dpf and 70 dpf) and having a corresponding q-value <0.01. Lastly, hyper- (increased methylation levels) and hypo-methylated (decreased methylation levels) sites/regions were identified per tissue for the two defined transitions. The promoter/enhancer regions overlapping with DMR/S were defined as the transcription start site (TSS) of the gene with an additional 1500 bp upstream and 200 bp downstream from the TSS, taking the strand into consideration for all genes.

Significantly differentially methylated sites and regions (DMS/DMR) were annotated with the Ensembl [Sus Scrofa 11.1 - release 102](#), where core promoter is defined as: TSS +500 bp upstream and 200 bp downstream (strand was considered for all genes). Intergenic regions are defined as a starting point 1000 bp upstream of TSS until next TSS. Hypo- and hyper-DMR are annotated using these custom regions files with bedtools intersect.

Identifying tissue specific promoters

Tissue specific promoters were defined in the following way. First, the Ensembl Sus Scrofa 11.1 - release 102 annotation was used to define promoters. Promoters from this annotation are calculated as follows: 1500 bp upstream of the TSS and 200 bp downstream of the TSS, we also accounted for strandedness in this calculation. This region is larger in comparison with the 'core promoter' region, as the aim is to identify potential enhancers close to promoters and in this way defining a larger region. We overlaid the promoter regions with both the DMR and DMS respectively, for each tissue at the developing transition. Thereafter all DMR/ DMS for a

4 DNA methylome in the developing pig

tissue were determined (thus all methylation changes occurring in both transitions). DMR/DMS which were specific to each tissue across all transitions were identified using bedtools intersect with the -v flag (e.g. bedtools intersect -v -a Tissue1_all_transitions -b all_other_tissues). Unique genes were extracted per tissue specific DMR/DMS and investigated for significant enrichment in GO terms using ShinyGO v0.77 [51].

Visualisation of methylation levels across genomic features

Relationship of methylation levels and various genomic regions were investigated using ViewBS 0.1.11 [52], where the methylation reports created from bismark were implemented together with custom region files. Distribution of methylation levels over promoter, gene-body and transcription termination sites were achieved by creating a region file using the Sus Scrofa 11.1 version 102 annotation file. Genes were classified based on their expression levels (TPM) as follows: (i) ≥ 0 to < 1 , (ii) ≥ 1 to < 10 , (iii) ≥ 10 to < 30 , (iv) ≥ 30 to > 75 , (v) ≥ 75 . The plot was generated using the MethOverregion functionality within ViewBS.

The repeat annotation was obtained from Warr *et al.*, (2020) [53]. Methylation patterns in different repeat classes were visualized using the MethOverregion function. The epigenomic states from the associated tissue were obtained from Pan *et al.*, (2021) [9], visualization was performed using the MethOverregion functionality.

Integrative analysis of expression and methylation

Gene expression data was curated within the project (available on the FAANG DCC) and TPM expression values per tissue, time-point in four replicates were available. DEG were also calculated over time, as for differential methylation, thus DEG were available per tissue for each transition in development. Promoter DMR and DMS as described above were overlaid with the corresponding DEG using bedtools intersect with the flag -wo to retain information regarding DEG and DMR/DMS. The DMEG were then categorized based on the methylation and expression changes: Up-regulated: $\log(\text{FC}) > 1.5$, down-regulated: $\log(\text{FC}) < -1.5$, Hyper-methylated: methylation $\text{FC} > 25$, and hypo-methylation: methylation $\text{FC} < 25$. All differentially methylated expressed genes with the specified categories, and all remaining were plotted in R. A GO (ShinyGO; [51]) analysis was performed for all genes per tissue at the developing transitions, terms were classified as significantly enriched for an $\text{FDR} < 0.05$. Top pathways were obtained, and all plotted together. Lastly the GO analysis was performed as stated above for the up-regulated and down regulated categories.

Supplementary information

Supplementary tables

Download: https://osf.io/kry5d/?view_only=c02a5c808a9749b8bfc26ce236f52ddf

(Supplementary_tables_pig.xls)

Supplementary data files

Download: https://osf.io/kry5d/?view_only=c02a5c808a9749b8bfc26ce236f52ddf

S1: Methylation over chromatin states per tissue time point. Notated as SS_tissue_timepoint_CG_MethOveRegion.pdf

S2: Methylation levels over repeat elements per tissue time point. Notated as SS_tissue_timepoint_CG_MethOveRegion.pdf

S3: Methylation levels per repeat region for all tissues. Notated as repeat_name_MethOverRegion_CG.pdf

S4: UMR region files. Notated as UMR_SS_tissue_timepoint.txt

S5: LMR region files. Notated as LMR_SS_tissue_timepoint.txt

S6: Tissue transition files and tissue transitions together with genes. Notated as tissue_transitions_f_bed.bed and genes_pos_tissue_transition.txt

S7: Differentially methylated sites per tissue per transition files, hypo - and hyper methylated sites files and genes associated with DMS files. Notated as Diff25p_OD_Ftest_tissue_time_points.bed, Diff25p_Hyper/Hypo_OD_Ftest_tissue_time_points.bed, genes_TSS_1500_tissue_time_points.txt

S8: Differentially methylated regions per tissue per transition files, hypo - and hyper methylated sites files and genes associated with DMS files. Notated as Diff25p_OD_Ftest_tissue_time_points.bed, Diff25p_Hyper/Hypo_OD_Ftest_tissue_time_points.bed, promoters_1500_Diff25p_OD_Ftest_tissue_time_points.txt

S9: Tissue specific DMS. Notated as Tissue_DMS_all_trans.bed

S10: Tissue specific DMR. Notated as Tissue_specific_dmr.bed

S11: DM sites associated with DEG and DMEG files. Notated as tissue_time_points_DMS_DEG.bed and as tissue_time_points_DM_DEG.bed

4 DNA methylome in the developing pig

S12: DMEG heatmaps

Additional information:

30 dpf pooling protocol:

https://data.faaang.org/api/fire_api/samples/INRA_SOP_GENESWITCH_D30_FETUS_POOLING_20200221.pdf

Sampling at 30 dpf protocol:

https://data.faaang.org/protocol/samples/INRA_SOP_GENESWITCH_D30_FETUS_SAMPLING_20200221.pdf

Sampling at 70 dpf protocol:

https://data.faaang.org/protocol/samples/INRA_SOP_GENESWITCH_D70_FETUS_SAMPLING_20200221.pdf

Sampling at newborn protocol:

https://data.faaang.org/protocol/samples/INRA_SOP_GENESWITCH_PIGLET_SAMPLING_20200221.pdf

DNA extraction protocol:

https://data.faaang.org/api/fire_api/assays/INRA_SOP_GENESWITCH_WP1_PIG_EXTRACTING_DNA_RNA_20201111.pdf

RRBS library preparation protocol:

https://data.faaang.org/protocol/samples/WU_SOP_GENESWITCH_WP1_RRBS_library_preparation_20201201.pdf

WGBS library preparation protocol:

https://data.faaang.org/api/fire_api/experiments/WU_SOP_GENESWITCH_WP1_WGBS_library_preparation_20210127.pdf

NuGEN Technologies' diversity trimming scripts:

<https://github.com/nugentechnologies/NuMetRRBS>

GSM-pipeline: <https://github.com/FAANG/GSM-pipeline>

Reference genome: https://ftp.ensembl.org/pub/release-104/fasta/sus_scofra/dna/

Annotation: https://ftp.ensembl.org/pub/release-102/gtf/sus_scofra/

References

1. Amin V, Onuchic V, Milosavljevic A. Chapter 7 - Mining the Epigenetic Landscape: Surface Mining or Deep Underground. In: Gray SG, editor. *Epigenetic Cancer Therapy* [Internet]. Boston: Academic Press; 2015 [cited 2023 Aug 7]. p. 141–65. Available from: <https://www.sciencedirect.com/science/article/pii/B9780128002063000070>
2. Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet.* 2014;15:453–68.
3. Moore LD, Le T, Fan G. DNA methylation and its basic function. *Neuropsychopharmacology.* 2013;38:23–38.
4. Bartolomei MS, Oakey RJ, Wutz A. Genomic imprinting: An epigenetic regulatory system. *PLOS Genetics.* 2020;16:e1008970.
5. Nasrullah, Hussain A, Ahmed S, Rasool M, Shah AJ. DNA methylation across the tree of life, from micro to macro-organism. *Bioengineered.* 2022;13:1666–85.
6. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol.* 2019;20:590–607.
7. Doherty R, Couldrey C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. *Frontiers in Genetics* [Internet]. 2014 [cited 2023 May 1];5. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2014.00126>
8. Zeng Y, Chen T. DNA Methylation Reprogramming during Mammalian Development. *Genes (Basel).* 2019;10:257.
9. Pan Z, Yao Y, Yin H, Cai Z, Wang Y, Bai L, et al. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat Commun.* 2021;12:5848.
10. Manghera M, Douville RN. Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? *Retrovirology.* 2013;10:16.
11. Li W, Liu S, Zhao J, Deng R, Liu Y, Li H, et al. The highly expressed ERV1 forms virus-like particles for regulating early embryonic development [Internet]. *bioRxiv*; 2022 [cited 2023 Aug 22]. p. 2022.06.21.496818. Available from: <https://www.biorxiv.org/content/10.1101/2022.06.21.496818v1>
12. Chen C, Wang W, Wang X, Shen D, Wang S, Wang Y, et al. Retrotransposons evolution and impact on lncRNA and protein coding genes in pigs. *Mob DNA.* 2019;10:19.
13. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* 2013;41:e155.
14. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev.* 2002;16:6–21.
15. Meissner A. Epigenetic modifications in pluripotent and differentiated cells. *Nat Biotechnol.* 2010;28:1079–88.
16. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet.* 2013;45:1198–206.
17. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* 2009;19:1044–56.

4 DNA methylome in the developing pig

18. Orozco LD, Morselli M, Rubbi L, Guo W, Go J, Shi H, et al. Epigenome-Wide Association of Liver Methylation Patterns and Complex Metabolic Traits in Mice. *Cell Metabolism*. 2015;21:905–17.
19. He Y, Hariharan M, Gorkin DU, Dickel DE, Luo C, Castanon R, et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature*. 2020;583:752–9.
20. Liu H, Zhou J, Tian W, Luo C, Bartlett A, Aldridge A, et al. DNA methylation atlas of the mouse brain at single-cell resolution. *Nature*. 2021;598:120–8.
21. Yang Y, Fan X, Yan J, Chen M, Zhu M, Tang Y, et al. A comprehensive epigenome atlas reveals DNA methylation regulating skeletal muscle development. *Nucleic Acids Research*. 2021;49:1313–29.
22. Wells JM, Melton DA. Early mouse endoderm is patterned by soluble factors from adjacent germ layers. *Development*. 2000;127:1563–72.
23. Lee HJ, Lowdon RF, Maricque B, Zhang B, Stevens M, Li D, et al. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nat Commun*. 2015;6:6315.
24. Huse SM, Gruppuso PA, Boekelheide K, Sanders JA. Patterns of gene expression and DNA methylation in human fetal and adult liver. *BMC Genomics*. 2015;16:981.
25. Šinkora M, Butler JE. The ontogeny of the porcine immune system. *Dev Comp Immunol*. 2009;33:273–83.
26. Lewis K, Yoshimoto M, Takebe T. Fetal liver hematopoiesis: from development to delivery. *Stem Cell Research & Therapy*. 2021;12:139.
27. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics & Chromatin*. 2018;11:37.
28. Csapo R, Gumpenberger M, Wessner B. Skeletal Muscle Extracellular Matrix – What Do We Know About Its Composition, Regulation, and Physiological Roles? A Narrative Review. *Front Physiol*. 2020;11:253.
29. Mukund K, Subramaniam S. Skeletal muscle: A review of molecular structure and function, in health and disease. *Wiley Interdiscip Rev Syst Biol Med*. 2020;12:e1462.
30. Wang Z, Grange M, Wagner T, Kho AL, Gautel M, Raunser S. The molecular basis for sarcomere organization in vertebrate skeletal muscle. *Cell*. 2021;184:2135–2150.e13.
31. Huang Y-Z, Sun J-J, Zhang L-Z, Li C-J, Womack JE, Li Z-J, et al. Genome-wide DNA Methylation Profiles and Their Relationships with mRNA and the microRNA Transcriptome in Bovine Muscle Tissue (*Bos taurine*). *Sci Rep*. 2014;4:6546.
32. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*. 2015;523:212–6.
33. Sliker RC, Roost MS, Iperen L van, Suchiman HED, Tobi EW, Carlotti F, et al. DNA Methylation Landscapes of Human Fetal Development. *PLOS Genetics*. 2015;11:e1005583.
34. Fan Y, Liang Y, Deng K, Zhang Z, Zhang G, Zhang Y, et al. Analysis of DNA methylation profiles during sheep skeletal muscle development using whole-genome bisulfite sequencing. *BMC Genomics*. 2020;21:327.
35. Ferretti E, Hadjantonakis A-K. Mesoderm specification and diversification: from single cells to emergent tissues. *Current Opinion in Cell Biology*. 2019;61:110–6.

4 DNA methylome in the developing pig

36. Rooke JA, Bland IM. The acquisition of passive immunity in the new-born piglet. *Livestock Production Science*. 2002;78:13–23.
37. Duarte ME, Kim SW. Intestinal microbiota and its interaction to intestinal health in nursery pigs. *Animal Nutrition*. 2022;8:169–84.
38. Dominique J-E, Gérard C. Myostatin regulation of muscle development: Molecular basis, natural mutations, physiopathological aspects. *Experimental Cell Research*. 2006;312:2401–14.
39. Chen M-M, Zhao Y-P, Zhao Y, Deng S-L, Yu K. Regulation of Myostatin on the Growth and Development of Skeletal Muscle. *Frontiers in Cell and Developmental Biology* [Internet]. 2021 [cited 2023 Aug 22];9. Available from: <https://www.frontiersin.org/articles/10.3389/fcell.2021.785712>
40. Aiello D, Patel K, Lasagna E. The myostatin gene: an overview of mechanisms of action and its relevance to livestock animals. *Animal Genetics*. 2018;49:505–19.
41. Prather RS. Pig genomics for biomedicine. *Nat Biotechnol*. 2013;31:122–3.
42. Walters EM, Prather RS. Advancing Swine Models for Human Health and Diseases. *Mo Med*. 2013;110:212–5.
43. Gutierrez K, Dicks N, Glanzner W, Agellon L, Bordignon V. Efficacy of the porcine species in biomedical research. *Frontiers in Genetics* [Internet]. 2015 [cited 2023 Apr 18];6. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2015.00293>
44. Bidarimath M, Tayade C. Pregnancy and spontaneous fetal loss: A pig perspective. *Molecular Reproduction and Development*. 2017;84:856–69.
45. Lunney JK, Van Goor A, Walker KE, Hailstock T, Franklin J, Dai C. Importance of the pig as a human biomedical model. *Science Translational Medicine*. 2021;13:eabd5758.
46. Ferenc K, Pietrzak P, Godlewski MM, Piwowarski J, Kiliańczyk R, Guilloteau P, et al. Intrauterine growth retarded piglet as a model for humans--studies on the perinatal development of the gut structure and function. *Reprod Biol*. 2014;14:51–60.
47. De Vos J, Derks MFL, Kurylo C, Groenen MAM, Madsen O. GSM-pipeline: GENE-SWitCH pipeline for comprehensive bisulfite sequencing analysis [Internet]. In Review; 2023 Jul. Available from: <https://www.researchsquare.com/article/rs-2984574/v1>
48. Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. [Internet]. Babraham Institute; 2015. Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
49. P McCullagh, Nelder JA. Binary data. *Generalized linear models*. US: Springer; 1989. p. 98–148.
50. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13:R87.
51. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*. 2020;36:2628–9.
52. Huang X, Zhang S, Li K, Thimmapuram J, Xie S, Wren J. ViewBS: a powerful toolkit for visualization of high-throughput bisulfite sequencing data. *Bioinformatics*. 2018;34:708–9.
53. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience*. 2020;9:giaa051.

4 DNA methylome in the developing pig

Supplementary figures

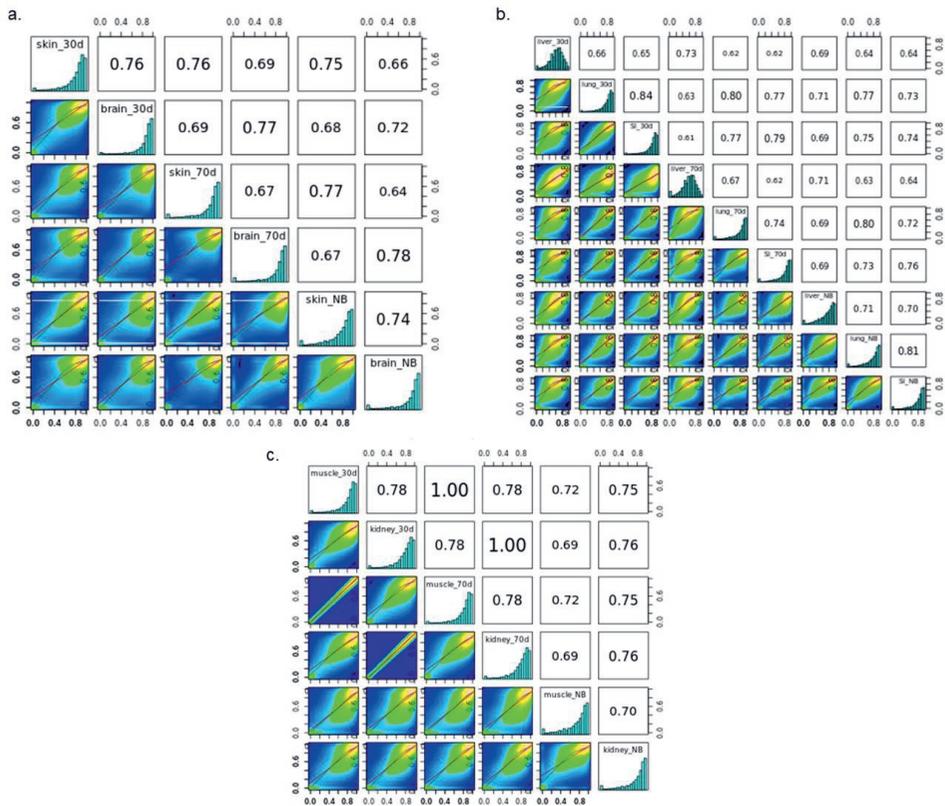


Figure S1 Correlations between CpG methylation patterns from WGBS data. Correlations are grouped based on germ layer: **a.** Ectoderm, **b.** Endoderm, **c.** Mesoderm.

4 DNA methylome in the developing pig

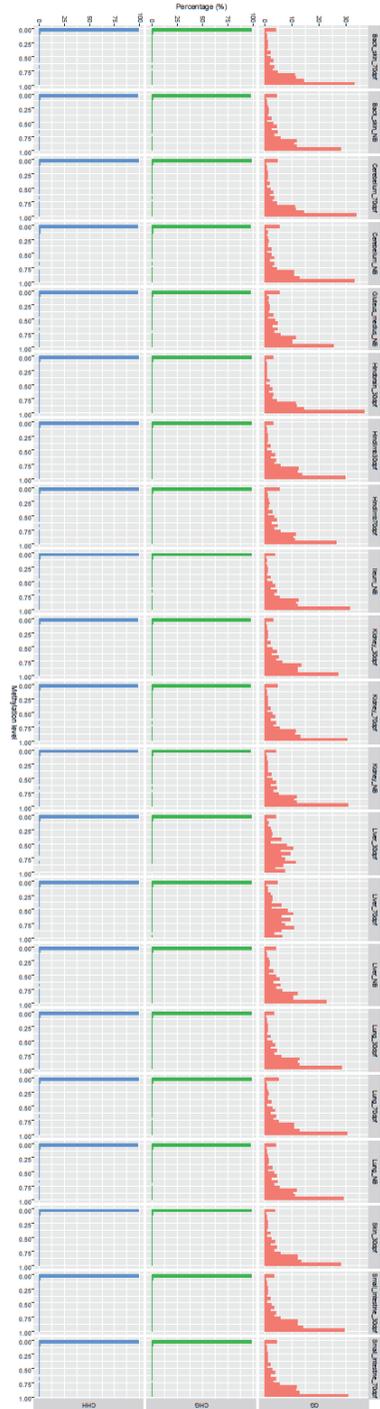


Figure S2 Distribution of methylation across tissues per developmental stages based on WGBS data. Percentage of sites is shown on the y-axis on the left, with the type of methylation shown on the y-axis on the right. Methylation levels is shown on the x-axis with the tissue-time point displayed at the top of the figure.

4 DNA methylome in the developing pig

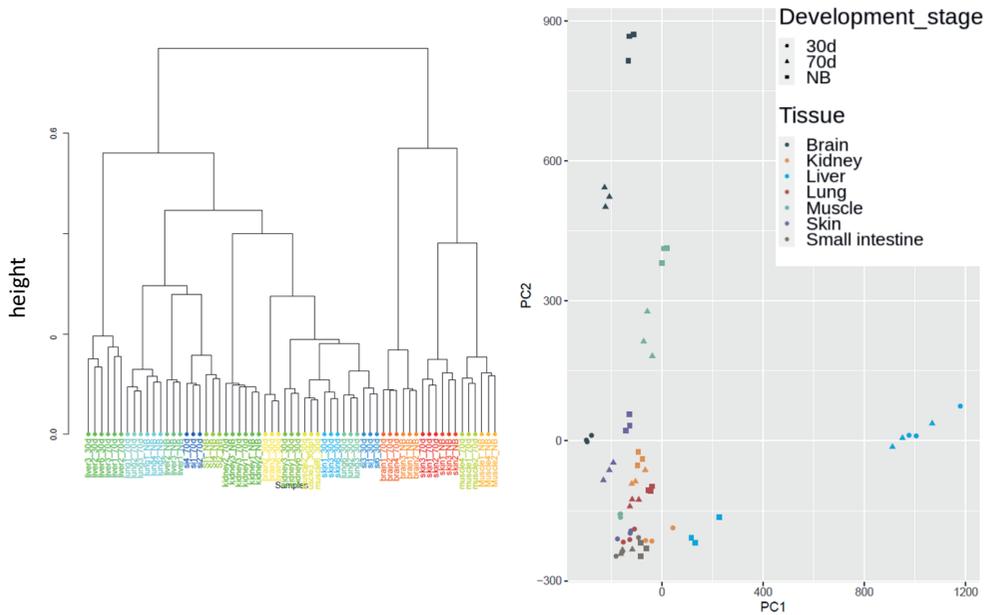


Figure S3 Clustering of all tissue samples at 3 developmental stages. Hierarchical cluster using Ward.D2 for RRBS data, with distances being measured from percent methylation per base (a). PCA clustering showing PC1 and PC2 with highest explained variance from RRBS data (b).

4 DNA methylome in the developing pig

Meth_state ■ FMR ■ LMR ■ UMR

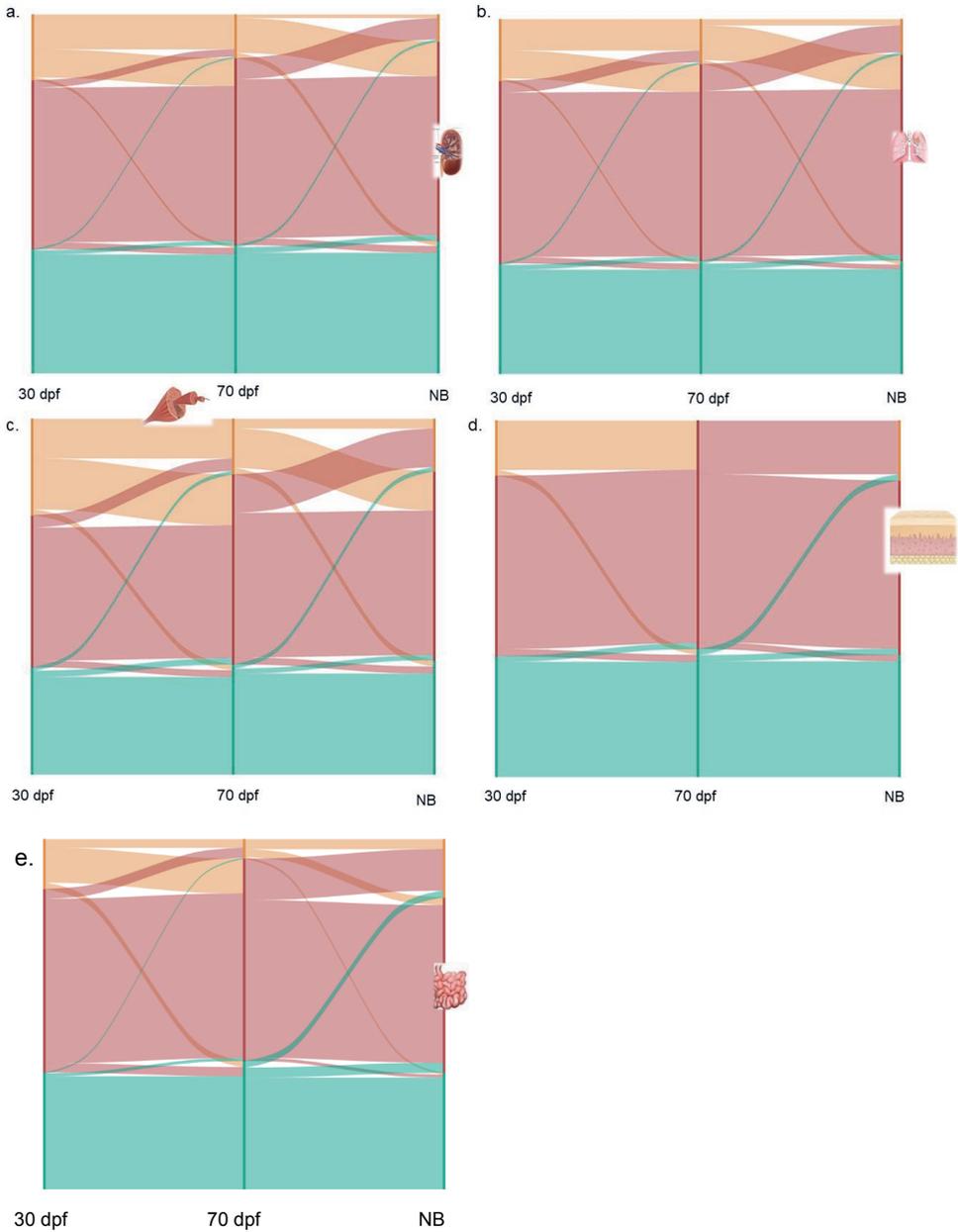


Figure S4 Dynamic changes for various tissues, with different colours depicting the methylation states (UMR, LMR and FMR). **a.** Kidney, **b.** Lung, **c.** Muscle, **d.** Skin, **e.** Small intestine.

4 DNA methylome in the developing pig 

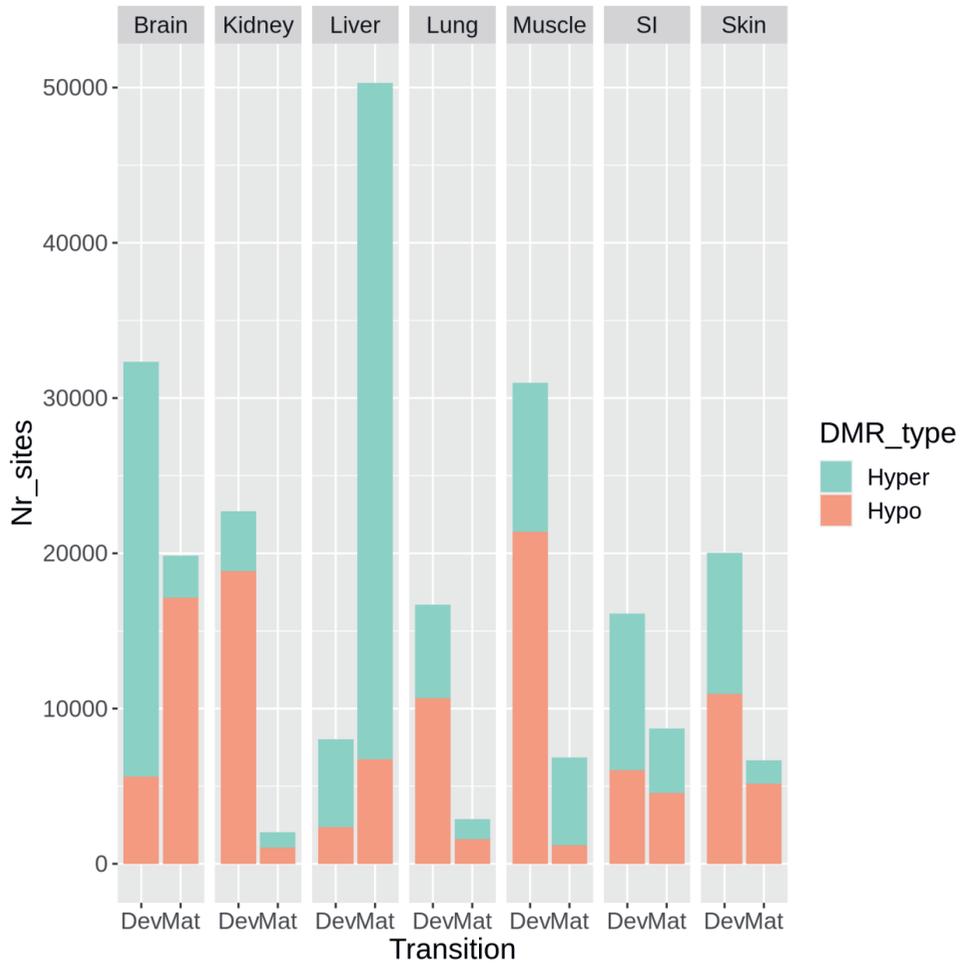
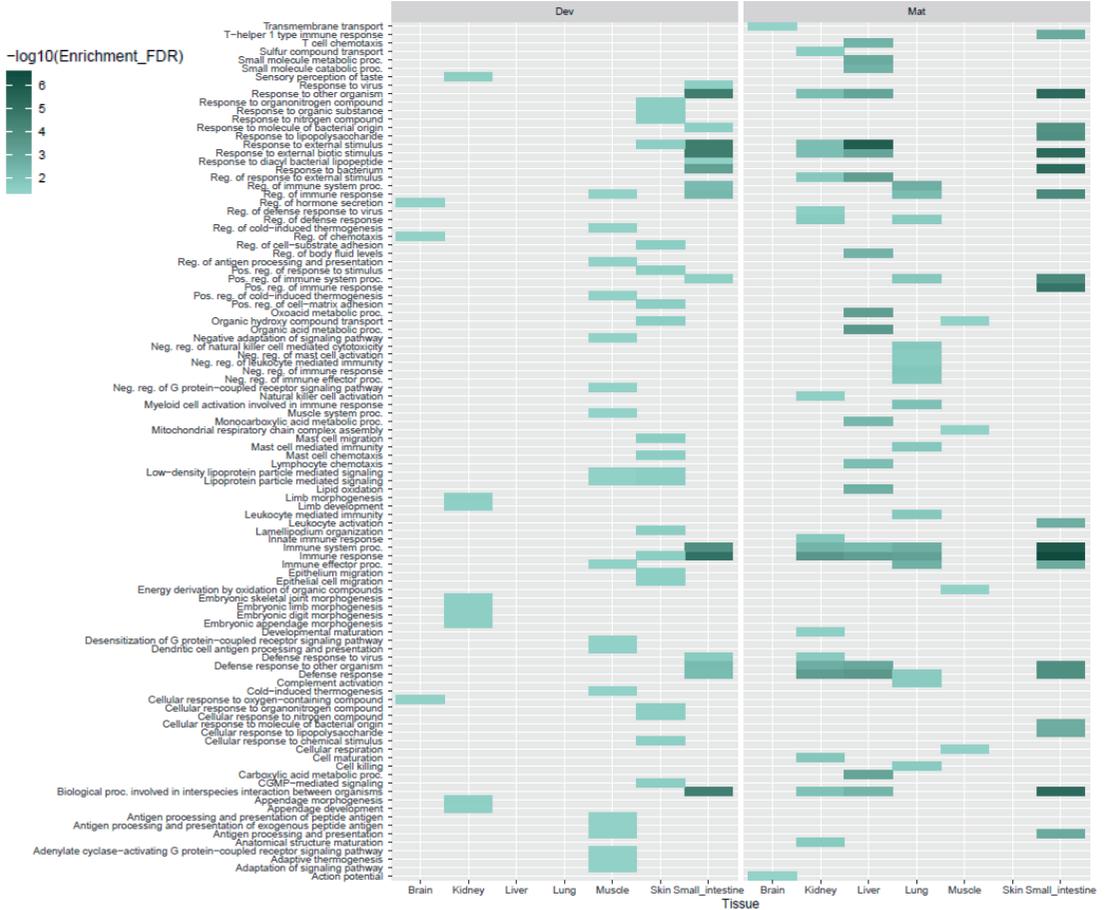


Fig. S5 Distribution of DMS identified using RRBS data, showing the number (y-axis) of hypo and hyper methylated regions per tissue (on top x-axis), for each fetal transition (bottom x-axis). Dev is defined as developmental transition and Mat defined as maturation transition. Hyper and hypo methylation are shown in orange and green as stacked bars.

4 DNA methylome in the developing pig

a

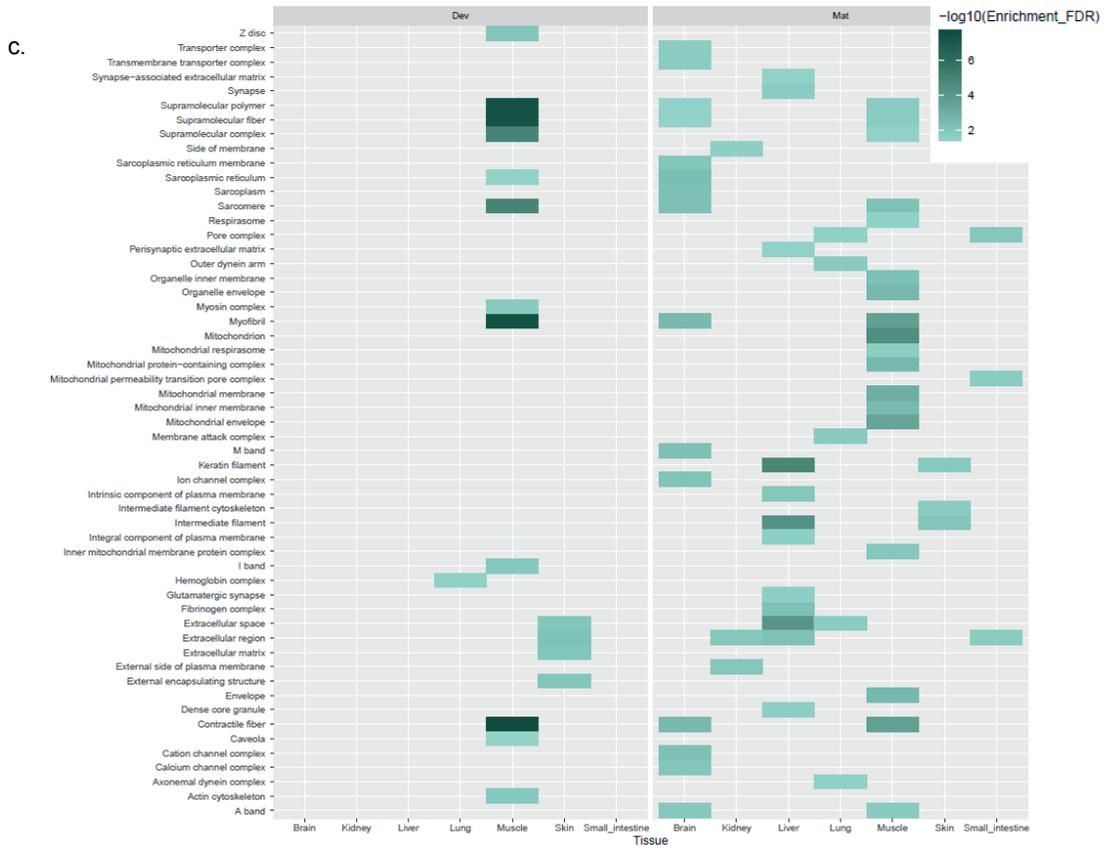


4 DNA methylome in the developing pig

b.



4 DNA methylome in the developing pig



4 DNA methylome in the developing pig

d.

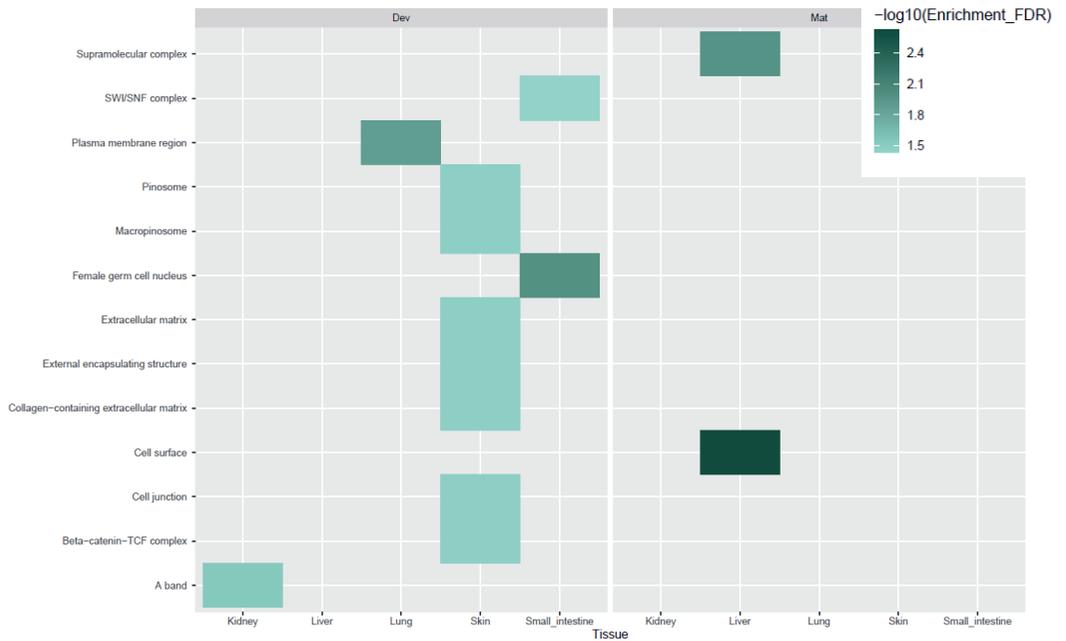


Figure S6 Enrichment analysis of DMEGs in the developmental and maturation transition of up- and down-regulated gene groups. With **a.** Biological processes (BP) gene ontology (GO) terms for up-regulated genes which are both hypo and hyper methylated. **b.** Biological processes GO terms for down-regulated genes which are both hyper and hypo methylated. Cellular component (CC) GO terms for **c.** up-regulated genes and **d.** down regulated genes. All terms are filtered for an FDR<0.05. Genes are filtered as follows for classifications: up-regulated $\log(FC) > 1.5$, and down-regulated $\log(FC) < -1.5$. All methylated regions have a minimum difference of 25%.

Chapter 5

Spatio-temporal methylome dynamics during development in chicken

J de Vos¹, M F.L. Derks¹, H. Acloque², S. Djebali³, S. Foissac⁴, C. Guyomar⁴, C. Kurylo⁴, E. Giuffra², M A.M. Groenen¹, O Madsen¹

¹Animal Breeding and Genomics, Wageningen University & Research, The Netherlands; ²Paris-Saclay University, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France.; ³IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, 31024 Toulouse, France; France; ⁴GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, 31326, France

(In preparation)



'An understanding of the natural world, and what's in it is a source of not only great curiosity but great fulfilment'

Sir David Attenborough



Abstract

Embryonic development is a dynamic process marked by germ layer differentiation into organs and tissues through organogenesis, which is regulated by intricate genomic and epigenomic interactions. In this study we used both reduced representation – and whole genome bisulfite sequencing data to unravel the role of DNA methylation in chicken development. We profiled the DNA methylomes from seven tissues (brain, kidney, liver, lung, muscle, skin, and small intestine) at three developmental stages, from embryo at 8 days, embryo at 15 days and at hatching. Methylome dynamics was investigated by 1. segmenting the whole methylome into classes based on the methylation levels and density of CpG sites, and 2. differential methylation analysis on both a per-site and region level. The number of differentially methylated sites across developmental stages within the various tissues ranged from 11,240 to 50,114 at the developmental transition (embryo 8 days to embryo 15 days) and from 4,276 to 25,661 at the maturation transition (embryo 15 days to hatch). Differentially methylated regions ranged from 29,848 to 263,44 at the developmental transition and from 50,993 to 201,024 at the maturation transition. Using methylome segregation, we visualize the dynamic methylome, revealing that most tissues exhibit similar patterns of methylation state changes. However, the brain displays distinct patterns in comparison to other tissues. We further combined the epigenome (DNA methylation) with the transcriptome of the same seven tissues and developmental stages. Differentially methylated expressed genes were investigated using functional enrichment analysis. In tissues such as brain and muscle, terms related to general growth, such as embryonic morphogenesis and animal organ development, are significantly enriched in down-regulated genes during the developmental transition. An interesting example is the up regulated genes in skin during the developmental transition which show enrichment for terms such as collagen fibril organization, and extracellular matrix organization. This provides a first look into the dynamics of the DNA methylome during development in chicken.

Background

Vertebrate development is a complex process that involves the coordinated activation and repression of numerous genes and signalling pathways [1]. Gastrulation is key during development in amniotes where the different germ layers are formed, namely the endoderm, mesoderm, and ectoderm [2]. These germ layers give rise to all the tissues and organs of the body through a series of cellular and molecular events, including morphogenesis, cell differentiation, and cell migration. The precise timing and regulation of these processes is critical for proper embryonic development, and even minor disruptions can lead to a wide range of congenital defects and developmental disorders. Development is orchestrated by gene expression, where alterations in gene expression levels at specific time points play a pivotal role in regulating tissue and cell differentiation [1,3,4]. Understanding the underlying molecular mechanisms of vertebrate development is therefore critical for a fundamental understanding of the evolution of development and for improving health and animal breeding.

The functional regulatory genome refers to the regions of the genome responsible for regulating gene expression. The functional regulatory genome consists of various elements such as promoters, enhancers, silencers, insulators, and other regulatory elements that interact with each other to regulate gene expression [5–7]. Promoters are regions of DNA located adjacent to the transcriptional start site of a gene and are necessary for transcription initiation [5,6]. Enhancers and silencers are regions of DNA that can be located either upstream, downstream or inside of a gene and can increase or decrease gene expression, respectively [8,9]. Epigenomic modifications are chemical modifications of DNA and its associated proteins that affect gene expression without altering the underlying DNA sequence [10,11]. There are several types of epigenomic modification, including DNA methylation, histone modifications, and long non-coding RNA. DNA methylation involves the addition of a methyl group to a cytosine nucleotide in DNA, in vertebrates typically at a CpG dinucleotide, which can cause gene silencing [4,12–14]. These epigenomic modifications are important for regulating normal cellular processes, such as tissue differentiation and development.

GENE-SWitCH is a project with the aim of investigating the functional regulatory genomes throughout embryonic development in the chicken and pig. Previous research investigating the dynamic nature of epigenomic modifications regulating the functional regulatory genome during fetal development have been conducted in mouse, and zebrafish [15–17]. However, similar studies in pig and chicken were limited in the number of tissues examined (e.g. only skeletal muscle) [18,19]. Our study contains data from seven key tissues in the chicken, representative of three embryological layers, and at three developmental stages namely

embryo at 8 days (early organogenesis), embryo at 15 days (late organogenesis) and at hatching. This is the first investigation into DNA methylation dynamics regulating gene expression during development in a wide variety of tissues in an avian species. Furthermore, we investigate the dynamics of development between a bird and a mammal by a comparative analysis between the chicken and pig.

Results

We investigated embryonic development and organogenesis by analysing seven different tissues from the three embryological germ layers, at different stages of development, embryo at day 8 (E8), embryo at day 15 (E15) and at hatching (HC). We generated 63 reduced representation bisulfite (RRBS) datasets (3 samples * 3 timepoints * 7 tissues) and 21 whole genome bisulfite (WGBS) datasets (1 sample * 3 timepoints * 7 tissues). By investigating the DNA methylome at these time-points we gain insight into the spatio-temporal dynamics of embryonic development.

Methylome during development

Based on WGBS data, global methylation levels ranged between 2-3%, CpG methylation between 59-71% and non-CpG levels between 0.2-0.4% (**Table S1**). Distribution of methylation levels (WGBS) across all tissues and time-points, is shown in **Fig. S2**. We observed a general pattern of methylation levels decreasing during development from E8 to E15 in all tissues, with brain showing the highest CpG methylation level at E8 (**Fig. 1a**). On a genome-wide level, brain showed a higher methylation level at E8, which decreased at E15, and then increased again at hatching (**Fig. S8** and **Data S12**). Interestingly, in kidney, the methylation levels decreased from E15 to hatching, whereas in the six remaining tissues a slight increase in methylation levels from E15 to hatching is seen.

A typical drop in methylation level (WGBS) at the promoter region is seen for all tissues at all stages of development (**Fig. 1b**). Lung at E15 had the lowest level of methylation at the promoter (~12%) while the highest level is observed for brain at E8 (~32%). Methylation levels within the gene body fluctuate between 50% and 70% for the different tissues at the three developmental stages. The level of methylation was also higher in the gene body compared to levels in intergenic regions (0-2000 bp downstream of the gene).

5 DNA methylome dynamics in chicken development

From the hierarchical clustering of WGBS samples (**Fig. 1c**), we observed two main clusters. The first main cluster divided further into two subclusters, one for kidney, liver, and skin at hatching, while all three brain tissues and kidney E8 are in another subcluster.

The second main cluster is divided into four subclusters. One subcluster contains lung, small intestine and muscle at hatching, with the latter forming a distinct branch within this subcluster. Interestingly, we also observed a subcluster of tissues derived from the endoderm germ layer, where liver at E15 and E8 group together and lung at E15 and E8 group together on separate branches within this subcluster. The remaining two subclusters show a separation of developmental stages E8 and E15, respectively. In the subcluster of E8, muscle and skin group on one branch, and small intestine on another. In the E15 subcluster, we found, as for E8, that muscle and skin group together and that small intestine now groups together with kidney on a separate branch with this E15 subcluster.

The principal component analysis (PCA) of WGBS data (**Fig. 1d**) generally showed a similar cluster pattern as the hierarchical clustering. Liver samples at E8 exhibited slight separation from those at E15 and hatch, forming a distinctive cluster. Kidney at E8 and E15 clusters with small intestine of the corresponding developmental stages.

Hierarchical clustering (**Fig. S1a**) and PCA clustering (**Fig. S1b**) for reduced representation bisulfite sequencing (RRBS) are presented in **Fig. S1**. Interestingly from these results the PCA showed a different pattern compared to the hierarchical clustering (**Fig. S1a**). In the PCA, brain samples cluster together at E8 and hatch, while E15 samples form a separate cluster along PC1 (**Fig. S1b**). In contrast, for the hierarchical clustering, brain and liver cluster on two distinct branches for liver and brain, respectively at all developmental stages.

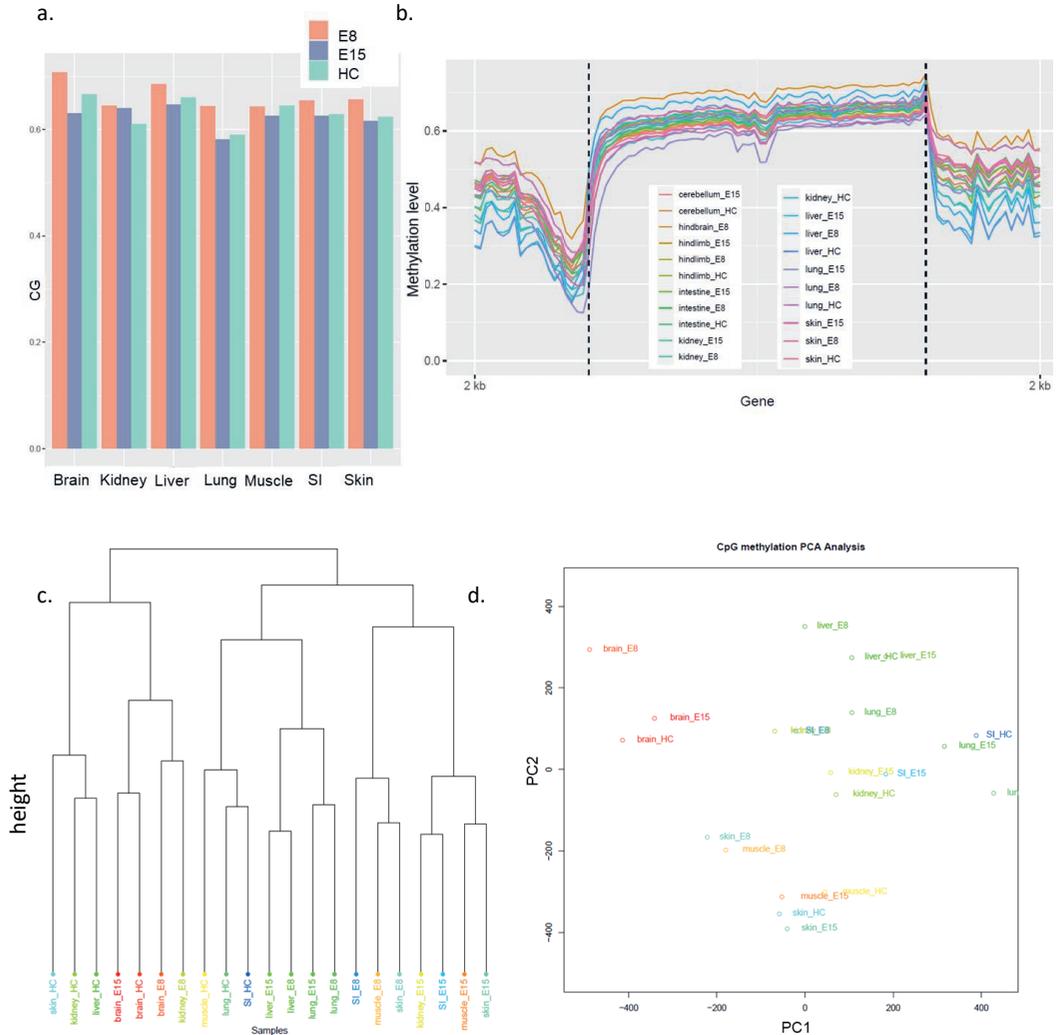


Fig. 1 Methylation patterns and distribution from WGBS data of the tissues at different developmental stages. **a.** CpG methylation across seven tissues per developmental stage. Developmental stages are shown as different colours, with orange indicating E8, purple E15 and green hatching (HC). **b.** Methylation levels across genes for each tissue per developmental stage from 2000 bp upstream of the transcription start site (TSS), in gene body and 2000 bp downstream of the gene. **c.** Hierarchical clustering of all tissue samples at three developmental stages in chicken. **d.** Principal component analysis (PCA) clustering showing PC1 and PC2. SI is small intestine.

Methylation levels across genomic regions during development

Methylation levels across genes in different classes of gene expression

To investigate the relationship between DNA methylation and gene expression, we examined transcriptome data obtained from the same tissues and timepoints, and we show an example of liver at hatching (**Fig. 2**). Our analysis revealed a negative correlation between the methylation levels at the transcription start site (TSS) and the corresponding gene expression levels. Genes with higher expression (TPM > 10) showed a decline in methylation levels to below 20% around the TSS. Conversely, unexpressed, or minimally expressed genes (0-1 TPM) exhibited significantly higher methylation levels (~50%) around the TSS. All categories of expressed genes demonstrated elevated levels of methylation within the gene-body.

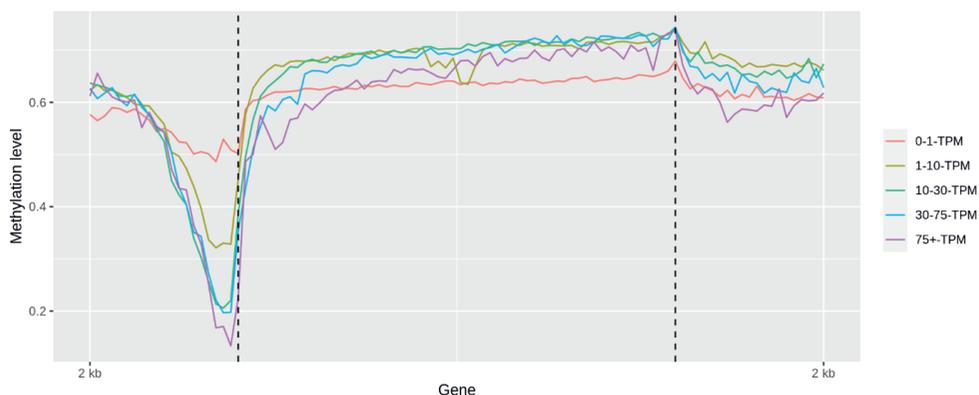


Fig. 2 Distribution of methylation levels across different regions relative to the gene-body in liver at hatching. The y-axis represents methylation levels, ranging from 0 (unmethylated) to 1 (fully methylated), while gene expression categories, spanning from unexpressed to lowly expressed and highly expressed genes, are depicted in distinct colours. The TSS is indicated as the first horizontal dotted line from the left, followed by the gene-body and regions downstream of the gene.

Dynamics of DNA methylation in regulatory and repeat elements

We further examined the methylation profile of repetitive elements and the methylation pattern associated with chromatin states as defined by Pan et al. [36]. Chromatin states show distinct patterns of methylation as shown for lung at E15 in **Fig. 3a**. The chromatin state with the lowest

methylation levels was E1, corresponding to actively transcribed TSS. Subsequent lowest methylation levels are observed in E3, E4, and E12, representing regions transcribed at genes, weakly transcribed at genes, and poised TSS, respectively. All four states showed a clear drop in methylation level compared to up- and downstream regions. Following these states, we observed E6 and E7, corresponding to strong active enhancers and medium enhancers with ATAC (assay for transposase-accessible chromatin) have a minor decrease in methylation compared to flanking regions. The E5 state, corresponding to transcribed regions without ATAC, and E2, representing flanking active TSS without ATAC showed a clear increase in methylation compared to up- and downstream regions. The four states with highest methylation levels are three enhancer states (E8, E9, E10) and ATAC islands (E11). Notably, the state with the highest methylation levels was E8, representing weak active enhancers. E12, corresponding to poised TSS, exhibited more varied methylation levels, likely due to the limited number of annotated regions in the genome for this state. Thus, the results showed that active and medium active enhancers are strongly methylated, while only strongly active enhancers showed a slight decline in methylation levels. This intriguing trend sheds light on the dynamic interplay between chromatin states and DNA methylation in gene regulation. (**Fig. 3a**). Methylation levels for chromatin states of other tissues are shown in **Data S1**.

LINE repeats are generally highly methylated compared to up-and downstream regions while other repeat families including LTR ERV1 elements showed a slight decline in methylation levels (**Fig. 3b, c and Data S2**). Interestingly, (hind)brain at E8 exhibited markedly higher methylation levels at repeat elements (**Fig. 3b**), while lung at hatching showed significantly lower methylation levels (**Fig. 3b**), most prominently at LINE CR1 repeats. Methylation levels across other DNA repeats are shown in **Data S2**.

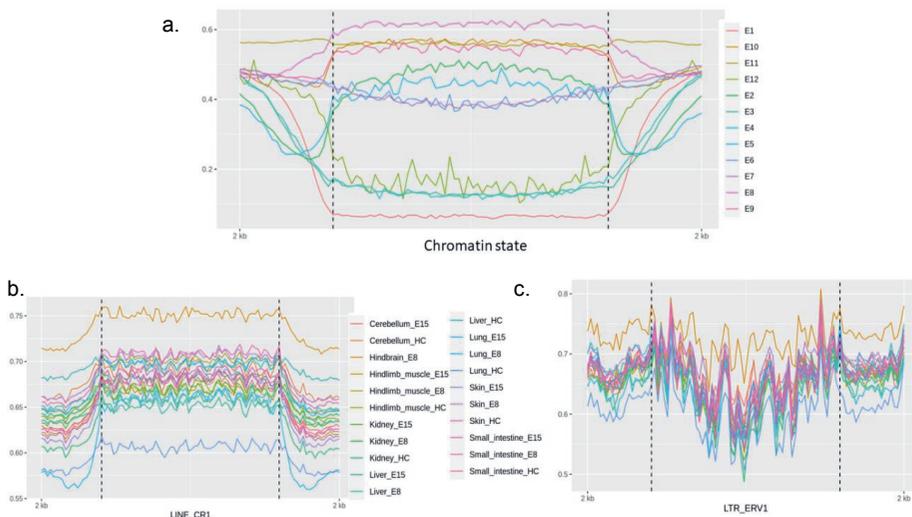


Fig. 3 a. Distribution of methylation levels within and surrounding different chromatin states in lung E15. Methylation levels are shown on the y-axis, and chromatin states are denoted as follows: E1: Actively transcribed TSS, E2: Flanking active TSS without ATAC, E3: Transcribed gene, E4: Weakly transcribed gene, E5: Transcribed region without ATAC, E6: Strong active enhancer, E7: Medium enhancer with ATAC, E8: Weak active enhancer, E9: Active enhancer without ATAC, E10: Poised enhancer, E11: ATAC Island, E12: Poised TSS. **b.** Methylation levels at LINE_CR1 and **c.** LTR_ERV1 repeats across samples. In all plots methylation levels ranging from 0 (unmethylated) to 1 (fully methylated) are shown on the y-axis

Dynamics of DNA methylation during developmental and maturation phases in 7 tissues

Methylome segregation

We identified unmethylated and lowly-methylated regions (UMR and LMR, respectively), which can be indicative of cis-regulatory elements, such as promoters and enhancers (see methods for details and criteria used to call UMR and LMR). The total number of UMR identified ranged from 3,285 to 13,780 (**Table S2**), with the lowest numbers of UMR found in brain at E8 and hatch, and the most found in lung at hatch. More LMR than UMR were identified, and the number ranged from 8,699 to 259,020 (**Table S2**), with the highest numbers of regions identified for small intestine at hatch and E15, and the lowest number of LMR identified in brain at hatch. Average lengths and number of regions overlapping genes are indicated in **Table S2**. Mean length of a UMR across all tissues was 3,216 bp (ranging from 2,371 to 4,717 bp) and 640 bp for LMR (ranging from 301 to 1,120 bp). Additionally, on average 5,568 (ranging from 2,425 to 9,462) and 2,860 (ranging from 368 to 6,557) genes are associated with UMR and LMR, respectively. Brain E8 and at hatching have the least number of genes associated with UMR and LMR. Lung E15 has the highest number of genes associated with UMR, while small intestine E15 has the highest number of genes associated with LMR. Regions and associated genes for UMR and LMR are shown in **Table S2** and **Data S3**, and **S4**, respectively.

Distance of UMR and LMR to the TSS was calculated and plotted for brain tissue at hatching (**Fig. S3**). We observed a relative similar proximity of UMR and LMR to TSSs, with the average UMR distance to TSSs being slightly shorter compared to LMR. The large similarity in the distribution of the distance of UMR and LMR to TSSs suggest an abundance of potential overlap of cis-regulatory elements (**Fig. S3**).

The dynamic nature of the DNA methylome, both at a temporal and spatial level, was revealed by observed transitional changes between UMR, LMR and fully methylated regions (FMR) during development, i.e. changes between E8 and E15 and between E15 and hatching. (**Fig. 4, Fig. S4**).

Fig. 4 depicts the proportions of transitional changes of UMR, LMR, and FMR in lung (a) and muscle (b). A substantial fraction of UMR remains unchanged from E8 to E15. A small portion of the UMR at E8 transitioned to LMR at E15, and only a small fraction shifted to FMR at E15. Similarly, LMR at E8 predominantly remained LMR at E15, with a smaller proportion transitioning to UMR, and an even smaller fraction transitioning to FMR at E15.

Notably, more than half of the FMR at E8 transitioned to LMR at E15, and most of the remaining staying unchanged, while a small fraction transitioning to UMR. Transitioning from E15 to hatching, a noteworthy trend emerged, where most of the FMR changed to LMR, with a small fraction changing to UMR. Moreover, the majority of LMR at E15 remained LMR at hatching, with the remaining fractions transitioning to FMR and UMR, respectively.

Interestingly, the largest proportion of UMR at E15 remained UMR at hatching, with a smaller fraction shifting to LMR, and an even smaller portion changing to FMR. These patterns are consistent for both lung and muscle tissues, with the main difference being that a smaller proportion of UMR changed to LMR at E15 to hatching in muscle compared to lung. Additionally, a larger proportion of LMR transitioned to FMR at E15 to hatching in muscle compared to lung, and a higher number of UMR transitioned to LMR from E8 to E15 in lung compared to muscle.

Transitions between methylation states of other tissues are shown in **Fig. S4**. Overall, the results revealed similar general patterns of methylation state transitions during development across different tissues. Some intriguing distinctions are observed in brain (**Fig. S4a**), where a substantial fraction of FMR transition to LMR from E8 to E15, and a prominent shift of FMR to LMR is seen from E15 to hatching. This suggests tissue-specific (distinct) epigenetic regulations in brain during chicken development. The number of genes found associated with transitioning regions are shown in **Table S3**.

Notable genes linked to tissue development include *MYLK4*, *PDIA5* and *HDAC2* in muscle transitioning from FMR at E8 and E15 to LMR at hatching. Similarly, the genes *HOXB1*, *GDF6*, and *MYB* transitioned from FMR at E8 and E15 to LMR at hatching in the small intestine. Results for transitions of all tissues during development are available in **Data S5**.

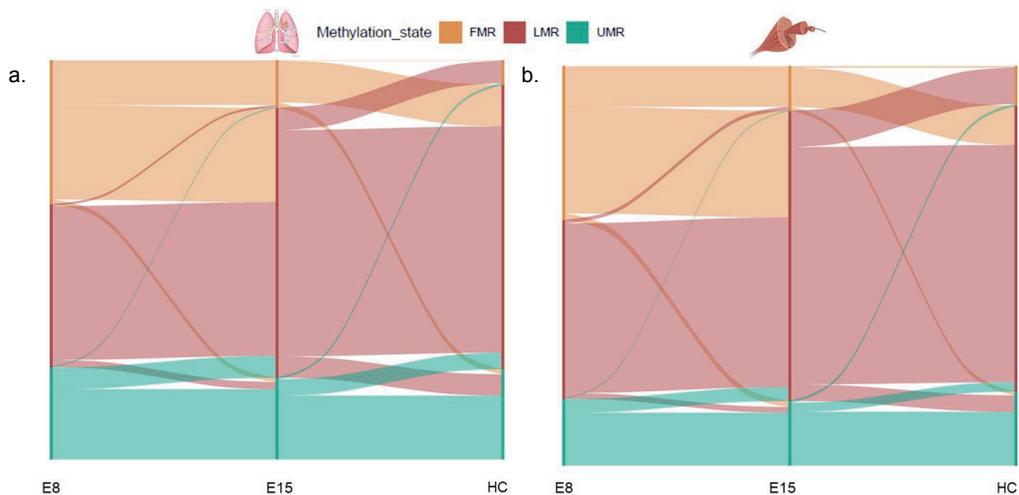


Fig. 4 Methylation transitions occurring in lung **(a)** and muscle **(b)**. Orange indicates FMR, red indicates LMR, and green indicates UMR over the different developmental stages. LMR and UMR were selected to show the dynamic nature of the methylome. FMR is only displayed in this figure if changing to LMR or UMR in at least one developmental stage. FMR that stay unchanged during the three stages of development are excluded from these plots.

Differential methylation analysis

To get a more detailed overview of the changes in methylation during chicken development, we performed per-site (RRBS, three samples per tissue per time-point) and region (WGBS, one sample per tissue per time-point) differential methylation (DM) analyses. Number of DM sites (DMS) and DM regions (DMR), together with associated genes are shown in **Table 1**. For these and subsequent analyses the E8 to E15 transition was defined as developmental transition and transition from E15 and hatching as maturation transition.

Kidney and muscle exhibited the lowest numbers of DMS during the developmental and maturation transitions. In contrast, the brain and small intestine showed the highest numbers of DMS at the developmental and maturation transitions. The number of DMS increased during the transition from development to maturation in the kidney and small intestine, while a decrease was observed in the skin, brain, muscle, liver, and lung.

For DMR, the brain exhibited the highest number of DMR during the developmental transition, while the kidney had the lowest. During the maturation transition, the small intestine displayed the highest number of DMR, and the lung had the lowest. Similarly to DMS, the number of

DMR decreased during the transition from development to maturation in the skin, brain, and muscle, while an increase is seen in the kidney, liver, lung, and small intestine.

We further divided both DMR and DMS into hyper and hypo-methylated changing regions based on the directional changes in the methylation, where hypo-methylation changes indicates a decrease in methylation level from one developmental stage to the next, and hyper-methylation changes indicates an increase (**Fig. 5**).

Table 1 Counts of DMS and DMR between developmental stages (E8 to E15 = Developmental transition (Dev); E15 to HC = Maturation (Mat) transition) for tissues. Thresholds of 15% difference and q-value <0.01 were applied.

Tissue	RRBS (DMS)				WGBS (DMR)			
	Developmental transition	Maturation transition	Promoter of genes		Developmental transition	Maturation transition	Promoter of genes	
			Dev	Mat			Dev	Mat
Skin	22,696	6,771	1,375	601	198,794	86,679	3,890	1,682
Brain	59,114	13,956	2,556	903	263,444	121,528	5,157	2,131
Muscle	24,394	4,276	1,370	479	212,851	79,900	4,138	1,468
Kidney	11,240	12,826	893	860	29,848	118,081	593	2,229
Liver	12,099	5,756	1,002	552	63,274	109,216	1,147	2,088
Lung	18,732	18,844	1,180	1,318	30,227	50,993	676	1,101
Small intestine	19,642	25,661	1,344	1,738	116,920	201,024	2,316	4,086

The largest numbers of hypo-methylated changing regions (WGBS) are observed in brain, muscle and skin at the developmental transition and in the small intestine at the maturation transition (**Fig. 5a**). DMS showed similar patterns in the number of hyper- and hypo-methylated changing sites for muscle, small intestine, and skin (**Fig. 5b**) compared to DMR patterns. Interestingly, the brain showed a larger proportion of hyper-methylated changing sites to hypo-methylated sites (**Fig.5b**), whereas for DMR a larger proportion of hypo-methylated changing regions to hyper-methylated regions is observed. Similarly, the lung has a higher proportion of hyper-methylated changing sites compared to the number of hyper-methylated changing regions. Tables of all DMR and DMS together with associated promoter of genes are shown in **Data S6** (DMS) and **S7** (DMR).

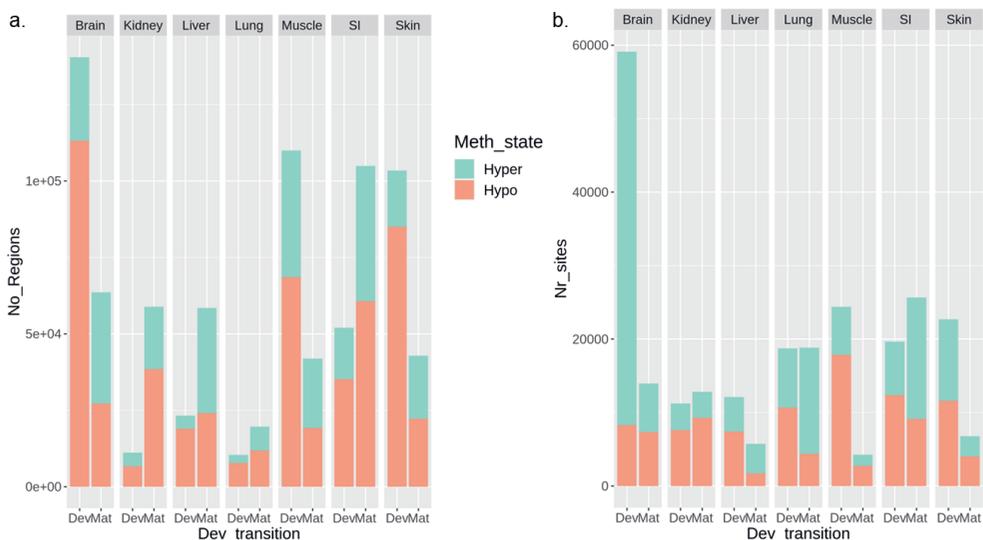


Fig. 5 Distribution of **a.** DMR (WGBS data) and **b.** DMS (RRBS data), showing the number (y-axis) of hypo- and hyper-methylated changing regions per tissue (on top x-axis), for each embryonic transition (bottom x-axis). Dev is defined as developmental transition and Mat defined as maturation transition. Hyper- and hypo-methylation are shown in orange and green stacked bars, respectively. SI indicates small intestine.

Tissue specific methylation

Tissue specific DMR and DMS (**Fig. 6a**), together with the associated gene promoter regions, were identified for each tissue (**Data S8**, and **S9**, respectively). The brain and small intestine showed the highest number of tissue specific DMR and DMS. The lowest number of tissue specific DMR is identified for lung, followed by liver and kidney. Lung had a high number of tissue specific DMS in comparison to DMR. Generally, there were more DMS than DMR in the brain, kidney, liver, lung, and small intestine.

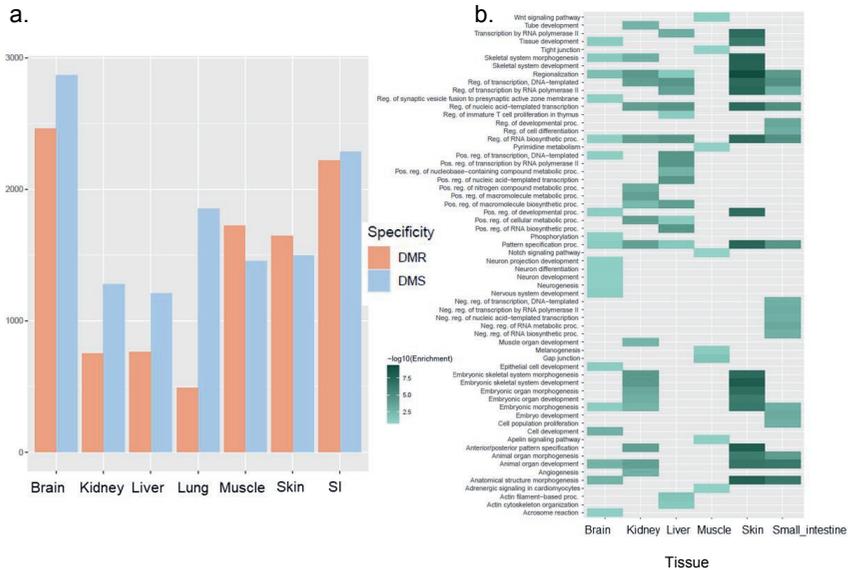


Fig. 6 a. Number of tissue specific DMR (WGBS) and DMS (RRBS) for each tissue. **b.** Significant enrichment of biological process terms (FDR<0.05) for genes associated with tissue specific DMR. Enrichment of cellular component terms for tissue specific DMR/S genes are shown in **Fig. S6** and **Table S6**, respectively.

Enrichment analysis of the genes associated with tissue specific DMR, showed significant enrichment of biological process terms in brain such as neuron projection development, neurogenesis, neuron differentiation, and tissue development while for kidney tube development is significantly enriched (**Fig. 6b**). Furthermore, various general growth and molecular GO biological process terms, such as regionalization, DNA-templated, animal organ development, and pattern specific processes, are enriched in multiple tissues (brain, kidney, liver, skin, and small intestine).

On an individual gene level, we investigated genes known for their roles in embryonic development, tissue differentiation, and organogenesis. For tissue specific DMS the following interesting genes are observed, *PAX3*, *PAX5*, *PAX6* and *PAX7* in brain, and *PAX3*, *PAX1*, *PAX7* in muscle. Furthermore, for tissue specific DMR, *HOXD11* and *HOXD3* were identified in kidney and skin, respectively.

Assessing functional enrichment from DM-associated differential expressed genes (DMEG) during development

Additional insight into the regulation of gene expression was gained by combining both differentially methylated regions and sites with differentially expressed genes, respectively. In this section, we focus primarily on the DMR integrated with differentially expressed genes (DMEG) (**Fig. 7a**). We observed the largest number of DMEG at the developmental transition in brain, and skin and the lowest number in kidney and lung (**Fig. 7a**). Additionally, we combined DMS with differentially expressed genes and the results for DMEG and DMS associated with differentially expressed genes can be found in **Data S10** and **S11**.

Fig. 7b shows DMEG in skin at the developmental transition with corresponding expression and methylation change. We observed more hypo-up and hypo-down genes in comparison to the hypermethylation categories (hyper-down, hyper-up). We then used functional enrichment analysis to investigate the up-regulated (combined hyper- and hypo-methylated) and down-regulated (combined hyper- and hypo-methylated) categories based on gene expression (TPM) for the identified genes. Results for skin are shown in **Fig. 7c,d** and significant results for other tissues are shown in **Fig. S7**. Biological processes such as collagen fibril organization, regulation of cell adhesion, positive regulation of response to stimulus were significantly enriched in the upregulated DMEG group (**Fig. 7c**). Secondly, for down-regulated DMEG, we observed significant enrichment of cellular component GO terms such as neuron projection (**Fig. 7d**). Interestingly, brain, liver, muscle, and small intestine showed down-regulation of many general developmental processes (**Fig. S7b**), including embryonic morphogenesis, embryo development, embryonic organ morphogenesis, anatomical structure morphogenesis, and animal organ morphogenesis mainly during the developmental transition. In the kidney during the developmental transition, we observed an enrichment of the A-band cellular component GO term in up-regulated genes (**Fig. S7a**).

On an individual gene level, we investigated genes that are known to play crucial roles in embryonic development, organogenesis, and germ layer formation. *GDF6*, and *GDF8* showed up-regulation in muscle, lung, and skin during the developmental transition, respectively. *WNT3A* is down-regulated in brain tissue during the developmental transition. Fibroblast Growth Factor (*FGF*) genes are down-regulated during the developmental transition in the small intestine (*FGF3*) and down-regulated during the maturation transition in lung, muscle, and the small intestine (*FGF6*). Among the identified *HOX* genes, *HOXD10* was down-regulated in muscle, *HOXA4* in skin, and *HOXA2* in lung and skin during the developmental

transition. Additionally, *BMP4* and *BMP11*, showed up-regulation in muscle and skin during the developmental transition.

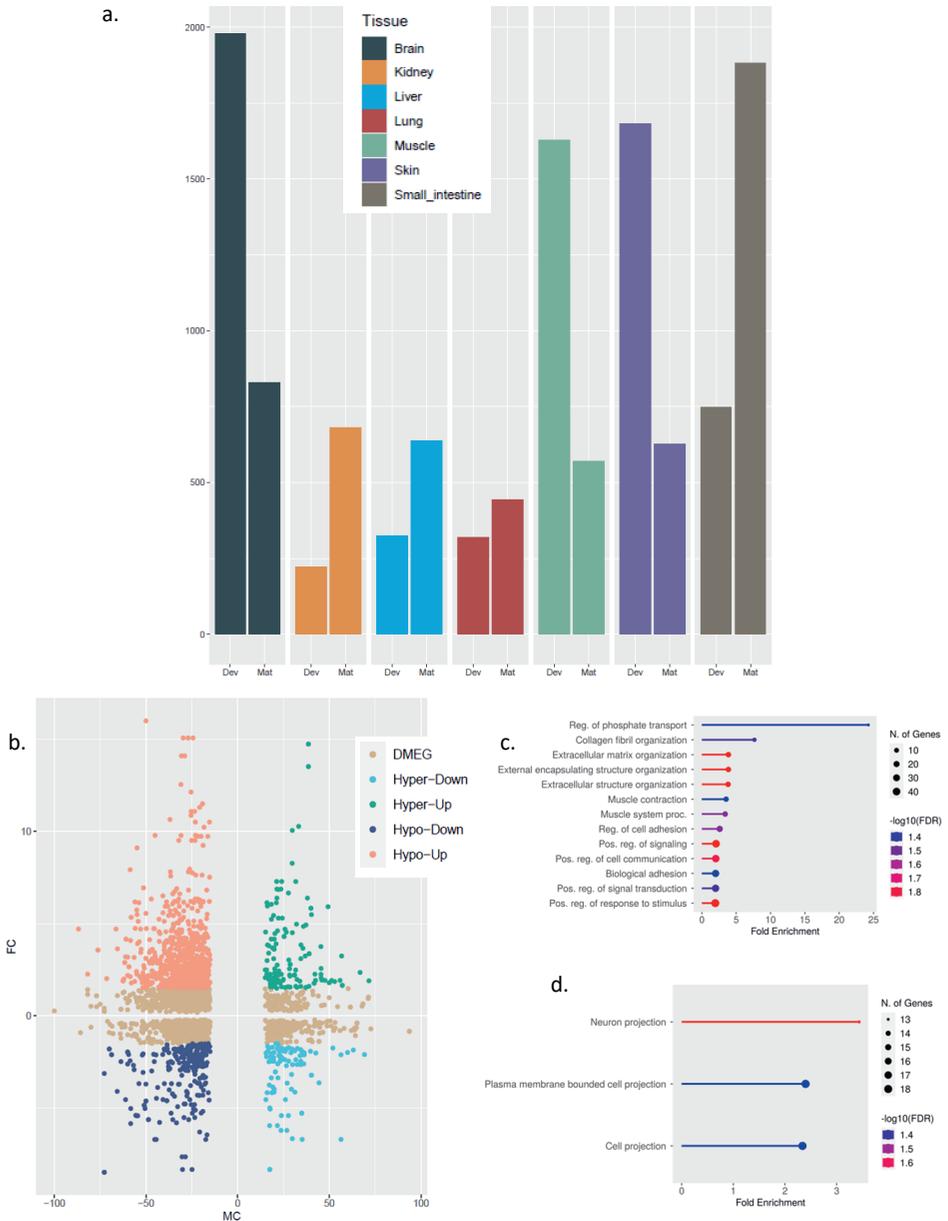


Fig. 7 a. Number of DMEG in chicken, shown per transition in the context of development. Various tissues are shown in corresponding colours. **b.** Scatterplot showing the relationship between expression and methylation in the skin at the developmental transition (E8 to E15),

5 DNA methylome dynamics in chicken development

with the expression fold change (FC) on the y-axis and methylation change (MC) on the x-axis. Categories of methylation change i.e. hypo/hyper and expression change (up- or down-regulated) are shown in different colours. DMEG category represents all differentially methylated expressed genes, which have an expression FC between -1.5 and 1.5 (filtered out). Further categories are defined as: up-regulated has $FC > 1.5$ and down-regulated has $FC < -1.5$. Hyper-methylation: $MC \Rightarrow 15\%$ and hypo-methylation: $MC \leq -15\%$. **c.** Biological processes gene ontology (GO) terms for up-regulated genes ($FC > 1.5$) in skin at the developmental transition. **d.** Cellular components (CC) GO terms for down-regulated genes ($FC < -1.5$) in skin at the developmental transition

Comparative analysis between pig and chicken

The dynamics of the methylome during fetal development in pig is reported in Chapter 4 of this thesis. Since the stages of development examined in pig (chapter 4) and chicken (this chapter) were chosen to represent similar stages of development, results of these two studies can be used for comparative analysis to gain some preliminary insight into similarities and differences in the methylome during vertebrate development. We first investigated the level of DNA conservation in DNA methylation defined regulatory elements (UMR and LMR). The UMR and LMR for the two species were compared to human by aligning the regions to the human reference sequence (**Fig. 8a**). We observe a high DNA sequence conservation for UMR (~90%) and LMR (~80%) between pig and human (**Fig 8a**). The DNA sequence conservation for UMR between chicken and human is moderate ranging from 50 to 65% (**Fig. 8a**), with the UMR in brain showing the highest sequence conservation. LMR sequences have low conservation between chicken and human, between 10 and 30%, with cerebellum showing the highest LMR conservation (~30%).

The conserved DNA-sequences between human-pig and human-chicken were used to compare the number of DNA methylation regions conserved between pig and chicken in the context of UMR and LMR. The conserved UMR and LMR between chicken and pig were then associated with genes (TSS + 1500 bp upstream) and the number of associated genes is shown in **Fig. 8b**. The highest variation among tissues and development stages was observed for LMR. For example, we observed no LMR-genes conserved between pig and chicken in liver at the third developmental stage. Generally, more LMR-genes are identified as conserved between pig and chicken at the second and third developmental stage, specifically for the skin, small intestine and brain. The number of UMR-genes is lower in comparison to the number of LMR-genes conserved across all tissues and developmental stages. Recent studies have

highlighted the association of LMR in both birds and mammals with transcription factor binding sites, suggesting its potential significance in epigenetic programming. Given these implications, the level of conservation of LMR between pig and chicken could indicate its conserved functional importance across species.

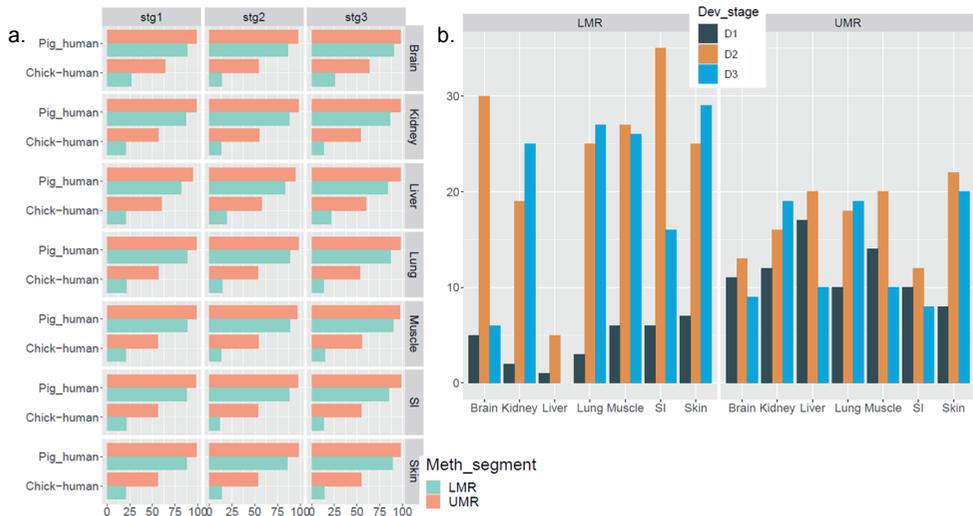


Fig. 8 Comparative analysis of UMR and LMR between pig and chicken. **a.** Alignment percentage of pig and chicken methylome segments (UMR and LMR) to human sequences. Tissue and time-point are shown as a label at the top and on the right y-axis, and on the x-axis the percentage of sequence aligning between each species and human is shown. **b.** Number of genes associated with the conserved UMR and LMR regions between pig and chicken.

Discussion

DNA methylation plays a role in the regulation of gene expression, and changes dynamically during development and during cell, and tissue differentiation. In this study, we investigated the spatio-temporal dynamics of the DNA methylome during chicken fetal development. Using key tissues at three developmental stages representing early and late organogenesis and hatching, we provided the first insight into the methylome's role in regulating avian fetal development and tissue differentiation. Similar studies have been performed in mouse [16] and pig (Chapter 4), both representatives of the mammalian clade, which provides the opportunity for insights into similarities and differences between species during vertebrate development. The pig methylation data are from the same tissues and sampled at comparable developmental

5 DNA methylome dynamics in chicken development

stages. The comparable stages between chicken and pig, respectively are E8 and 30 dpf (days post fertilization) representing early organogenesis, E15 and 70 dpf representing late organogenesis, and lastly hatching and new-born.

On a global methylation scale we observed similar levels as previously described in adult chicken [20], during development of skeletal muscle in chicken [19], and in adult great tit [21,22]. In contrast, we observe lower global methylation levels in liver and muscle in comparison to a study in chicken embryos (day 2 until 21 days) [23]. CpG methylation levels are similar between the different tissues at the same developmental stage, and there is a general trend in decreased levels of methylation from early stages of development (E8) to hatching stage, which is similar to observations during Great tit development [24]. Global CpG methylation levels across all tissues and time-points are lower in comparison to the observations in pig (Chapter 4) and mouse [16].

Methylation levels at the promoter and TSS show the expected drop, followed by increased methylation in the gene-body (GB) as previously observed in birds such as chicken and great tit, and mammals such as pig, mouse and human [16,20–22,25]. However, methylation levels within the gene-body are lower in chicken (~0.6) compared to pig (~0.8). This is in concordance with previous studies, where avian methylation levels are lower and frequencies of fully methylated sites are lower compared to mammals in similar tissues during development and in adult stages [16,21,23,25]. The higher methylation levels within the gene-body in both mammalian and avian species suggest this as a likely mechanism which is conserved among species. However, the differences in the high level of methylation in the gene-body of mammals and birds have been suggested to be due to transposable elements whose activity is suppressed by methylation and thereby inhibiting transposable element insertions, which can disrupt the genome. Mammals have a higher amount of transposable elements than birds, and thus more methylation is required within the gene bodies [26–29].

In our study both UMR and LMR, are found in close proximity to the TSS. These findings contrast with previous results in pig (Chapter 4) and human [20], as well as the common assumption that UMR are indicative of promoters and LMR of enhancers. This could be due to the method of methylome segregation [21], and assumptions which are based on the mammalian genome, while here we are investigating a bird genome. Furthermore, the chicken genome is more gene dense, and smaller in size in comparison to mammalian genomes which could explain the closer proximity of both UMR and LMR to TSS in chicken.

MYLK4, *PDIA5* and *HDAC2* genes are identified in muscle transitioning from FMR at E8 and E15 to LMR at hatching. The *MYLK4* [30] and *HDAC2* [31] have specific roles in muscle

development, with the latter identified in skeletal muscle development in chickens. *HOXB1*, transition from FMR at E8 and E15 to LMR at hatching in the small intestine. *HOXB1* is important in morphogenesis, and has been identified in intestinal epithelial cells, as well as in the mesoderm and gut endoderm [32,33].

Remarkable differences in methylation patterns are evident in the number of hyper- and hypo-methylated regions across different tissues and developmental transitions in both pig and chicken. Notably, at the developmental transition, pig liver has only a few hyper- and hypo-methylated regions. However, during the maturation transition, a considerable number of hyper-methylated regions are identified in pig liver. In contrast, the liver tissue in chicken does not show any remarkable differences compared to other tissues. Differences between the methylation dynamics of liver during fetal development between the pig and chicken can be due to differences in tissues differentiation between mammals and birds during organogenesis. In eutherian mammals the yolk sac is smaller, and nutrients are primarily received from the mother through the placenta. In contrast, the yolk sac in birds is much larger and nutrients are primarily stored and utilised here [34]. During early developmental stages the yolk sac has a hematopoietic function, which is replaced by the liver during the fetal development stages in eutherian mammals, explaining why the liver is one of the first organs to develop fully [1,3,34]. In contrast, birds have a different embryonic development pattern where the liver develops simultaneously with other organs and tissues.

Brain development in chickens (birds) displays more changes in the methylome during fetal development than in pigs (mammals). Specifically, a larger proportion of hypo-methylated regions are identified in the chicken methylome during the developmental transition, in comparison to pigs. Functional enrichment analyses show no brain-specific GO terms for tissue-specific DMR in pigs, whereas in chickens, tissue-specific terms such as neurogenesis and neuron development are identified.

In skin, tissue specific functions start developing from late organogenesis (E15) onwards, e.g. enrichment of extracellular matrix organization and collagen fibril organization in upregulated genes during the developmental transition. Skin has 3 layers, namely the epidermis (top layer), dermis (middle layer) and subcutaneous tissue. Collagen and keratinocytes are found in the epidermis layer of the skin, and the extracellular matrix is found in the connective tissue of the subcutaneous layer [35]. Skin and brain develop from the ectoderm germ layer, and the downregulation of neuron projection from late organogenesis (E15) shows the differentiation of these tissues from late organogenesis stage. This observation is seen in other tissues (muscle, small intestine) and is also seen in the pig during development in kidney and muscle [1].

5 DNA methylome dynamics in chicken development

The identification of well-known genes (e.g. *HOX*, *PAX*, *WNT*), that play a crucial role in embryonic development, tissue differentiation, and growth across multiple tissue types shows the importance of these genes in shaping the developmental trajectory of these tissues in both chicken and pig. Furthermore, our observations of the differential expression and methylation of these genes together with their developmental roles further accentuate their significance in regulating tissue development.

Conclusion

Our study provides, to our knowledge, the most comprehensive analysis of the DNA methylation dynamics during development of seven tissues in chicken, expanding on previous research on adult tissues and a single embryonic tissue [19,36]. We reveal methylome dynamics by segmenting the methylome, and performing differential methylation analyses, which show sites and regions changing during development. The regulatory function of methylation for gene expression is shown by combining our results with gene expression data. Remarkable differences in the methylome patterns between chickens and pigs are observed, with the most notable differences found in the liver tissue during development. Liver is fully developed at an early stage of development in pig in comparison to chicken, which is due to the hematopoietic function of the liver in mammals during development in comparison to birds. The spatio-temporal epigenomic data sets described here, are an invaluable resource for addressing fundamental questions about gene regulation during chicken tissue and organ development.

Methods

Experimental model and tissue collection

Tissues, including the brain (hindbrain at E8, and cerebellum at E15 and hatching), liver, kidney, skin, muscle (hindlimb muscle at E8 and E15 and *Gluteus Medius* at hatching), lung, and small intestine (small intestine at E8 and E15 and ileum at hatching), were sampled at dissected to remove and dissect the organs. Tissues were processed using a standard procedure of cleaving the organ and then each organ was stored individually. The embryo was sexed using a standard PCR procedure. Hatching samples were collected from fertilized eggs incubated for 21 days, using a standard procedure and the hatched chick was euthanized. Thereafter the organs were dissected and, small pieces were stored for DNA and RNA extraction.

DNA and RNA were extracted from the stored samples following a standard protocol which is available on the [FAANG DCC](#). DNA was extracted following the protocol from AllPrep DNA/RNA/miRNA from Qiagen. The protocol was the same for all tissues and developmental stages, except for kidney where optimizations were necessary and those are explained in more detail in the protocol. Sexes selected for methylation sequencing were as follows, RRBS: E8 – pooled sample, E15 – female, male, male, hatching – female, male, male; and WGBS: E8 – pooled, E15 - female, hatching - female.

Bisulfite sequencing (RRBS and WGBS)

[RRBS](#) (three samples x tissue x developmental stage) and [WGBS](#) (one sample x tissue x developmental stage) library preparations are described in the protocol available on the FAANG DCC. In brief, the Tecan Ovation RRBS Methyl-Seq kit and the Illumina NovaSeq 6000 were used to construct and sequence the RRBS libraries. MspI enzyme was used to cleave the isolated DNA, and fragments between 50 and 500 bp were selected. The bcl2fastq v2.17.1.14 conversion software was used to generate and demultiplex fastq files following the sequencing of all samples. Illumina NovaSeq 6000 at Novogene, was used for the creation and sequencing of the WGBS libraries, furthermore EZ DNA Methylation Gold Kit was used for bisulfite conversion and fragments were selected and quantified using the KAPA HiFi HotStart ReadyMix. The libraries were then sequenced with a PE150 strategy.

Raw data analysis & clustering analysis

NuGEN Technologies' diversity trimming [scripts](#) were used for initial trimming of the raw RRBS FASTQ files to remove extra sequences added by the diversity adaptors. Afterward, reads were filtered to include only those beginning with the expected YGG trinucleotide sequence. The resulting dataset averaged 38 million reads for each library of RRBS (ranging from 28.5 million to 49 million) and 154 million reads for WGBS data (ranging from 114 million to 193 million). The coverage of all reads was investigated, yielding an average of 46X and 22X for RRBS and WGBS data, respectively. Quality statistics can be found in **Tables S4** (WGBS) and **S5** (RRBS). Additionally, we determined the bisulfite conversion rate (~99.5%) of samples by calculating the ratio of unmethylated to the total number of reads for covered cytosines on the mitochondrial genome for RRBS datasets. This rate was used as a quality metric to determine the efficiency of bisulfite conversion. Bisulfite conversion rate for WGBS datasets (~99.5%) were provided by sequence provider. The datasets were processed using the [GSM-pipeline](#) v1.0 used for RRBS and v2.0 for WGBS, on the Gallus gallus GRCg6a reference genome assembly (https://ftp.ensembl.org/pub/release-104/fasta/gallus_gallus/dna/). The reads were initially processed using the TrimGalore tool [37] with default settings for trimming. Bismark aligner was used within the pipeline, with the parameter --rrbs or --bam specified for RRBS

and WGBS data, respectively. The pipeline subsequently converts binary alignment files to CGmap files, where reports of methylation rates for all covered CpGs are provided. The information is also provided in extensive bismark methylation reports. Additionally, global methylation rates per sample were reported. The CGmap file was converted to CpG report text files, providing methylation distributions per sample, including correlations and clustering between various tissues at the three time points, correlations between tissues from the same germ layer, and developmental stage. PCA and hierarchical (Ward.D2 cluster method) clustering and correlations were performed using the generated CpG text report files in MethylKit v1.18.0 [38].

Methylome segmentation

Firstly, CGmap files from WGBS data were filtered to only include CpGs with a coverage of at least 10 reads per sample. UMR and LMR were identified by using the filtered CGmap files as input into the MethylSeekR v1.32.0 package in R [39]. The segmentation criteria for the methylome involved (1) a false discovery rate (FDR) of less than 5% for regions, (2) DNA methylation levels ranging from 0-50% for UMR and 10-50% for LMR, and (3) more than 5 CpGs per region. Further distinction between UMR and LMR was made based on the CpG site density within the regions. UMR were found to be CpG-rich (>30 CpG sites per region) whereas LMR were CpG-poor (<30 CpG sites per region) [39]. FMR were further identified using custom scripts, with the following requirements: (1) CpG sites with methylation levels >0.75, (2) regions of 1000 bp with more than 10 CpG sites, and (3) removal of redundant regions repeated in UMR/LMR. Methylome dynamics were determined using the intersect function in bedtools v2.30.0 [40] and regions were required to have at least 10% overlap. FMR at all stages of development were excluded, and the results were plotted in R. Scripts are available on request.

Differential methylation analysis

The CGmap report text files generated from the RRBS and WGBS data were filtered to exclude CpG sites with less than 10 reads and bases/regions with a higher coverage than the 99.5 percentile. The read coverage distributions between samples were normalized using the MethylKit (R) package v1.18.0 [38]. Reads on both strands of a CpG nucleotide were merged to obtain better counts of cytosines at a specific CpG site. Regions in WGBS were calculated using a sliding window approach with a window size of 500 bp and a step size of 250 bp. A methylation difference calculation between development phases for each tissue was performed using a pairwise comparison with an F-test, together with basic correction for overdispersion (MN) [38,41]. Pairs of E8 to E15 were defined as developmental transition, and E15 to hatching as maturation transition. Methylation differences for RRBS data were

calculated on a per site level, whereas for WGBS data, methylation differences were calculated on a regional level. CpG sites/regions were considered DM if there was >15% difference in methylation rate for each transition (e.g. E8 and E15) and having a corresponding q-value <0.01. Lastly, hyper- and hypo-methylated sites/regions were identified per tissue for the two defined transitions.

DMS and DMR were annotated using the Ensembl Gallus gallus GRCg6a - *release 104* annotation (https://ftp.ensembl.org/pub/release-104/gtf/gallus_gallus/). The core promoter regions were defined as the transcription start site (TSS) of the gene with an additional 500 bp upstream and 200 bp downstream from the TSS, taking the strand into consideration for all genes. Intergenic regions were defined as starting 1000 bp upstream of the gene TSS, until the next TSS. We annotated the hypo- and hyper-DMR using the described custom region files with bedtools intersect.

Identifying tissue specific promoters

Tissue-specific promoters were identified by first using the Ensembl Gallus gallus GRCg6a - *release 104* annotation to determine promoters. Promoters were defined as 1200 bp upstream and 200 bp downstream of the TSS with a consideration for strandedness. This region is larger in comparison with the 'core promoter' region, as the aim is to identify potential enhancers close to promoters and in this way defining a larger region. These promoter regions were then combined with associated DMR from WGBS and DMS from RRBS for each tissue during the developmental transition. Thereafter all DMR and DMS for a given tissue were determined (thus all methylation changes occurring in both transitions). Next, DMR/DMS that were specific to each tissue across all transitions were identified using bedtools intersect with the -v flag (e.g. bedtools intersect -v -a Tissue1_all_transitions -b all_other_tissues). Unique genes were extracted from each tissue-specific DMR and DMS and investigated for significant enrichment in gene ontology (GO) terms using ShinyGO v0.77 [42].

Visualisation of methylation levels and transposable elements

Relationship of methylation levels and various genomic regions were investigated using ViewBS v0.1.11 [43], where the methylation reports created from bismark were implemented together with custom region files of repeats and chromatin states. Distribution of methylation levels over promoter, gene-body and transcription termination sites (TSS) were achieved by creating a region file using the GRCg6a version 104 annotation file. Genes were classified based on their expression levels (TPM) as follows: (i) ≥ 0 to < 1 , (ii) ≥ 1 to < 10 , (iii) ≥ 10 to < 30 , (iv) ≥ 30 to > 75 , (v) ≥ 75 . The plot was generated using the MethOverregion functionality within ViewBS.

5 DNA methylome dynamics in chicken development

Methylation patterns in different repeat classes were visualised using the MethOverregion function. The epigenomic states from the associated tissue were obtained from Pan, *et al* (2023) [36] and visualisation was performed using the MethOverregion functionality. The methylation level distribution was obtained using the ViewBS MethLevDist utility. Genome wide methylation patterns were obtained using the ViewBs MethGeno utility.

Integrative analysis of expression and methylation

The relationship between gene expression and DNA methylation was investigated by using curated gene expression data (available on the FAANG DCC) in the context of TPM expression values per tissue and time-point in 4 replicates. Furthermore, differentially expressed genes were available per tissue for each transition in development, while promoter DMR and DMS were overlaid with the corresponding differentially expressed genes using bedtools intersect with the -wo flag to retain information regarding differentially expressed genes and DMR/DMS. The categorization of DMEG were performed based on the methylation and expression changes, where DMEG were classified as up-regulated with a $\log(\text{FC}) > 1.5$, down-regulated with $\log(\text{FC}) < -1.5$, hyper-methylated: methylation change $> 15\%$, and hypo-methylation if methylation change < 15 . Lastly, GO analysis was performed for all genes per tissue at the developing transitions and terms were significantly enriched with a $\text{FDR} < 0.05$. Top pathways were obtained and plotted together, followed by GO (ShinyGO; [42]) analysis for the up-regulated and down regulated categories as stated above.

Comparative analysis of regulatory regions between species

For alignments of methylome segments (UMR, LMR), we implemented a similar method as Roller *et al.* [44]. The regions were aligned using LastZ algorithm from Ensembl release 90 (pig) and release 95 (chicken). Human was used as the reference species for alignment of UMR and LMR in the two species (pig, chicken). The aligned regions of corresponding tissue-time points in human were then compared (bedtools) between pig and chicken for identification of conserved methylated regions involved in regulating development. Lastly the promoter regions of genes were associated with the conserved regions to determine the number of associated genes.

Supplementary information

Supplementary tables

Download: https://osf.io/xn5vq/?view_only=bdf728b6c28546d693d4c65cab9a20a3

(Excel file – Supplementary_tables.xls)

Supplementary data files

Download: https://osf.io/xn5vq/?view_only=bdf728b6c28546d693d4c65cab9a20a3

S1: Methylation over chromatin states per tissue time point. Notated as GG_tissue_timepoint_CG_MethOveRegion.pdf

S2: Methylation levels over repeat elements per tissue time point and methylation levels per repeat region for all tissues. Notated as GG_tissue_timepoint_CG_MethOveRegion.pdf and repeat_name_MethOverRegion_CG.pdf

S3: UMR region files. Notated as UMR_GG_tissue_timepoint.txt

S4: LMR region files. Notated as LMR_GG_tissue_timepoint.txt

S5: Tissue transition files. Notated as tissue_transitions.bed

S6: Differentially methylated sites per tissue per transition files, hypo - and hyper methylated sites files and genes associated with DMS files. Notated as Diff15p_OD_Ftest_tissue_time_points.bed,

Promoter_{Hyper/Hypo}_OD_Ftest_tissue_time_points.txt

S7: Differentially methylated regions per tissue per transition files, hypo - and hyper methylated sites files and genes associated with DMS files. Notated as Diff15p_OD_Ftest_tissue_time_points.bed,

Promoter_{Hyper/Hypo}_OD_Ftest_tissue_time_points.txt

S8: Tissue specific DMS. Notated as Tissue_specific_dms.bed and genes_Tissue_specific_dms.bed

S9: Tissue specific DMR together with genes. Notated as GENES_Tissue_specific_dmr.bed

S10: DMEG files. Notated as tissue_time_points_DM_DEG.bed

S DNA methylome dynamics in chicken development

S11: DM sites associated with DEG files. Notated as tissue_time_points_DM_DEG.bed

S12: Methylation levels over the whole genome for each sample. Notated as tissue_Genome_Methylation_MethGeno_CG.pdf

Additional information:

E8 pooling protocol:

https://data.faaug.org/api/fire_api/samples/ROSLIN_SOP_GENESWITCH_E8_EMBRYO_POOLING_20200915.pdf

Sampling at E8 protocol:

https://data.faaug.org/api/fire_api/samples/ROSLIN_SOP_GENESWITCH_E8_EMBRYO_SAMPLING_20200915.pdf

Sampling at E15 protocol:

https://data.faaug.org/api/fire_api/samples/ROSLIN_SOP_GENESWITCH_E15_EMBRYO_SAMPLING_20200915.pdf

Sampling at hatch protocol:

https://data.faaug.org/api/fire_api/samples/ROSLIN_SOP_GENESWITCH_HATCHED_CHICK_SAMPLING_20200915.pdf

DNA extraction protocol:

https://data.faaug.org/api/fire_api/assays/ROSLIN_SOP_GENESWITCH_WP1_CHICK_EXTENSION_DNA_RNA_20201111.pdf

RRBS library preparation protocol:

https://data.faaug.org/api/fire_api/experiments/WU_SOP_GENESWITCH_WP1_RRBS_library_preparation_20201201.pdf

WGBS library preparation protocol:

https://data.faaug.org/api/fire_api/experiments/WU_SOP_GENESWITCH_WP1_WGBS_library_preparation_20210127.pdf

NuGEN Technologies' diversity trimming scripts:

<https://github.com/nugentechnologies/NuMetRRBS>

GSM-pipeline: <https://github.com/FAANG/GSM-pipeline>

Reference genome: https://ftp.ensembl.org/pub/release-104/fasta/gallus_gallus/dna/

Annotation: https://ftp.ensembl.org/pub/release-104/gtf/gallus_gallus/

References

1. Gilbert SF, Barresi MJF. *Developmental Biology*. 11th ed. Sunderland, Massachusetts: Sinauer Associates, Inc.; 2016.
2. Hine R, editor. *A Dictionary of Biology* [Internet]. 8th ed. Oxford University Press; 2019 [cited 2023 May 4]. Available from: <https://www.oxfordreference.com/display/10.1093/acref/9780198821489.001.0001/acref-9780198821489-e-1810>
3. Kardong KV. *Vertebrates: Comparative Anatomy, Function, Evolution*. 8th ed. New York, NY, U.S.A: McGraw-Hill Education; 2019.
4. Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet*. 2014;15:453–68.
5. Heintzman ND, Ren B. Finding distal regulatory elements in the human genome. *Current Opinion in Genetics & Development*. 2009;19:541–9.
6. Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*. 2006;7:29–59.
7. Liu Y, Liu XS, Wei L, Altman RB, Batzoglu S. Eukaryotic Regulatory Element Conservation Analysis and Identification Using Comparative Genomics. *Genome Res*. 2004;14:451–8.
8. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. *Genome Biology*. 2021;22:108.
9. Smallwood A, Ren B. Genome organization and long-range regulation of gene expression by enhancers. *Current Opinion in Cell Biology*. 2013;25:387–94.
10. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*. 2019;20:207–20.

5 DNA methylome dynamics in chicken development

11. John RM, Rougeulle C. Developmental Epigenetics: Phenotype and the Flexible Epigenome. *Frontiers in Cell and Developmental Biology* [Internet]. 2018 [cited 2023 May 1];6. Available from: <https://www.frontiersin.org/articles/10.3389/fcell.2018.00130>
12. Goldberg ML, Fischer JA, Hood LE, Hartwell L. *Genetics: from genes to genomes*. New York, NY: McGraw Hill; 2012.
13. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol*. 2019;20:590–607.
14. Greenberg MVC. Get Out and Stay Out: New Insights Into DNA Methylation Reprogramming in Mammals. *Frontiers in Cell and Developmental Biology* [Internet]. 2021 [cited 2023 Mar 28];8. Available from: <https://www.frontiersin.org/articles/10.3389/fcell.2020.629068>
15. Yizhar-Barnea O, Valensisi C, Jayavelu ND, Kishore K, Andrus C, Koffler-Brill T, et al. DNA methylation dynamics during embryonic development and postnatal maturation of the mouse auditory sensory epithelium. *Sci Rep*. 2018;8:17348.
16. He Y, Hariharan M, Gorkin DU, Dickel DE, Luo C, Castanon R, et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature*. 2020;583:752–9.
17. van der Velde A, Fan K, Tsuji J, Moore JE, Purcaro MJ, Pratt HE, et al. Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol*. 2021;4:1–15.
18. Yang Y, Fan X, Yan J, Chen M, Zhu M, Tang Y, et al. A comprehensive epigenome atlas reveals DNA methylation regulating skeletal muscle development. *Nucleic Acids Research*. 2021;49:1313–29.
19. Ran J, Li J, Yin L, Zhang D, Yu C, Du H, et al. Comparative Analysis of Skeletal Muscle DNA Methylation and Transcriptome of the Chicken Embryo at Different Developmental Stages. *Frontiers in Physiology* [Internet]. 2021 [cited 2023 May 18];12. Available from: <https://www.frontiersin.org/articles/10.3389/fphys.2021.697121>
20. Al Adhami H, Bardet AF, Dumas M, Cleroux E, Guibert S, Fauque P, et al. A comparative methylome analysis reveals conservation and divergence of DNA methylation patterns and functions in vertebrates. *BMC Biology*. 2022;20:70.
21. Derks MFL, Schachtschneider KM, Madsen O, Schijlen E, Verhoeven KJF, van Oers K. Gene and transposable element methylation in great tit (*Parus major*) brain and blood. *BMC Genomics*. 2016;17:332.
22. Laine VN, Gossmann TI, Schachtschneider KM, Garraway CJ, Madsen O, Verhoeven KJF, et al. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun*. 2016;7:10474.
23. Li Q, Li N, Hu X, Li J, Du Z, Chen L, et al. Genome-Wide Mapping of DNA Methylation in Chicken. *PLoS One*. 2011;6:e19428.
24. Watson H, Salmón P, Isaksson C. Dynamic changes in DNA methylation during embryonic and postnatal development of an altricial wild bird. *Ecology and Evolution*. 2019;9:9580–5.
25. Schachtschneider KM, Madsen O, Park C, Rund LA, Groenen MAM, Schook LB. Adult porcine genome-wide DNA methylation patterns support pigs as a biomedical model. *BMC Genomics*. 2015;16:743.
26. Zhou W, Liang G, Molloy PL, Jones PA. DNA methylation enables transposable element-driven genome expansion. *Proceedings of the National Academy of Sciences*. 2020;117:19359–66.

27. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*. 2017;114:E1460–9.
28. Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res*. 2018;26:25–43.
29. Jansz N. DNA methylation dynamics at transposable elements in mammals. *Essays in Biochemistry*. 2019;63:677–89.
30. Terry EE, Zhang X, Hoffmann C, Hughes LD, Lewis SA, Li J, et al. Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. *eLife*. 2018;7:e34613.
31. Shahjahan Md, Liu R, Zhao G, Wang F, Zheng M, Zhang J, et al. Identification of Histone Deacetylase 2 as a Functional Gene for Skeletal Muscle Development in Chickens. *Asian-Australas J Anim Sci*. 2016;29:479–86.
32. Huang D, Chen SW, Langston AW, Gudas LJ. A conserved retinoic acid responsive element in the murine Hoxb-1 gene is required for expression in the developing gut. *Development*. 1998;125:3235–46.
33. Lewis SL, Tam PPL. Definitive endoderm of the mouse embryo: Formation, cell fates, and morphogenetic function. *Developmental Dynamics*. 2006;235:2315–29.
34. Ross C, Boroviak TE. Origin and function of the yolk sac in primate embryogenesis. *Nat Commun*. 2020;11:3760.
35. Frandson RD, Spurgeon TL. *Anatomy and Physiology of Farm Animals*. 5th ed. Pennsylvania, U.S.A.: Lea & Febiger; 1992.
36. Pan Z, Wang Y, Wang M, Wang Y, Zhu X, Gu S, et al. An atlas of regulatory elements in chicken: A resource for chicken genetics and genomics. *Science Advances*. 2023;9:eade1204.
37. Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. [Internet]. Barbraham Institute; 2015. Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
38. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13:R87.
39. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res*. 2013;41:e155.
40. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
41. P McCullagh, Nelder JA. *Binary data. Generalized linear models*. US: Springer; 1989. p. 98–148.
42. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*. 2020;36:2628–9.
43. Huang X, Zhang S, Li K, Thimmapuram J, Xie S, Wren J. ViewBS: a powerful toolkit for visualization of high-throughput bisulfite sequencing data. *Bioinformatics*. 2018;34:708–9.
44. Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, et al. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biology*. 2021;22:62.

Supplementary figures

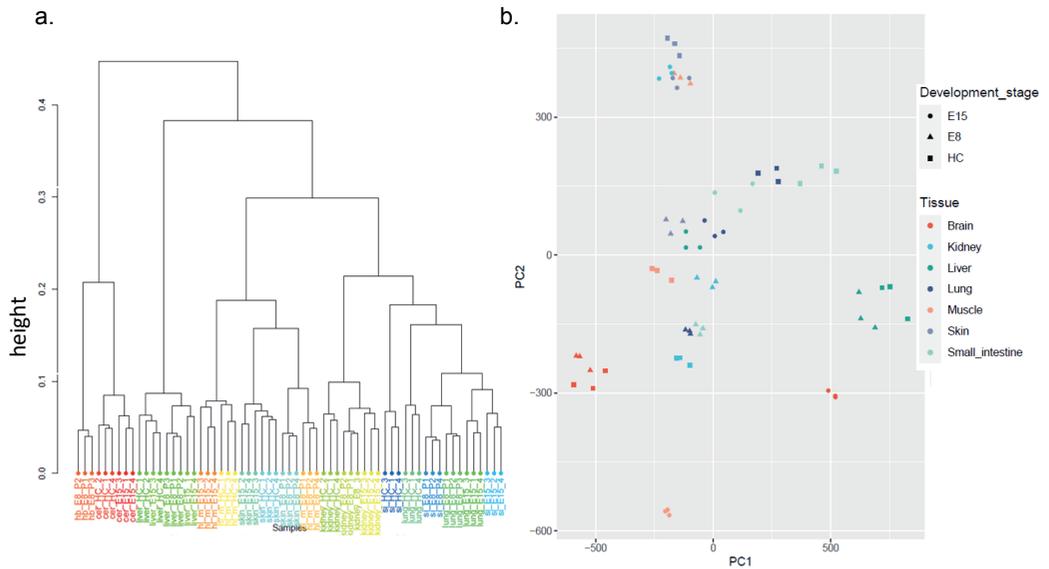


Figure S1 Clustering of all tissue samples at 3 developmental stages. Hierarchical cluster using Ward.D2 for RRBS data, with distances being measured from percent methylation per base (a). PCA clustering showing PC1 and PC2 with highest explained variance from WGBS data (b).

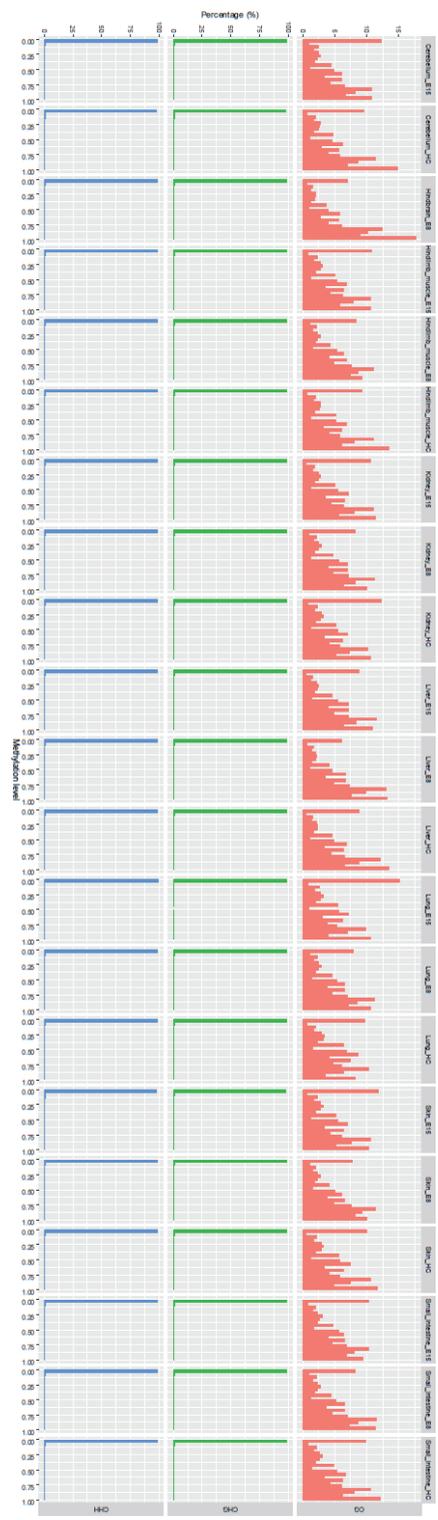


Figure Distribution of methylation across tissues per developmental stages based on WGBS data. Percentage of sites is shown on the y-axis on the left, with the type of methylation shown on the y-axis on the right. Methylation levels is shown on the x-axis with the tissue-time point displayed at the top of the figure

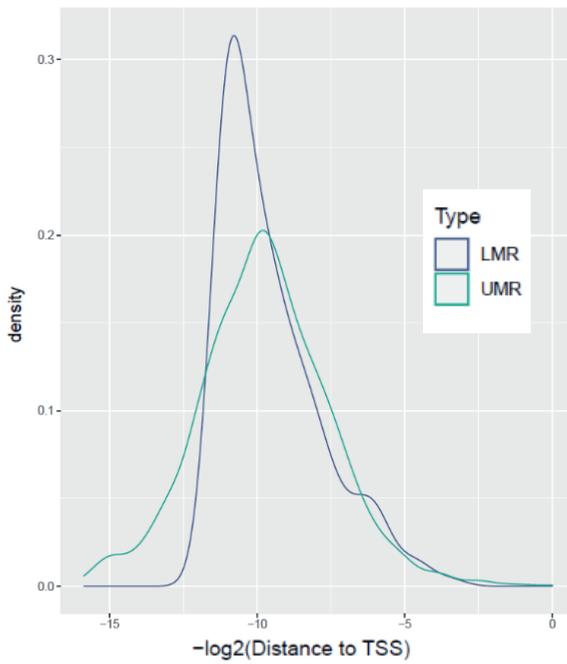
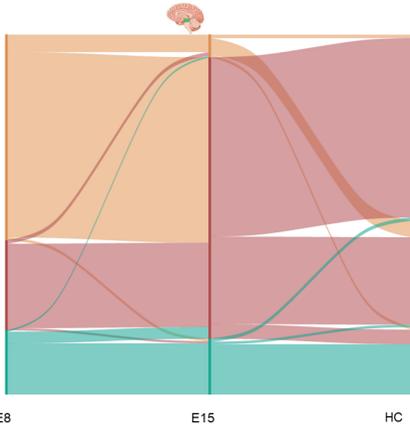


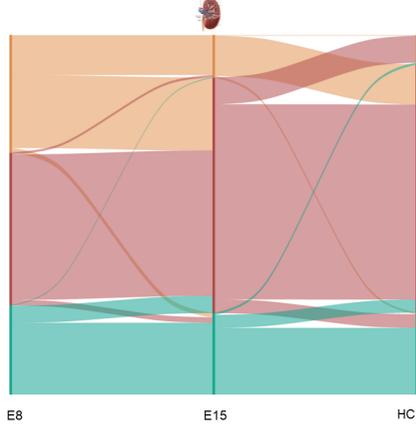
Figure S3 Density plot showing the distribution of the UMR (green) and LMR (blue) from the TSS in brain tissue at hatching. Y-axis shows the densities of the regions (UMR and LMR) and x-axis is the log distance of the regions to the TSS.

5 DNA methylome dynamics in chicken development 

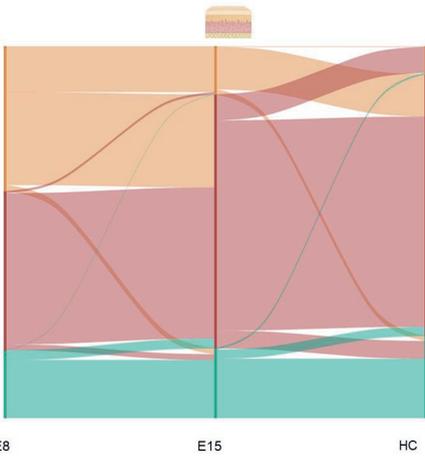
a.



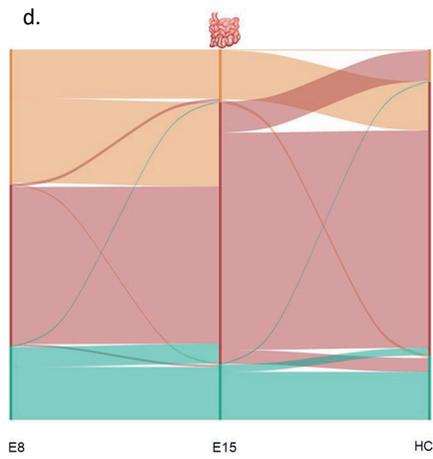
b.



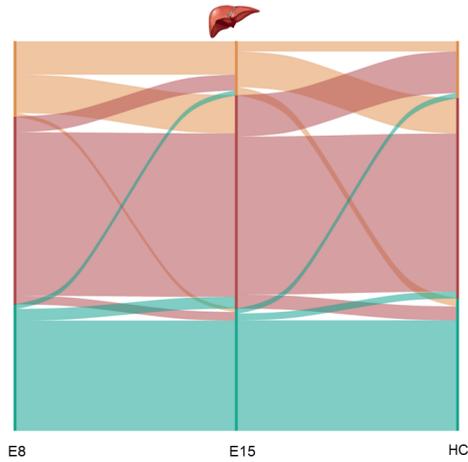
c.



d.



e.



5 DNA methylome dynamics in chicken development 

Figure S4 Dynamic changes for various tissues, with different colours depicting the methylation states (UMR, LMR and FMR). a. brain, b. kidney, c. skin, d. small intestine, e. liver.

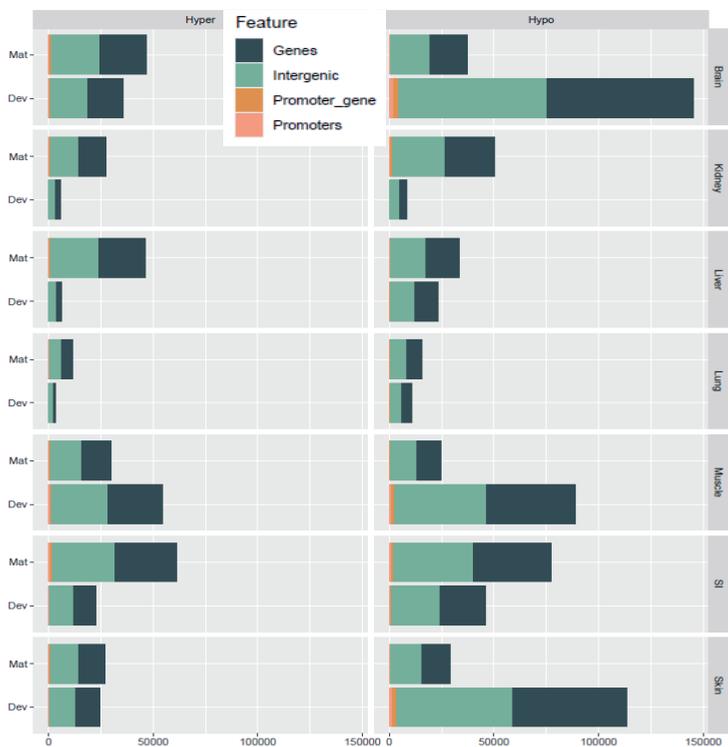


Figure S5 Annotation of the hyper and hypo methylated regions from WGBS. Features are defined as follows, intergenic is 1000 bp upstream of TSS until next TSS. Genes are defined as the gene body. Promoter is the region 500 bp upstream of the TSS. Promoter_gene are regions which are found in both promoter and gene body.

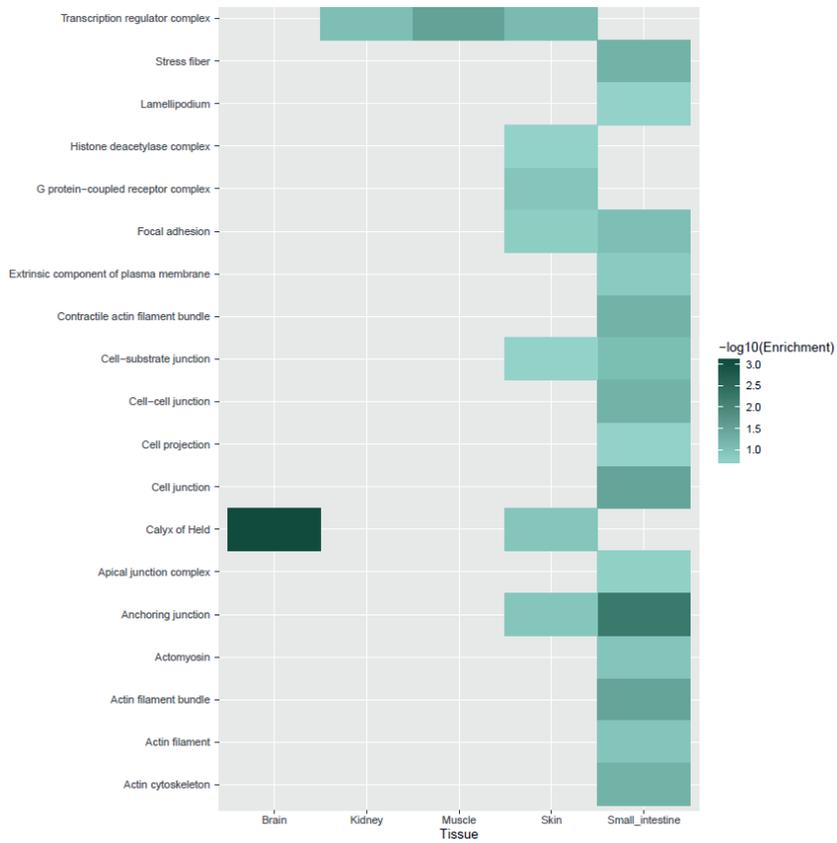


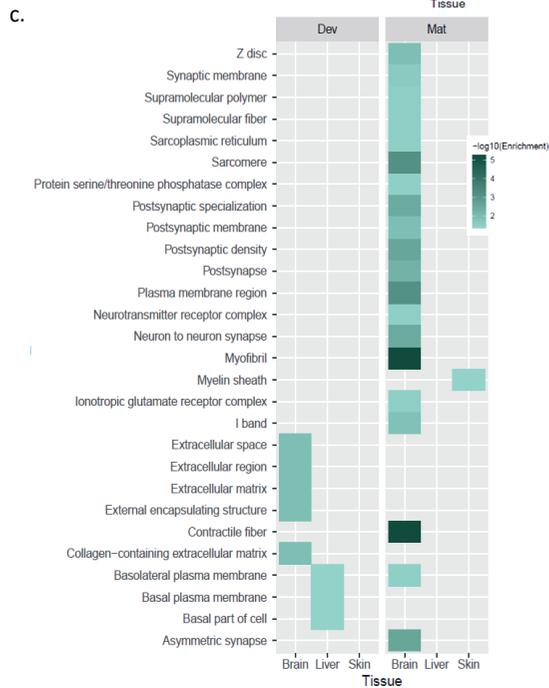
Figure S6 Enrichment analysis of cellular component GO terms for tissue specific DMR genes.

5 DNA methylome dynamics in chicken development

a.



5 DNA methylome dynamics in chicken development



5 DNA methylome dynamics in chicken development

Figure S7 Enrichment analysis of DMEG in the developmental and maturation transition of up- and down-regulated gene groups. With **a.** Biological processes (BP) gene ontology (GO) terms for up-regulated genes which are both hypo and hyper methylated. **b.** Biological processes GO terms for down-regulated genes which are both hyper and hypo methylated. Cellular component (CC) GO terms for **c.** up-regulated genes. All terms are filtered for an $FDR < 0.05$. Genes are filtered as follows for classifications: up-regulated $\log(FC) > 1.5$, and down-regulated $\log(FC) < -1.5$. All methylated regions have a minimum difference of 15%.

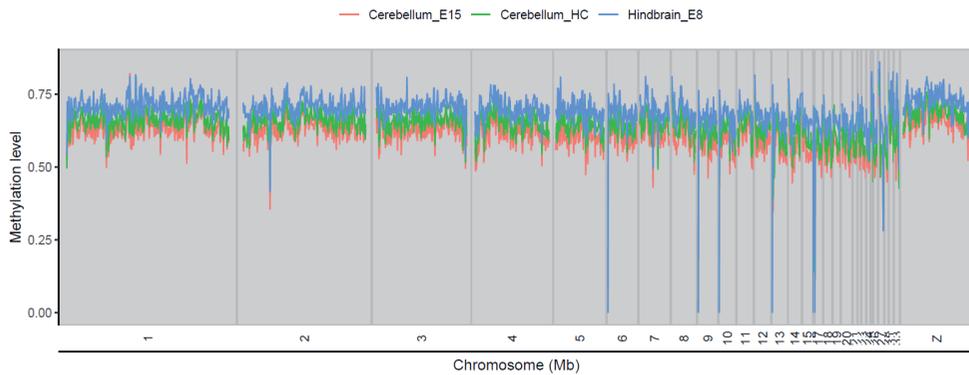
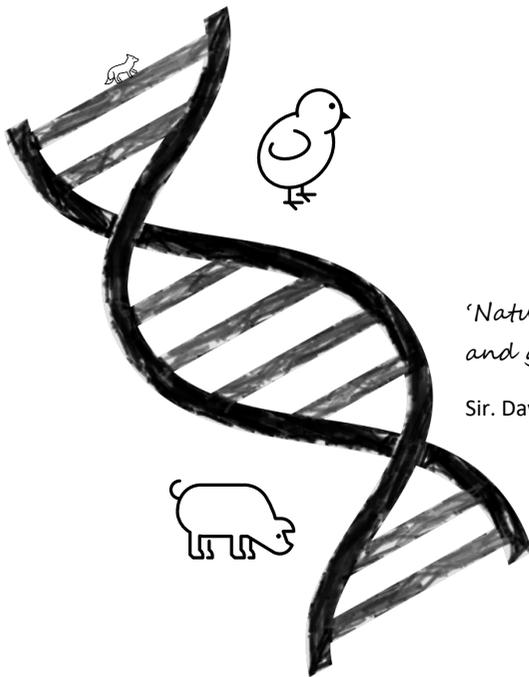


Fig. S8 Methylation distribution per chromosome in the brain during development. Results for all tissues are shown in **Data S12**.

Chapter 6

General Discussion



*'Nature is our biggest ally
and greatest inspiration'*

Sir. David Attenborough

Outline

Gene expression is regulated by a combination of epigenomic modifications acting at (genomic) functional regulatory elements. However, our understanding of the dynamic nature of gene expression regulation, especially during embryonic development, remains limited. Embryonic development is a complex and dynamic process that is essential in the formation of all multicellular organisms.

In this thesis, I investigated gene expression regulation through an exploration of the functional regulatory genome and epigenome. Specifically, I investigated gene regulation in two distinct ways: first, by a molecular characterization of two functionally relevant cell lines from pig and chicken, and second, by focusing on the role of DNA methylation in regulating fetal development in pig and chicken, utilizing innovative combinations of well-established methodologies.

In the first part of my thesis, I characterized the functional regulatory genome and epigenome of two cell lines in pig and chicken (**Chapter 2**), and in this discussion I will provide insight into the opportunities and limitations of utilising cell lines and organoids in future research, specifically for animal research.

I will then discuss the challenges in research to ensure analytical reproducibility, and the challenges to create and maintain pipelines ensuring reproducibility. Reproducibility of results is important in research, and it is often difficult to replicate analyses as described in literature. As we focussed on DNA methylation data, we describe our contributions to a pipeline for the analysis of methylation data, and the implementation thereof for the analysis of data used in this thesis (**Chapter 3**).

To understand the role of the epigenome and the functional regulatory genome during development, I investigated the methylome dynamics of seven key tissues in both pig (**Chapter 4**) and chicken (**Chapter 5**) during fetal development. By integrating RNA-seq expression data with methylation data, I gained further insight into the dynamic nature of the methylome during fetal development.

Subsequently, I will explore the potential of implementing functional genome data, such as the methylome maps generated in this thesis, for fine mapping QTL. I also discuss the integration of various data types for identification of regulatory genome and epigenomic modifications. This includes histone modifications, open chromatin, DNA methylation, and gene expression, which are utilised to identify epigenomic states and gain further insights into the developmental processes. Furthermore, I examine the potential of creating a reference of epigenomic states

6 General discussion

and regulatory elements for each tissue, as well as the application of functional genomic data in applied breeding and genomic prediction models. Finally, I consider advancements in technologies like single-cell sequencing, long-read sequencing, and their relevance in projects like FAANG and future animal breeding (**Figure 1**).

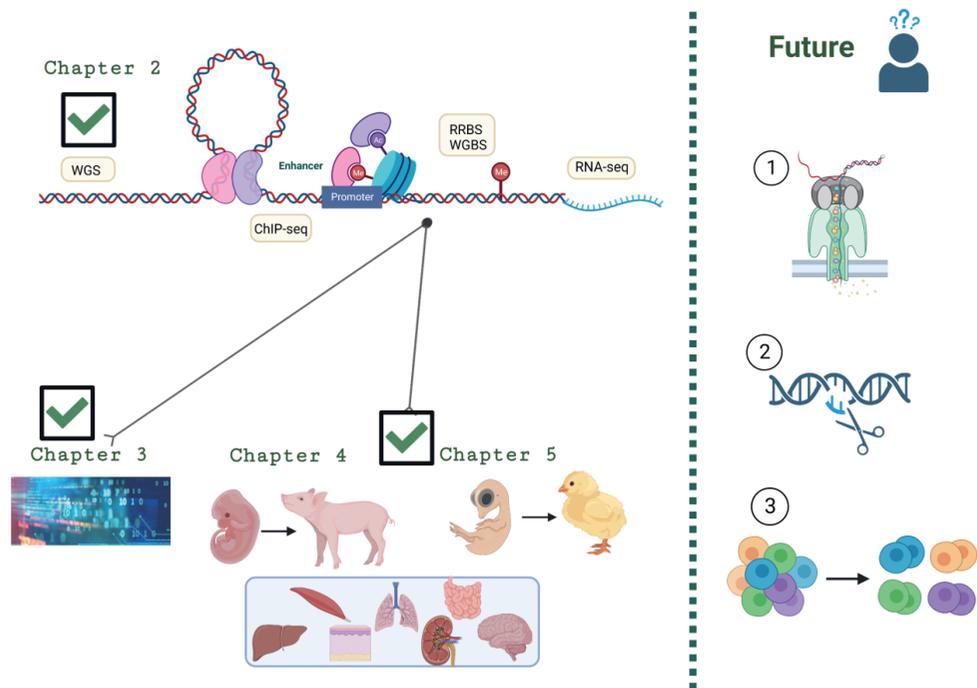


Figure 1 Research questions, achievements, and prospects of this thesis. The left panel illustrates the aims of the thesis with corresponding chapters. **Chapter 2** involves the molecular characterization of two cell lines in pig and chicken, utilizing various assays shown in yellow boxes. **Chapter 3** presents the development of a bioinformatics pipeline for DNA methylation data. **Chapters 4** and **5** explore the dynamics of the DNA methylome during pig and chicken fetal development in seven tissues. The right panel depicts future potential due to technological developments such as long-read sequencing with Oxford nanopore (1), epigenome-editing (2) and single cell sequencing (3). (Created with BioRender.com)

6.1 Cell lines in research: A cautionary tale and future considerations

In **Chapter 2** we provided a comprehensive analysis of the genome architecture of two cell lines commonly used in animal research (pig IPECJ2 and chicken SL-29). This integrative omics approach was a novel endeavour that aimed to deepen our understanding of the (epi)genetic makeup of cell lines and shed light on their usability as models for *in vivo* research. Advantages of cell lines include the ability to produce large quantities of cells for experiments and ease of manipulation (Verma, 2014; Verma et al., 2020). Importantly, cell lines provide an ethical alternative for experimental animals in research.

In human research, cell lines have been extensively utilized to investigate various diseases such as cancer (Chatterjee et al., 2018; Izquierdo-Torres et al., 2019) and rheumatoid arthritis (Kolarz and Majdan, 2017; Nakano et al., 2013), drug development (Belter et al., 2020; Chiappinelli et al., 2016; Heerboth et al., 2014; Pfister and Ashworth, 2017; Ruofß et al., 2019; Tomaselli et al., 2020) and exposure to toxins such as atrazine, a popular herbicide in the U.S (Ataei and Abdollahi, 2022; Lopez-Suarez et al., 2022; Sánchez et al., 2020). Primate cell lines have been instrumental in studying epigenomic patterns, evolution (García-Pérez et al., 2021; Mitalipov, 2006), and serving as a biomedical model for investigating diseases and their treatments (Juan et al., 2023; Rodriguez-Polo and Behr, 2022).

However, in farm animal research in species such as pig, chicken, cattle, and sheep, there have been a limited focus on utilizing cell lines for studying epigenomic alterations in response to environmental changes. Most studies using livestock cell lines focussed on disease susceptibility (Meekins et al., 2020), immune response (Laude and Gelfi, 1979; Leighton et al., 2015; Su et al., 2013), nutrition (Lallès, 2016) and biomedical modelling (Roura et al., 2016). In our research we utilised the pig IPECJ2 cell line: an intestinal epithelium cell line used for e.g. digestibility research, and the chicken SL-29 cell line: a fibroblast cell line which can be used for research related to immunity.

During the characterization of the pig IPECJ2 and chicken SL-29 cell lines, one of the main results we encountered is the aneuploidy of some chromosomes, despite these being untransformed cells. From this research (**Chapter 2**), it remains unclear whether the specific chromosomes are aneuploid from the establishment of the cell line or if aneuploidy has occurred after a certain number of passages. In a mice fibroblast cell line no aneuploidy is observed after three passages, however by passage seven this cell line exhibits a proliferation defect (Sheltzer and Amon, 2011). Through further investigation into the aneuploidy in the

IPECJ2 cell line (**Appendix A**), we observed that the IPECJ2 cell line exhibited aneuploidy in specific chromosomes and changes in aneuploidy related to number of passages. Chromosome 17 was found to be triploid from earlier passages (67 passages), while chromosome 16 was diploid at earlier passages (67 passages) and transitioned to triploid after prolonged culture (87-91 passages). Thus, the number of passages contributes to the rise of aneuploidy, as observed in mouse, and since the IPECJ2 cells in the **Appendix A** study were purchased and grown independently in different laboratories, it is likely that chromosome 17 is aneuploid from the cell stock.

An important distinction should be made between whole-organismal aneuploidy and somatic aneuploidy. Whole-organismal aneuploidy is a condition where the entire organism has an abnormal number of chromosomes. In such cases, every cell in the organism carries the same chromosomal abnormality, resulting in an overall imbalance of genetic material throughout the body. In contrast somatic aneuploidy is a condition where some cells in the organism have an abnormal number of chromosomes. Somatic aneuploidy has been found at low levels in healthy human tissues such as skin, brain, liver (Li and Zhu, 2022). This could provide a possible explanation for the observed aneuploidy in the cell lines investigated in **Chapter 2**.

On one hand, aneuploidy may provide a fitness advantage to cells due to alterations in gene dosage, resulting in changes in gene expression and protein levels. This could lead to a selective advantage, allowing cells to adapt better to their environment and proliferate more rapidly (Rosenkrantz and Carbone, 2017; Sheltzer and Amon, 2011). In the developing brain and embryonic neural cells, the percentage of aneuploidy is high and declines in the adult brain; this leads to the formation of neuronal networks and brain plasticity (Rosenkrantz and Carbone, 2017). On the other hand, aneuploidy can be detrimental to cell fitness by disrupting gene expression balance, interfering with critical cellular processes, and causing cell cycle defects or cell death (Ben-David and Amon, 2020; Rosenkrantz and Carbone, 2017; Sheltzer and Amon, 2011). In contrast, a study implementing single cell sequencing showed aneuploidy to be a rare occurrence in somatic tissues of mammals, and to have negative impacts on cell fitness (Knouse et al., 2014). However, their experiments involved a remarkably limited number of analysed cells, with as few as nine cells per experiment.

These findings serve as a cautionary tale when working with cell lines, as aneuploidy can introduce variability and affect the reproducibility of experimental results particularly for genome editing of genes located on aneuploidy chromosomes. Therefore, careful characterization and monitoring of cell lines are essential to ensure their genetic stability and

reliability for research purposes (Hynds et al., 2018; Sheltzer and Amon, 2011; Vcelar et al., 2018).

In addition to cell lines, organoids have emerged as a promising avenue of investigation. Organoids offer both functional and structural similarities to tissues and organs, surpassing the limitations of traditional cell lines (Clevers, 2016; Lancaster and Knoblich, 2014; van der Hee et al., 2020). There is an expectation that organoids will advance research, in particular biomedical and clinical research, providing a bridge between cell lines and *in vivo* studies (Silva-Pedrosa et al., 2023). Organoids provide new opportunities for studying fundamental cellular and gene activities. They also hold significant potential for translational research in areas such as diseases, toxicology, and cancer (Clevers, 2016), and in animal science (Kar et al., 2021). However there are still some limitations associated with organoid models, including incomplete differentiation, the inability to grow beyond a certain size and the cell organization remaining basic (Bhaduri et al., 2020; Yin et al., 2021; Zhao et al., 2022). Lastly, new technologies such as single cell sequencing will provide insights into cell heterogeneity and identify the compositions of cell populations. I will discuss this in further detail in section 6.6a.

Despite these limitations of cell lines and organoids, the advantages they offer are substantial. They provide a more ethical method for conducting research on animals and offer solutions to dilemmas such as small sample sizes in animal studies.

6.2 Ensuring reproducibility in research

In recent years the reproducibility of scientific research has been a topic of much debate (Allison et al., 2018; Botvinik-Nezer et al., 2020; Fanelli, 2018; Shiffrin et al., 2018; Stodden et al., 2018). A survey published in *Nature* revealed that 52% of researchers believe there is a 'significant crisis of reproducibility' (Baker, 2016). Surprising results were reported by Stodden et al. (2018) where corresponding authors of computational articles were emailed requesting data and code associated with the articles. The most intriguing responses obtained were as follows: 11% Contact another person, 11% Asked for a reason for providing code, 7% Refused to share code, 36% Shared code, 26% No responses. To address the reproducibility challenge problem, I propose the implementation of computational pipelines, and publicly available data as a possible solution. As part of the GENE-SWitCH project, we have implemented pipelines for all bioinformatic analyses. Furthermore, we have aimed to adhere to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) for both code and data produced. All data is openly accessible on the FAANG data portal under the Toronto agreement, which ensures that obtained datasets are published by the data owners before public dissemination.

A computational pipeline refers to a series of interconnected tasks or processes executed in a specific order to achieve a particular objective. By utilizing well-documented pipelines, such as the GSM-pipeline described in **Chapter 3**, the versions of software, tools, and parameters used can be transparently referenced (Grüning et al., 2018; Mangul et al., 2019; Wratten et al., 2021). The pipeline we developed in **Chapter 3** employs a workflow manager called nextflow (Di Tommaso et al., 2017), which uses domain-specific language (DSL). Additionally, the software used is containerized in Docker (Silver, 2017), ensuring that the versions of tools remain consistent and stable.

However, pipelines also present challenges, such as the need for updates to programming languages like DSL, Python, and R. If pipelines are not properly maintained, they can quickly become outdated. This necessitates researchers dedicated to both creating new pipelines and maintaining established ones, which in itself poses a challenge. Another challenge is the constant development of new tools and software. As an example, the **bio.tools registry** reports a staggering 17,000 entries of life science tools in 2020 (Marx, 2020). Therefore, it is important to consider whether pipelines should be continually developed and how to deal with the challenges associated with this (Cohen-Boulakia et al., 2017). Therefore, I conclude, pipelines offer utilities in providing data analysis reproducibility, however the time involved in development and maintenance thereof is a constraint. In cases where adherence to stringent pipeline standards and extensive review are required, pipelines might not yield substantial benefits due to prolonged timelines. However, for pipeline development within organizations lacking stringent requirements, simplified maintenance and quick development are feasible.

Based on this thesis and previous research, I propose several solutions to address the reproducibility problem. Firstly, journal policies should require structured and transparent documentation of methods and data (Allison et al., 2018; Baker, 2016; Christian et al., 2020). An example of a journal policy that promotes transparency is STAR methods (Tonzani and Fiorani, 2021). In conjunction with revised journal policies, the publication of all code used in research papers should be implemented. Secondly, while the utilization of pipelines offers a promising solution, it is essential to carefully consider the complexity and effort required for their creation and maintenance. Developing pipelines can be time-consuming, and some pipelines may become outdated quickly. Additionally, users need a comprehensive understanding of all steps involved in the pipeline for data processing to ensure correct data interpretation, effective problem-solving, and quality control. While concerns about development remain, pipelines present a promising solution to address the issue of reproducibility.

6.3 Epigenomic regulation of tissue differentiation during fetal development

6.3.1 The dynamic methylome in mammals and birds

The primary focus of this thesis is to characterise the DNA methylome of tissues related to complex traits such as meat production, feed efficiency, and egg production in two monogastric species, namely pig (**Chapter 4**) and chicken (**Chapter 5**). These species have been identified as the primary sources of meat worldwide according to a report by OECD-FAO in 2022 (OECD/FAO, 2022).

Our investigation began by exploring the dynamics of DNA methylation during fetal development in pig. DNA methylation has been studied extensively in mouse tissue development (He et al., 2020; Yizhar-Barnea et al., 2018) and in adult tissues of pig (Pan et al., 2021; Schachtschneider et al., 2015; Yang et al., 2021), cattle (Huang et al., 2014), mouse (Hon et al., 2013; H. Liu et al., 2021; Orozco et al., 2015), human (Brunner et al., 2009; Nakano et al., 2013; Schultz et al., 2015) and other primates (Blake et al., 2020).

Additionally, we explored the dynamics of the methylome in chickens during development. However, the available studies in birds, especially during development are limited (Derks et al., 2016; Laine et al., 2016; Li et al., 2011; Lindner et al., 2021; Pan et al., 2023; Ran et al., 2021; Watson et al., 2019). Several studies have investigated the mechanisms of DNA methylation and other epigenomic modifications in zebrafish, shedding light on their influence during various stages of development (Balasubramanian et al., 2019; Goll and Halpern, 2011).

This thesis is the first research characterizing the methylome in different tissues during fetal development in non-model species such as pig and chicken. The results have multiple important applications: (1) comparative analysis offering the possibility to gain insights into the genetic background of human development through biomedical models such as pig and chicken, (2) contributing to well-annotated reference genomes for FAANG and animal science research, and (3) enhancing the efficacy of genomic selection in the pig and poultry sectors.

a. Methylome of tissues

The liver exhibited intriguing differentiation patterns for methylation level in pigs compared to chicken, indicating differences in organogenesis between the two species. In pigs, the liver develops rapidly due to its role in early haematopoiesis caused by the short transient function

of the mammalian yolk sac in haematopoiesis (Ross and Boroviak, 2020). This is in contrast to birds, where the yolk sac maintains its hematopoietic function until hatching (Guedes et al., 2014; Wong and Uni, 2021). The developmental differences in the liver between these species may be attributed to the distinct amniotic environment in eutherian mammals, where the liver undergoes multiple waves of haematopoiesis before this role is eventually transferred to the bone marrow just before birth (Brunner et al., 2009; Farlik et al., 2016; Huse et al., 2015; Lewis et al., 2021; Ross and Boroviak, 2020; Waterland et al., 2009). From these results I have come to the conclusion that the liver matures later in birds (Guedes et al., 2014) compared to the development in the pig, human (Bonder et al., 2014; Huse et al., 2015) and mouse (He et al., 2020).

In terms of non-CpG methylation levels in brain, our study reports lower levels compared to previous research on adult stages in vertebrates (~0.2-8%) (de Mendoza et al., 2021; Derks et al., 2016; Guo et al., 2014, 2014; Jang et al., 2017; Lee et al., 2017; J.-H. Lee et al., 2020; H. Liu et al., 2021). This pattern is similar to fetal brain tissue in humans and mice, where non-CpG methylation levels is negligible, and an accumulation is observed during adulthood (Lister et al., 2013). The low non-CpG methylation levels observed in our study may reflect the developmental stages and section of the brain tissue analysed. Postnatal environmental stimuli and synaptic development contribute to brain maturation and plasticity during early adult stages in vertebrates (Bonfanti and Charvet, 2021, 2021; Coelho-Santos and Shih, 2020; Lister et al., 2013; Nakafuku and del Águila, 2020; Pizzorusso and Tognini, 2020). These findings shed light on the complex dynamics of brain development, specifically the role of CpG methylation during embryological development in brain and the assumed role of non-CpG methylation after birth.

Another tissue of interest is the small intestine, which exhibited dynamic methylation changes during both developmental transitions in pig (**Chapter 4**). In eutherian mammals like pig, newborns receive passive immunity via IgG from the colostrum of the immune competent mother. This transfer of intestinal macromolecules is prominent during the first 1-2 days after birth, followed by a decline in transfer (Weström et al., 2020). This environmental cue contributes to changes in methylation patterns in the intestine of pig (mammals) (Pinho and Maga, 2021; Rakoff-Nahoum et al., 2015; Yu et al., 2015). In contrast, non-mammalian vertebrates such as birds receive maternal macromolecule immunity from the yolk sac, which plays a crucial role in supplying passive immunity to the developing chick (Wong and Uni, 2021; Zhang and Wong, 2019).

Overall, the methylome profiles observed in this study provide insights into the regulation of tissue-specific differentiation and development. The distinct methylation dynamics in various tissues highlight their roles in orchestrating precise timing and the development of organs. These findings have implications for understanding species-specific physiological processes and may contribute to advancements in biomedical research, animal breeding, and genomic selection in the pig and poultry sectors. Conducting research in human development poses significant challenges due to ethical considerations and issues related to embryo quality, which limit the accuracy and scope of genome representation. Due to the abovementioned limitations, the results presented in this thesis hold valuable insights that could be extrapolated to humans. Moreover, given that certain complex traits are established during fetal development, the identification of genes and regulatory elements in animals holds promise for enhancing genomic selection in pig and chicken. Further details on this aspect are discussed in section 6.5.

b. Methylome of embryological layers

Gastrulation is a crucial stage of embryonic development during which three germ layers are formed, and subsequent organogenesis leads to the development of organs and tissues with specific functions (Kiecker et al., 2016; Muhr and Ackerman, 2023). In this study, our aim was to investigate the patterns of DNA methylation that distinguish tissues originating from the same embryological layer. While previous research in model organisms has identified genes involved in gastrulation, understanding of the epigenomic regulation still remains limited (He et al., 2020; Kiecker et al., 2016; Salehin et al., 2022; Xiang et al., 2020; Zorn and Wells, 2007). By utilizing gene enrichment analysis, we gained insights into the biological processes underlying the epigenomic regulation and gene expression of germ layers.

We concluded that the germ layer was not completely segregated at early developmental stages between brain and skin in chicken, as well as between muscle and kidney in pig. Both the skin and brain originate from the ectoderm germ layer, while the muscle and kidney derive from the mesoderm germ layer (Kiecker et al., 2016). However, we did not observe this pattern in other tissues, suggesting that the development of certain organs, such as liver and lung in both pig and chicken, occurs earlier than the first development stages sampled in this research (30dpf in pig and E8 in chicken). Furthermore, the use of 'bulk' tissues may limit the detection of similarities between tissues originating from the same germ layer (discussed further in section 6.5 a).

c. Limitations and improvements of the study

One limitation of our study is the absence of whole-genome DNA methylation results in post-natal tissue samples. This prevented us from investigating the complete DNA methylome, specifically UMR and LMR, during the development of tissues like the brain and small intestine in response to post-natal environmental stimuli. While publicly available RRBS data for certain tissues, such as cerebellum, ileum, and liver in adult pig and chicken (Pan et al., 2023, 2021), is accessible, future studies should utilise this data to explore targeted promoter regions of the methylome.

Additionally, it is important to note that our sample sizes for whole-genome bisulfite sequencing (WGBS) were small, with only one sample available, and for reduced representation bisulfite sequencing (RRBS), we had three samples. Smaller sample sizes decrease the statistical power of a study (Button et al., 2013; Hackshaw, 2008; Ioannidis, 2005; Lakens, 2022). However, increasing sample sizes in animal development and functional studies pose challenges due to the high costs of functional assays, and ethical concerns (Arifin and Zahiruddin, 2017; Bacchetti et al., 2005; Festing, 2018) regarding the number of animals sacrificed for research during fetal development (Bacchetti, 2013; Hackshaw, 2008). Efforts are being made to address these limitations, such as the European FAANG GEroNIMO project (<https://www.geronimo-h2020.eu/project>). This project utilizes a cost efficient method called GBS-MeDIP (Genotype by Sequencing – Methylated DNA Immunoprecipitation, Rezaei et al., 2022) to investigate DNA methylation in large populations of pigs and chickens. Additionally, in the GEroNIMO project, sampling of sperm, muscle biopsy, and reuse of samples collected for other projects are used, avoiding the need for animal sacrifice. However, when studying development, alternative sampling strategies are required, and this remains a challenging limitation to address.

Another limitation worth mentioning is the suboptimal gene annotations for both pig and chicken (Derks and Groenen, 2022), especially when compared to the well-established annotation for human. This may have contributed to the lack of significant enrichment findings and the limited identification of embryological germ layer similarities and changes between tissues (**Chapters 4 & 5**). Improving gene annotations in non-model species through projects like GENE-SWitCH significantly advances future studies in this field. ENSEMBL has incorporated tissue-specific developmental data from GENE-SWitCH for gene annotation of the pig and the three chicken genome assemblies (updated Red Jungle Fowl, new Broiler, and White Leghorn references) (Martin et al., 2023; <https://projects.ensembl.org/gene-switch/>).

Finally, our research utilized bulk tissue sequencing, which is a common method for investigating genome-wide methylation and transcriptomic variations. However, it should be noted that this approach provides an average methylation level across various cell types within a tissue, potentially obscuring the contribution of specific cell types with distinct functions. Single-cell sequencing is an emerging method that allows for the investigation of specific cell types, and its potential application is considered in a later section (6.6 a) of this thesis.

6.3.2 Integrative analysis to understand epigenome dynamics during development

The functional regulatory genome refers to the regions of the genome that contain regulatory elements that control gene expression. By integrating data from various functional assays that relate to regulation of gene expression, we can gain insights into the architecture of the functional regulatory genome and its associated epigenome. Studies conducted in human and mouse have demonstrated the potential of creating comprehensive epigenomic maps (Madhani et al., 2008; Roadmap Epigenomics Consortium et al., 2015; Satterlee et al., 2019; van der Velde et al., 2021; Yang et al., 2021).

Projects like EpiMap (Boix et al., 2021) and the Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010) showcase the ability to annotate epigenomic maps across many tissues, cells and different developmental stages (embryonic to adult). These maps have been associated with specific diseases, organs, and cell types. In the context of animal breeding, the potential of having reference epigenomic maps across tissues related to important traits can be envisioned. This would enable a deeper understanding of the epigenomic basis of traits and facilitate targeted breeding strategies to improve desired traits in animals (discussed in further detail 6.5).

Such epigenomic maps can provide valuable resources for researchers, allowing them to explore the regulatory landscape of the genome and investigate how specific epigenomic modifications are associated with gene expression and phenotype. These maps could potentially aid in identifying key regulatory elements, deciphering the mechanisms underlying gene regulation, and even predicting the effects of genetic and environmental factors on the epigenome and gene expression patterns. Examples for this have been shown in human (Roadmap Epigenomics Consortium et al., 2015; Satterlee et al., 2019; Wang et al., 2021), mouse (Orozco et al., 2015; van der Velde et al., 2021), pig (Kern et al., 2021; Pan et al., 2021; Yang et al., 2021) and chicken (Kern et al., 2021; Pan et al., 2023).

a. Case study: Epigenome map in chicken SL-29 cell line

6 General discussion

I explored the potential for creating an epigenomic map using omics data available from the chicken SL-29 cell line that I analysed in **Chapter 2**. I demonstrate how the incorporation of DNA methylation data can be used to identify regions and sites of repression and activation, when combined with ChIP-seq data for histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3) to pinpoint promoters, enhancers, and gene silencers. Methylation classes were defined by MethylSeekR (Burger et al., 2013) as described in **Chapters 4 & 5**. In **Figure 2** I present a method for integrating various functional omics data and annotating epigenomic maps within the genome.

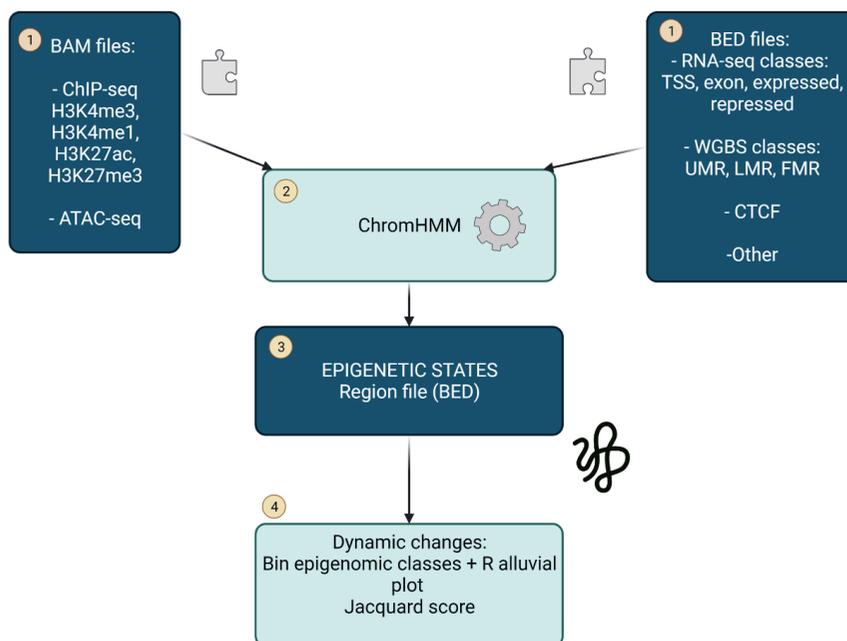
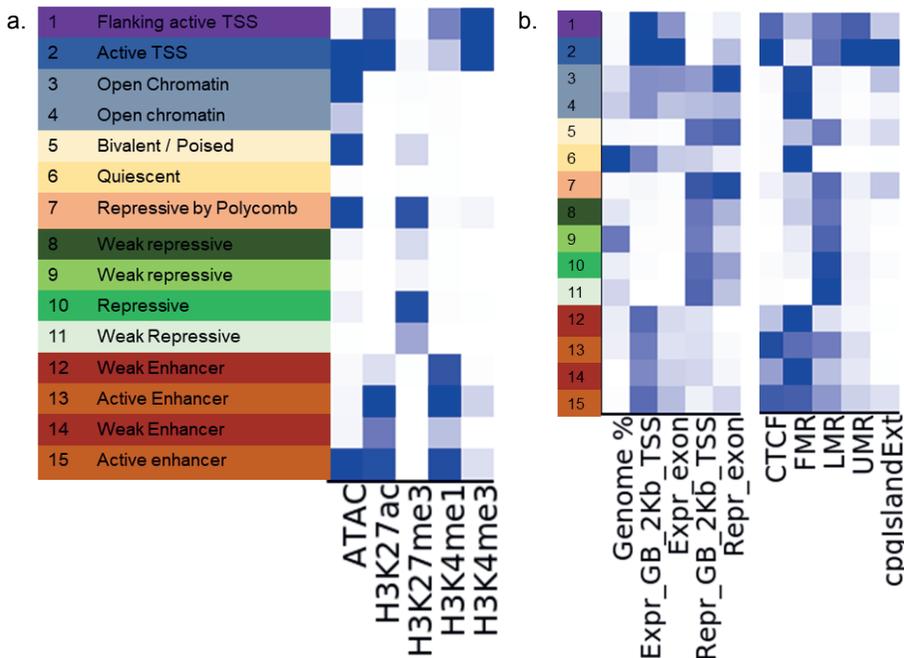


Figure 2 Schematic representation for integrating different functional data sets such as ChIP-seq and ATAC-seq together with DNA methylation, expression data and CTCF. In this diagram I suggest using the ChromHMM tool (2) (Ernst and Kellis, 2017), which implements a hidden Markov model to (a) detect the presence or absence of marks and (b) prove a level of the presence of marks. From this analysis a region file (3) is provided from which regions in the genome are annotated as e.g. a poised enhancer, active transcription start site (TSS). Investigating the dynamic nature of development or similar tissues can be achieved by using an alluvial plot for showing changes in epigenomic states (4). (Created with BioRender.com)

I applied the aforementioned method (**Figure 2**) to the functional assays we generated (ATAC-seq, WGBS, CTCF, RNA-seq and ChIP-seq of four histone marks: H3K4me1, H3K4me3, H3K27ac, H3K27me3) for the chicken SL-29 cell line (**Chapter 2**). This allowed me to define 15 epigenomic states, which are shown in **Figure 3**. It is important to note that the defined epigenomic states in this study build upon previous annotations from different resources, such as Gorkin et al., (2020), Roadmap Epigenomics Consortium et al., (2015), and van der Velde et al., (2021).

The defined model consisted of nine active states and six repressive states that showed distinct levels of DNA methylation, DNA accessibility, and regulatory binding. The active states (associated with expressed genes) consist of a flanking active TSS, open chromatin and enhancer states. Repressive states (associated with unexpressed genes) include weak repressive, quiescent and the unique poised/bivalent and repressive by polycomb states. For example, the active TSS state has an enrichment (dark blue) of ATAC-seq, H3K27ac, H3K4me3, and lower enrichment of H3K4me1 (**Figure 3a**). Additionally, there is a strong enrichment of expressed genes and exons, UMR, LMR, and CpG islands for this state (**Figure 3b**). UMR is typically indicating promoter regions and LMR enhancer regions. By incorporating the methylation states into the model, we gain an additional layer of information that contributes to the annotation of the epigenome (**Figure 3b**). Furthermore, as previously shown (**Chapters**



4 and 5), methylation has an inhibitory effect on gene expression. Therefore, FMR is indicative of repressed genes and UMR/LMR indicative of expressed genes (**Figure 3b**).

Figure 3 Fifteen epigenomic states identified by ChromHMM implementing functional omics data from the chicken SL-29 cell line. *a. Histone mark probabilities for each epigenomic state. b. Genome coverage and enrichment of the categories in overlapping genomic features such as expressed, repressed within the gene-body and 2 kb upstream of the TSS, expressed and repressed exons, CTCF, classes of methylation (FMR, LMR, UMR) and CpG islands.*

The integration of multiple omics datasets will enhance our understanding of the complex interplay between epigenomic modifications and gene regulation in the chicken genome. By understanding the epigenomic states associated with specific histone marks, DNA methylation patterns, and regulatory elements, researchers can gain a deeper understanding of the functional genomic architecture and the regulatory mechanisms underlying gene expression in this cell line.

6.4 Implementation of functional regulatory genome maps for fine mapping

From the 1990's the potential of integrating genomic information with phenotypic records for improved breeding strategies began to emerge (Meuwissen et al., 2016). Genome-wide association studies (GWAS) have gained popularity in livestock breeding for mapping quantitative trait loci (QTL) associated with economically important traits such as muscle composition, growth, and methane emissions (Sharma et al., 2015). The Animal QTL database (<https://www.animalgenome.org/cgi-bin/QTLdb/index>) serves as comprehensive resource capturing trait mapping information in livestock species, with the largest number of QTL identified for cattle (193,898), chicken (18,411) and pig (48,844) (accessed May 2023). However, a major challenge lies in the fact that the majority (>90%) of identified QTL are located in non-coding regions of the genome, making it difficult to decipher the underlying mechanisms regulating gene expression and their influence on traits (Orozco, 2022; Schaid et al., 2018; Weikard et al., 2017). To address this, fine mapping approaches implementing functional genomics data provide a way to unravel the mechanisms underlying important traits (Cano-Gamez and Trynka, 2020; Orozco, 2022; Weikard et al., 2017). Expression QTL studies can be used to investigate the regulation of gene expression and has been one of the first methods to investigate gene regulation. In human studies, the incorporation of functional data

such as open chromatin information (caQTL) and histone modifications (hQTL) alongside QTL has been reported by Orozco, 2022. This directly incorporates regulatory elements for investigating the relationship between gene regulation and traits.

Functional maps generated by the GENE-SWitCH project, and the methylome maps presented in this thesis offer valuable resources for fine mapping QTL identified through GWAS. By utilizing this functional data, valuable insights can be gained into the regulation of gene expression that influences specific traits. For instance, QTL located in enhancer, promoter, or open chromatin regions play a critical role in regulating genes.

6.5 Implementing functional genomics into applied animal breeding

The integration of genomic information with phenotypic records has proven valuable for improving animal breeding (Meuwissen et al., 2016), but the full potential of functional maps in this context is still being explored. Currently, QTL are incorporated into genomic breeding models, and researchers are currently investigating how functional maps can enhance animal breeding outcomes (Derks and Groenen, 2022; Ernst and Steibel, 2013; Johnsson, 2023). Ongoing investigations study the integration of functional layers, such as the DNA methylation changes identified in this thesis. A first example of using these DNA methylation changes showed only marginal improvements in genomic prediction accuracy (Mollandin, 2022).

Another strategy within the GENE-SWitCH project, explored predicting gene expression levels from genome sequence and methylation data using a genomic prediction model. (B. Perez, 2023, personal communication). Ten genes were chosen based on their association with tissues as identified in an eQTL study based on whole-genome sequence data, and gene expression in liver, duodenum, and muscle tissue from 300 slaughter pigs (100 Duroc, 100 Landrace and 100 Large White) (Crespo-Piazuelo et al., 2022). A set of 65,486 single nucleotide polymorphisms (SNPs) were pre-selected within regulatory regions (2 Kb upstream and 0.2 Kb downstream of these ten genes TSS). Corresponding methylation states at 30 dpf and NB (**Chapter 4**) in the liver, duodenum and muscle were used to differentially weigh SNPs in constructing the genomic relationship matrix. This was used in GBLUP (Genomic best linear unbiased prediction) either including the methylation scores or without it. Sites in FMR were given a weight of 0 as these were assumed not to contribute to the expression of genes.

Preliminary results of this study are shown in **Figure 4**. The first analysis involved investigating the accuracy of predicted gene expression in muscle within a breed (**Figure 4a**). We observed higher accuracy with methylation data from muscle than other tissues, and this increase in

accuracy differed between the ten genes. This indicates the benefit of methylation in the prediction models to be highly tissue and gene dependent. Interestingly, a few genes exhibited lower genomic prediction for methylation scores compared to unweighted models.

A second analysis involved the predictive accuracy of gene expression from SNPs across different breeds. In this instance (**Figure 4b**), the Duroc and Landrace breeds were used as the reference population and the validation was conducted in the Large White breed. Notably, these results show a range of patterns across the tissues and various genes. Across all three tissues, certain genes show a negative accuracy in the genomic prediction of gene expression. These results suggest a benefit, albeit small, of methylation scores to genomic prediction, more specifically within a breed in comparison to across breed.

Drawing insights from these findings, the following considerations emerge. These examples employed a single layer of functional annotation, namely methylation, which also displays a degree of tissue specificity in genomic prediction. Consequently, this prompts a subsequent question regarding the potential advantages of incorporating multiple layers of functional annotation, such as ChIP-seq data on histone modifications, chromatin states, and subsequently integrating these multiple functional layers into a comprehensive scoring system. Finally, it may be that tissue and or breed specific functional layers must be made before they are suitable to be used in genomic prediction.

Combined Annotation-Dependent Depletion (CADD) is an integrative annotation that combines layers of genomic annotation information from various sources, such as evolutionary constraints and gene model annotations, to score and rank variants (Kircher et al., 2014; Rentzsch et al., 2019). CADD has been developed for several livestock species including pig (pCADD), and their utility for trait variant classification has been demonstrated (Derks et al., 2021). Bink et al., 2022 found that scoring SNPs based on CADD scores resulted in slightly higher genomic prediction accuracy for three traits in layer chickens. Likewise, another related method, Functional-And-Evolutionary Trait Heritability (FAETH), has shown promise in improving genomic prediction accuracy in cattle, (Xiang et al., 2019).

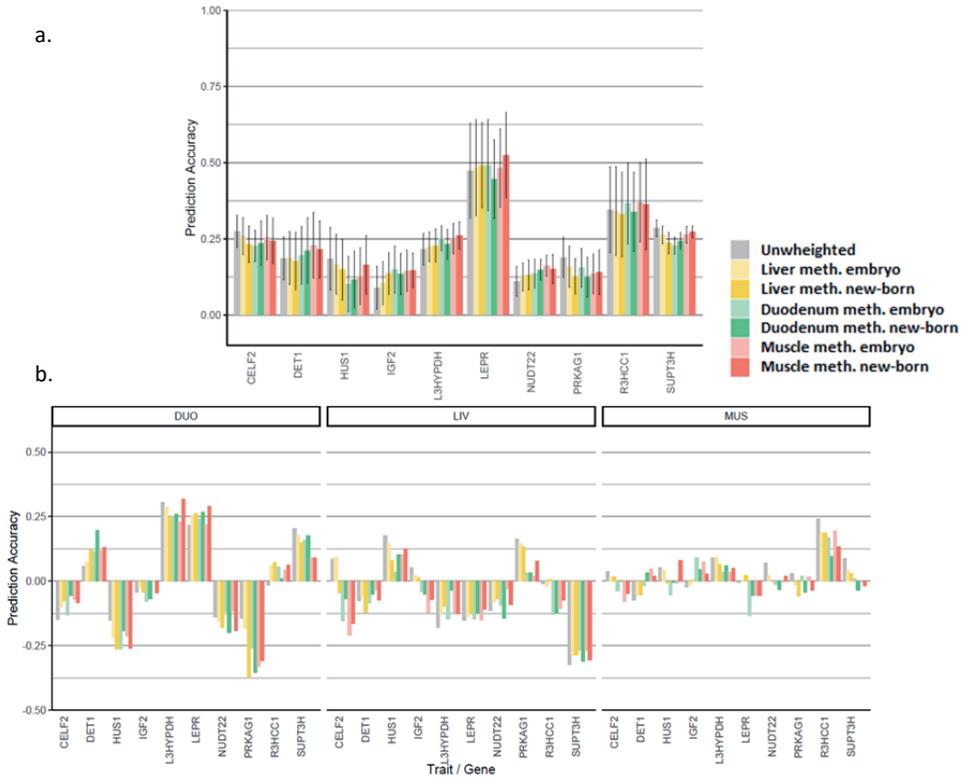


Figure 4 Utilization of methylation data to improve prediction of gene expression for ten selected genes in two different scenarios. *a.* Accuracy of genomic prediction of gene expression in ten genes in liver with the methylation included as a score within a breed. Standard error bars are included. *b.* Accuracy of genomic prediction of gene expression in ten genes in liver, muscle and duodenum. This method investigates prediction accuracy across breeds with the Duroc and landrace used in the reference population and Large White used as the validation. (B. Perez, 2023, personal communication).

Interpreting functional data in the context of genomic models remains a challenge due to its complexity. Ideally functional maps would be updated regularly, and scores such as CADD or FAETH updated based on additional functional data. The most significant variants identified by these methods could then be incorporated into genomic prediction models. These developments in utilizing functional data for applied animal breeding present exciting opportunities.

6.6 Future potential

During the last 20 years the field of genomics has seen many technological developments, and exciting opportunities for new research have arisen such as long-read sequencing (Amarasinghe et al., 2020; Pollard et al., 2018), single cell sequencing (Nawy, 2014), implementation of pan-genomes (Golicz et al., 2020), and (epi)genome editing (Kungulovski and Jeltsch, 2016; Nakamura et al., 2021). Since the FAANG project was established in 2014 (Andersson et al., 2015), there have been subsequent projects with the aim of improving our knowledge of the functional regulatory annotation in different livestock species, such as cattle (Moreira et al., 2022), chicken and pig (Acloque et al., 2022) and in aquaculture (Baudement et al., 2022). The first goal of the FAANG consortium was to use high-depth functional assays on an individual level, as described in this thesis for pig and chicken. Due to recent advancements in genomics, a promising new era of research for FAANG has emerged (Clark et al., 2020) and in this section I will discuss the impact of new technologies on the findings presented in this thesis and propose potential improvements to address its limitations.

a. Single cell sequencing

In 2013, a cutting edge molecular biology technique called single cell sequencing was developed (Nawy, 2014). This technique allows researchers to investigate genetic information of individual cells and provides a high-resolution view of the cellular heterogeneity. It facilitates the study of different cell types that constitute a tissue or organ (Schwartzman and Tanay, 2015).

Traditional methods of genomic and transcriptomic sequencing involve measuring a mixture of cells (called a “bulk”) comprising a tissue or organ. “Bulk” sequencing analyses, as employed in this thesis, enable the measurement of the average methylation levels in a population of cells. For example, RNA-sequencing of muscle tissue (Rubenstein et al., 2020) produces transcripts from various types of cells such as smooth muscle cells, myeloid cells, satellite cells, and endothelial cells, which are analysed together.

Implementing single-cell sequencing allows the researcher to characterize individual cells, leading to an improved understanding of cellular functions and interactions. Additionally, this can enhance current research by facilitating a better comprehension of organ differentiation from germ layers, as seen in the reported examples in mice (Argelaguet et al., 2019; Chan et al., 2019; He et al., 2020; Liu et al., 2021), zebrafish (Farrell et al., 2018) and frog (Briggs et al., 2018). Clark et al. (2022) reported how DNA methylation dynamics in specific mice cells

can regulate embryogenesis, using single cell sequencing. These examples demonstrate the importance of utilising single-cell epigenomic sequencing for an improved functional annotation of developmental trajectories, germ layer differentiation and cell fates in pig and chicken as well. Clark et al. (2020) propose creating single-cell atlases for key tissues in livestock species, which is one of the future priorities of FAANG (5-10 years). In pig research utilising single cell transcriptomic analysis have included identification of different cell types for cerebral cortex and hypothalamus (Chen et al., 2023; Zhu et al., 2021), skeletal muscle development (Cai et al., 2023; Chen et al., 2023), and embryo implantation site (Tian et al., 2022), while for chicken it included limb patterning at embryonic development (Feregrino et al., 2019), and heart development (Mantri et al., 2021). Furthermore, this technology can be used to determine the cell composition of organoids, thus quantifying the similarity of organoid cells to primary tissue counterparts and reveal cell-specific reactions to environmental variables and disease conditions. The research reported in this thesis provides the foundation for future implementation of such technologies.

b. Long-read sequencing

Long-read sequencing technologies enable sequencing longer DNA fragments, making this useful for studying repetitive regions of the genome, resolving complex structural variations, improving de novo assemblies and identifying new transcripts (Amarasinghe et al., 2020). Short read sequencing technologies typically provide up to 300 - 800 bases (Sanger and Illumina sequencing technologies), whereas long-read sequences can generate reads more than 10 kb. An additional advantage of long-read sequencing is the ability to detect DNA methylation from the electrical readout of the instrument, and thus avoiding treatments with e.g. bisulfite (Gouil and Keniry, 2019; Y. Liu et al., 2021; Simpson et al., 2017). Better characterization of DNA methylation patterns will be provided, due to longer sequencing reads in comparison to short reads. This will create an opportunity to perform both methylation analysis and genomic analysis, which provides the opportunity to investigate different variants and modifications on the DNA.

Haplotypes represent specific combinations of alleles located closely together on a chromosome, while allele-specific methylation refers to differential methylation of the two alleles. This phenomenon can potentially influence gene expression patterns and is often associated with genes exhibiting allele-specific expression (Zhang et al., 2009). Understanding haplotypes and allele-specific methylation play an important role in deciphering genetic variation, epigenomic regulation of biological processes and traits, and the inheritance of (un)favourable alleles from parents. Long-read sequencing facilitates the investigation of

haplotype-dependent allele-specific methylation (Akbari et al., 2022, 2021; Gigante et al., 2019), enabling the identification of specific haplotypes associated with important traits and serving as molecular markers in breeding. This aspect of epigenomics contributes to our understanding of genetic regulation and its implementation to select desirable traits, ultimately leading to improved livestock and animal performance. (Triantaphyllopoulos et al., 2016).

c. (Epi)genetic editing

Gene editing, a well-established method, presents ethically complex considerations, particularly in the context of livestock breeding. In academic research, the implementation of gene editing poses fewer ethical concerns, especially when testing variations in cell lines or organoids. The emergence of (epi)genetic editing technologies, such as CRISPR-based (epi)genetic editors, has empowered researchers to selectively modify specific (epi)genetic marks at genomic loci (K. Lee et al., 2020). Kang et al., 2019 demonstrated the value of epigenome editing to confirm the effects of epigenomic mechanisms on gene expression in mice. Moreover, in the field of aquaculture, CRISPR-Cas9 gene editing has already proven successful in species like Atlantic salmon, finfish, and molluscs (Clark et al., 2020; Houston et al., 2020).

Epigenome editing operates by modifying the chromatin state, without altering the DNA sequence (Kungulovski and Jeltsch, 2016). This technique specifically offers the advantage of modifying mechanisms that regulate gene expression, providing an enhanced understanding of the functional roles of epigenomic modifications through direct interrogation (Kungulovski and Jeltsch, 2016; Nakamura et al., 2021). The construction of functional maps, as demonstrated in this thesis, that depict the genome architecture of various tissues and cells can significantly improve the accuracy of epigenomic gene editing by accounting for potential pleiotropic effects. However, despite its potential, the implementation of this technology remains constrained by ethical considerations, challenges and unknown effects (Bishop and Van Eenennaam, 2020; K. Lee et al., 2020; Raza et al., 2022).

d. Pan-genomes

A pan-genome refers to the collective genomic diversity of a population or a species, encompassing more than just the reference genome, which is a single representation of a haplotype for that species. The emergence of the pan-genomes concept stems from the recognition that there is significant genetic variability among individuals of the same species (Golicz et al., 2020). Traditional genome sequencing methods primarily focus on generating a

single reference genome that represents a consensus or average genetic makeup of the species. However, this approach fails to capture the full extent of genetic diversity present in populations.

More recently, the utilization of pan-genomes has gained momentum in various vertebrate species such as humans (Sherman et al., 2019; Sherman and Salzberg, 2020), and pigs (Derks et al., 2022). In the context of research presented in this thesis, the utilization of pan-genomes as a 'reference' genome during analysis enhances the accuracy of short-read mapping as compared to using a single reference, enabling the detection of variants specific to different breeds which play a role in regulating traits established during development. As discussed in this thesis, a significant portion of phenotypic variation can be attributed to genomic regions outside of genes, implying that changes in gene regulation may underlie many crucial traits. Consequently, it can be postulated that, in the future, the adoption of a regulatory build for each pan-genome will become customary, facilitating more precise variant detection, and enabling a detailed elucidation of the underlying biological regulatory mechanisms.

References

- Acloque, H., Harrison, P. w., Lakhali, W., Martin, F., Archibald, A. I., Beinat, M., Davey, M., Djebali, S., Foissac, S., Guizard, S., Guyomar, C., Kuo, R., Kurylo, C., Madsen, O., Miedzinska, K., Mongellaz, M., Smith, J., Smith, J., Sokolov, A., de Vos, J., Giuffra, E., Watson, M., 2022. 550. Extensive functional genomics information from early developmental time points for pig and chicken, in: Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). Wageningen Academic Publishers, pp. 2281–2284. https://doi.org/10.3920/978-90-8686-940-4_550
- Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M.A., Hirst, M., Jones, S.J., 2022. Genome-wide detection of imprinted differentially methylated regions using nanopore sequencing. *eLife* 11, e77898. <https://doi.org/10.7554/eLife.77898>
- Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M.A., Hirst, M., Jones, S.J.M., 2021. Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. *Genome Biology* 22, 68. <https://doi.org/10.1186/s13059-021-02283-5>
- Allison, D.B., Shiffrin, R.M., Stodden, V., 2018. Reproducibility of research: Issues and proposed remedies. *Proceedings of the National Academy of Sciences* 115, 2561–2562. <https://doi.org/10.1073/pnas.1802324115>
- Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21, 30. <https://doi.org/10.1186/s13059-020-1935-5>
- Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., Dalrymple, B.P., Elisk, C.G., Foissac, S., Giuffra, E., Groenen, M.A., Hayes, B.J., Huang, L.S., Khatib, H., Kijas, J.W., Kim, H., Lunney, J.K., McCarthy, F.M., McEwan, J.C., Moore, S., Nanduri, B., Notredame, C., Palti, Y., Plastow, G.S., Reecy, J.M., Rohrer, G.A., Sarropoulou, E., Schmidt, C.J., Silverstein, J., Tellam, R.L., Tixier-Boichard, M., Tosser-Klopp, G., Tuggle, C.K., Vilkki, J., White, S.N., Zhao, S., Zhou, H., The FAANG Consortium, 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology* 16, 57. <https://doi.org/10.1186/s13059-015-0622-4>
- Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C.W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P.J., Kelsey, G., Dean, W., Nichols, J., Stegle, O., Marioni, J.C., Reik, W., 2019. Multi-omics profiling of mouse gastrulation at single cell resolution. *Nature* 576, 487–491. <https://doi.org/10.1038/s41586-019-1825-8>

- Arifin, W.N., Zahiruddin, W.M., 2017. Sample Size Calculation in Animal Studies Using Resource Equation Approach. *Malays J Med Sci* 24, 101–105. <https://doi.org/10.21315/mjms2017.24.5.11>
- Ataei, M., Abdollahi, M., 2022. A systematic review of mechanistic studies on the relationship between pesticide exposure and cancer induction. *Toxicology and Applied Pharmacology* 456, 116280. <https://doi.org/10.1016/j.taap.2022.116280>
- Bacchetti, P., 2013. Small sample size is not the real problem. *Nat Rev Neurosci* 14, 585–585. <https://doi.org/10.1038/nrn3475-c3>
- Bacchetti, P., Wolf, L.E., Segal, M.R., McCulloch, C.E., 2005. Ethics and Sample Size. *American Journal of Epidemiology* 161, 105–110. <https://doi.org/10.1093/aje/kwi014>
- Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454. <https://doi.org/10.1038/533452a>
- Balasubramanian, S., Raghunath, A., Perumal, E., 2019. Role of epigenetics in zebrafish development. *Gene* 718, 144049. <https://doi.org/10.1016/j.gene.2019.144049>
- Baudement, M.-O., Gillard, G.B., Podgorniak, T., Grnvoid, L., Lien, S., Kent, M.P., 2022. 546. European project AQUA-FAANG: the epigenetic landscape of the Atlantic Salmon; focus on liver tissue, in: *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*. Wageningen Academic Publishers, pp. 2265–2268. https://doi.org/10.3920/978-90-8686-940-4_546
- Belter, A., Barciszewski, J., Barciszewska, A.-M., 2020. Revealing the epigenetic effect of temozolomide on glioblastoma cell lines in therapeutic conditions. *PLOS ONE* 15, e0229534. <https://doi.org/10.1371/journal.pone.0229534>
- Ben-David, U., Amon, A., 2020. Context is everything: aneuploidy in cancer. *Nat Rev Genet* 21, 44–62. <https://doi.org/10.1038/s41576-019-0171-x>
- Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A., 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28, 1045–1048. <https://doi.org/10.1038/nbt1010-1045>
- Bhaduri, A., Andrews, M.G., Kriegstein, A.R., Nowakowski, T.J., 2020. Are Organoids Ready for Prime Time? *Cell Stem Cell* 27, 361–365. <https://doi.org/10.1016/j.stem.2020.08.013>
- Bishop, T.F., Van Eenennaam, A.L., 2020. Genome editing approaches to augment livestock breeding programs. *Journal of Experimental Biology* 223, jeb207159. <https://doi.org/10.1242/jeb.207159>

6 General discussion

- Blake, L.E., Roux, J., Hernando-Herraez, I., Banovich, N.E., Perez, R.G., Hsiao, C.J., Eres, I., Cuevas, C., Marques-Bonet, T., Gilad, Y., 2020. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Res* 30, 250–262. <https://doi.org/10.1101/gr.254904.119>
- Boix, C.A., James, B.T., Park, Y.P., Meuleman, W., Kellis, M., 2021. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 590, 300–307. <https://doi.org/10.1038/s41586-020-03145-z>
- Bonder, M.J., Kasela, S., Kals, M., Tamm, R., Lökk, K., Barragan, I., Buurman, W.A., Deelen, P., Greve, J.-W., Ivanov, M., Rensen, S.S., van Vliet-Ostaptchouk, J.V., Wolfs, M.G., Fu, J., Hofker, M.H., Wijmenga, C., Zhernakova, A., Ingelman-Sundberg, M., Franke, L., Milani, L., 2014. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* 15, 860. <https://doi.org/10.1186/1471-2164-15-860>
- Bonfanti, L., Charvet, C.J., 2021. Brain Plasticity in Humans and Model Systems: Advances, Challenges, and Future Directions. *International Journal of Molecular Sciences* 22, 9358. <https://doi.org/10.3390/ijms22179358>
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., Avesani, P., Baczkowski, B.M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R.G., Berkers, R.M.W.J., Bhanji, J.P., Biswal, B.B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K.L., Bowring, A., Braem, S., Brooks, H.R., Brudner, E.G., Calderon, C.B., Camilleri, J.A., Castellon, J.J., Cecchetti, L., Cieslik, E.C., Cole, Z.J., Collignon, O., Cox, R.W., Cunningham, W.A., Czoschke, S., Dadi, K., Davis, C.P., Luca, A.D., Delgado, M.R., Demetriou, L., Dennison, J.B., Di, X., Dickie, E.W., Dobryakova, E., Donnat, C.L., Dukart, J., Duncan, N.W., Durnez, J., Eed, A., Eickhoff, S.B., Erhart, A., Fontanesi, L., Fricke, G.M., Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J.J., Golowin, S.A.E., González-García, C., Gorgolewski, K.J., Grady, C.L., Green, M.A., Guassi Moreira, J.F., Guest, O., Hakimi, S., Hamilton, J.P., Hancock, R., Handjaras, G., Harry, B.B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.-P., Huettel, S.A., Hughes, M.E., Iacovella, V., Iordan, A.D., Isager, P.M., Isik, A.I., Jahn, A., Johnson, M.R., Johnstone, T., Joseph, M.J.E., Juliano, A.C., Kable, J.W., Kassinosopoulos, M., Koba, C., Kong, X.-Z., Kosciak, T.R., Kucukboyaci, N.E., Kuhl, B.A., Kupek, S., Laird, A.R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M.Y.C., Lim, P.C., Lintz, E.N., Liphardt, S.W., Losecaat Vermeer, A.B., Love, B.C., Mack, M.L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J.T., Melero, H., Méndez Leal, A.S., Meyer, B., Meyer, K.N., Mihai, G., Mitsis, G.D., Moll, J., Nielson, D.M., Nilsson, G., Notter, M.P., Olivetti, E., Onicas, A.I., Papale, P., Patil, K.R., Peelle, J.E., Pérez, A., Pischedda, D., Poline, J.-B., Prystauka, Y., Ray, S., Reuter-Lorenz, P.A., Reynolds, R.C., Ricciardi, E., Rieck, J.R., Rodriguez-Thompson, A.M., Romyn, A., Salo, T., Samanez-Larkin, G.R., Sanz-Morales, E., Schlichting, M.L., Schultz, D.H., Shen, Q., Sheridan, M.A., Silvers, J.A., Skagerlund, K., Smith, A., Smith, D.V., Sokol-Hessner, P., Steinkamp,

- S.R., Tashjian, S.M., Thirion, B., Thorp, J.N., Tinghög, G., Tisdall, L., Tompson, S.H., Toro-Serey, C., Torre Tresols, J.J., Tozzi, L., Truong, V., Turella, L., van 't Veer, A.E., Verguts, T., Vettel, J.M., Vijayarajah, S., Vo, K., Wall, M.B., Weeda, W.D., Weis, S., White, D.J., Wisniewski, D., Xifra-Porxas, A., Yearling, E.A., Yoon, S., Yuan, R., Yuen, K.S.L., Zhang, L., Zhang, X., Zosky, J.E., Nichols, T.E., Poldrack, R.A., Schonberg, T., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 84–88. <https://doi.org/10.1038/s41586-020-2314-9>
- Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., Klein, A.M., 2018. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360, eaar5780. <https://doi.org/10.1126/science.aar5780>
- Brunner, A.L., Johnson, D.S., Kim, S.W., Valouev, A., Reddy, T.E., Neff, N.F., Anton, E., Medina, C., Nguyen, L., Chiao, E., Oyolu, C.B., Schroth, G.P., Absher, D.M., Baker, J.C., Myers, R.M., 2009. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* 19, 1044–1056. <https://doi.org/10.1101/gr.088773.108>
- Burger, L., Gaidatzis, D., Schübeler, D., Stadler, M.B., 2013. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res* 41, e155. <https://doi.org/10.1093/nar/gkt599>
- Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14, 365–376. <https://doi.org/10.1038/nrn3475>
- Cai, S., Hu, B., Wang, X., Liu, T., Lin, Z., Tong, X., Xu, R., Chen, M., Duo, T., Zhu, Q., Liang, Z., Li, E., Chen, Y., Li, J., Liu, X., Mo, D., 2023. Integrative single-cell RNA-seq and ATAC-seq analysis of myogenic differentiation in pig. *BMC Biology* 21, 19. <https://doi.org/10.1186/s12915-023-01519-z>
- Cano-Gamez, E., Trynka, G., 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* 11.
- Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G., Khodaverdian, A., Yosef, N., Meissner, A., Weissman, J.S., 2019. Molecular recording of mammalian embryogenesis. *Nature* 570, 77–82. <https://doi.org/10.1038/s41586-019-1184-5>
- Chatterjee, A., Rodger, E.J., Eccles, M.R., 2018. Epigenetic drivers of tumourigenesis and cancer metastasis. *Seminars in Cancer Biology, Epigenetics in cancer* 51, 149–159. <https://doi.org/10.1016/j.semcancer.2017.08.004>
- Chen, L., Li, H., Teng, J., Wang, Z., Qu, X., Chen, Z., Cai, X., Zeng, H., Bai, Z., Li, Jinghui, Pan, X., Yan, L., Wang, F., Lin, L., Luo, Y., Sahana, G., Lund, M.S., Ballester, M., Crespo-Piazuelo, D., Karlakov-Mortensen, P., Fredholm, M., Clop, A., Amills, M., Loving, C., Tuggle, C.K., Madsen, O., Li, Jiaqi, Zhang, Z., Liu, G.E., Jiang, J., Fang, L., Yi, G., 2023. Construction of a multi-tissue cell atlas reveals cell-type-

6 General discussion

specific regulation of molecular and complex phenotypes in pigs.
<https://doi.org/10.1101/2023.06.12.544530>

Chiappinelli, K.B., Zahnow, C.A., Ahuja, N., Baylin, S.B., 2016. Combining Epigenetic and Immunotherapy to Combat Cancer. *Cancer Research* 76, 1683–1689. <https://doi.org/10.1158/0008-5472.CAN-15-2125>

Christian, T.-M., Gooch, A., Vision, T., Hull, E., 2020. Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLOS ONE* 15, e0230281. <https://doi.org/10.1371/journal.pone.0230281>

Clark, E.L., Archibald, A.L., Daetwyler, H.D., Groenen, M.A.M., Harrison, P.W., Houston, R.D., Kühn, C., Lien, S., Macqueen, D.J., Reecy, J.M., Robledo, D., Watson, M., Tuggle, C.K., Giuffra, E., 2020. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biol* 21, 285. <https://doi.org/10.1186/s13059-020-02197-8>

Clark, S.J., Argelaguet, R., Lohoff, T., Krueger, F., Drage, D., Göttgens, B., Marioni, J.C., Nichols, J., Reik, W., 2022. Single-cell multi-omics profiling links dynamic DNA methylation to cell fate decisions during mouse early organogenesis. *Genome Biology* 23, 202. <https://doi.org/10.1186/s13059-022-02762-3>

Clevers, H., 2016. Modeling Development and Disease with Organoids. *Cell* 165, 1586–1597. <https://doi.org/10.1016/j.cell.2016.05.082>

Coelho-Santos, V., Shih, A.Y., 2020. Postnatal development of cerebrovascular structure and the neurogliovascular unit. *WIREs Developmental Biology* 9, e363. <https://doi.org/10.1002/wdev.363>

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsén, K., Larmande, P., Bras, Y.L., Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., Blanchet, C., 2017. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems* 75, 284–298. <https://doi.org/10.1016/j.future.2017.01.012>

Crespo-Piazuelo, D., González-Rodríguez, O., Mongellaz, M., Acloque, H., Mercat, M.-J., Bink, M. c. a. m., Huisman, A. e., Ramayo-Caldas, Y., Sánchez, J. p., Ballester, M., 2022. 524. Deciphering genetic variants from whole genome affecting duodenum, liver and muscle transcriptomes in pigs, in: *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*. Wageningen Academic Publishers, pp. 2174–2177. https://doi.org/10.3920/978-90-8686-940-4_524

de Mendoza, A., Poppe, D., Buckberry, S., Pflueger, J., Albertin, C.B., Daish, T., Bertrand, S., de la Calle-Mustienes, E., Gómez-Skarmeta, J.L., Nery, J.R., Ecker, J.R., Baer, B., Ragsdale, C.W., Grützner, F., Escriva, H., Venkatesh, B., Bogdanovic, O., Lister, R., 2021. The emergence of the brain non-CpG methylation system in vertebrates. *Nat Ecol Evol* 5, 369–378. <https://doi.org/10.1038/s41559-020-01371-2>

- Derks, M. f. l., Boshove, A., Harlizius, B., Sell-Kubiak, E., Lopes, M. s., Grindflek, E., Knol, E., Groenen, M. a. m., Gjuvslund, A. b., 2022. 564. A pan-genome of commercial pig breeds, in: Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). Wageningen Academic Publishers, pp. 2337–2341. https://doi.org/10.3920/978-90-8686-940-4_564
- Derks, M.F.L., Groenen, M.A.M., 2022. Use of Genome Sequencing for Improved Pig Breeding. Next-Generation Sequencing and Agriculture, CABI Biotechnology Series 139–151. <https://doi.org/10.1079/9781789247848.0006>
- Derks, M.F.L., Gross, C., Lopes, M.S., Reinders, M.J.T., Bosse, M., Gjuvslund, A.B., de Ridder, D., Megens, H.-J., Groenen, M.A.M., 2021. Accelerated discovery of functional genomic variation in pigs. *Genomics* 113, 2229–2239. <https://doi.org/10.1016/j.ygeno.2021.05.017>
- Derks, M.F.L., Schachtschneider, K.M., Madsen, O., Schijlen, E., Verhoeven, K.J.F., van Oers, K., 2016. Gene and transposable element methylation in great tit (*Parus major*) brain and blood. *BMC Genomics* 17, 332. <https://doi.org/10.1186/s12864-016-2653-y>
- Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C., 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 35, 316–319. <https://doi.org/10.1038/nbt.3820>
- Ernst, C.W., Steibel, J.P., 2013. Molecular advances in QTL discovery and application in pig breeding. *Trends in Genetics* 29, 215–224. <https://doi.org/10.1016/j.tig.2013.02.002>
- Ernst, J., Kellis, M., 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 12, 2478–2492. <https://doi.org/10.1038/nprot.2017.124>
- Fanelli, D., 2018. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences* 115, 2628–2631. <https://doi.org/10.1073/pnas.1708272114>
- Farlik, M., Halbritter, F., Müller, F., Choudry, F.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., Uppal, R., Stunnenberg, H.G., Ouwehand, W.H., Laurenti, E., Lengauer, T., Frontini, M., Bock, C., 2016. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* 19, 808–822. <https://doi.org/10.1016/j.stem.2016.10.019>
- Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., Schier, A.F., 2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* 360, eaar3131. <https://doi.org/10.1126/science.aar3131>
- Feregino, C., Sacher, F., Parnas, O., Tschopp, P., 2019. A single-cell transcriptomic atlas of the developing chicken limb. *BMC Genomics* 20, 401. <https://doi.org/10.1186/s12864-019-5802-2>
- Festing, M.F., 2018. On determining sample size in experiments involving laboratory animals. *Lab Anim* 52, 341–350. <https://doi.org/10.1177/0023677217738268>

6 General discussion

- García-Pérez, R., Esteller-Cucala, P., Mas, G., Lobón, I., Di Carlo, V., Riera, M., Kuhlilm, M., Navarro, A., Blancher, A., Di Croce, L., Gómez-Skarmeta, J.L., Juan, D., Marquès-Bonet, T., 2021. Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. *Nat Commun* 12, 3116. <https://doi.org/10.1038/s41467-021-23397-1>
- Gigante, S., Gouil, Q., Lucattini, A., Keniry, A., Beck, T., Tinning, M., Gordon, L., Woodruff, C., Speed, T.P., Blewitt, M.E., Ritchie, M.E., 2019. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Res* 47, e46. <https://doi.org/10.1093/nar/gkz107>
- Golicz, A.A., Bayer, P.E., Bhalla, P.L., Batley, J., Edwards, D., 2020. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics* 36, 132–145. <https://doi.org/10.1016/j.tig.2019.11.006>
- Goll, M.G., Halpern, M.E., 2011. DNA Methylation in Zebrafish. *Prog Mol Biol Transl Sci* 101, 193–218. <https://doi.org/10.1016/B978-0-12-387685-0.00005-6>
- Gorkin, D.U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J.S., Davidson, J.M., Qiu, Y., Afzal, V., Akiyama, J.A., Plajzer-Frick, I., Novak, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Mannion, B.J., Lee, E.A., Fukuda-Yuzawa, Y., He, Y., Preissl, S., Chee, S., Han, J.Y., Williams, B.A., Trout, D., Amrhein, H., Yang, H., Cherry, J.M., Wang, W., Gaulton, K., Ecker, J.R., Shen, Y., Dickel, D.E., Visel, A., Pennacchio, L.A., Ren, B., 2020. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* 583, 744–751. <https://doi.org/10.1038/s41586-020-2093-3>
- Gouil, Q., Keniry, A., 2019. Latest techniques to study DNA methylation. *Essays in Biochemistry* 63, 639–648. <https://doi.org/10.1042/EBC20190027>
- Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A., Taylor, J., 2018. Practical Computational Reproducibility in the Life Sciences. *Cell Systems* 6, 631–635. <https://doi.org/10.1016/j.cels.2018.03.014>
- Guedes, P.T., Oliveira, B.C.E.P.D. de, Manso, P.P. de A., Caputo, L.F.G., Cotta-Pereira, G., Pelajo-Machado, M., 2014. Histological Analyses Demonstrate the Temporary Contribution of Yolk Sac, Liver, and Bone Marrow to Hematopoiesis during Chicken Development. *PLOS ONE* 9, e90975. <https://doi.org/10.1371/journal.pone.0090975>
- Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G., Song, H., 2014. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci* 17, 215–222. <https://doi.org/10.1038/nn.3607>
- Hackshaw, A., 2008. Small studies: strengths and limitations. *European Respiratory Journal* 32, 1141–1143. <https://doi.org/10.1183/09031936.00136408>

- He, P., Williams, B.A., Trout, D., Marinov, G.K., Amrhein, H., Berghella, L., Goh, S.-T., Plajzer-Frick, I., Afzal, V., Pennacchio, L.A., Dickel, D.E., Visel, A., Ren, B., Hardison, R.C., Zhang, Y., Wold, B.J., 2020. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* 583, 760–767. <https://doi.org/10.1038/s41586-020-2536-x>
- He, Y., Hariharan, M., Gorkin, D.U., Dickel, D.E., Luo, C., Castanon, R., Nery, J.R., Lee, A.Y., Zhao, Y., Huang, H., Huang, H., Williams, B.A., Trout, D., Amrhein, H., Fang, R., Chen, H., Li, B., Visel, A., Pennacchio, L.A., Ren, B., Bing Ren, Ren, B., Ecker, J.R., 2020. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature* 583, 752–759. <https://doi.org/10.1038/s41586-020-2119-x>
- Heerboth, S., Lapinska, K., Snyder, N., Leary, M., Rollinson, S., Sarkar, S., 2014. Use of Epigenetic Drugs in Disease: An Overview. *Genet Epigenet* 6, 9–19. <https://doi.org/10.4137/GEG.S12270>
- Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., Ren, B., 2013. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet* 45, 1198–1206. <https://doi.org/10.1038/ng.2746>
- Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M., Davie, A., Robledo, D., 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. *Nat Rev Genet* 21, 389–409. <https://doi.org/10.1038/s41576-020-0227-y>
- Huang, Y.-Z., Sun, J.-J., Zhang, L.-Z., Li, C.-J., Womack, J.E., Li, Z.-J., Lan, X.-Y., Lei, C.-Z., Zhang, C.-L., Zhao, X., Chen, H., 2014. Genome-wide DNA Methylation Profiles and Their Relationships with mRNA and the microRNA Transcriptome in Bovine Muscle Tissue (*Bos taurine*). *Sci Rep* 4, 6546. <https://doi.org/10.1038/srep06546>
- Huse, S.M., Gruppuso, P.A., Boekelheide, K., Sanders, J.A., 2015. Patterns of gene expression and DNA methylation in human fetal and adult liver. *BMC Genomics* 16, 981. <https://doi.org/10.1186/s12864-015-2066-3>
- Hynds, R.E., Vladimirov, E., Janes, Sam.M., 2018. The secret lives of cancer cell lines. *Disease Models & Mechanisms* 11, dmm037366. <https://doi.org/10.1242/dmm.037366>
- Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. *PLOS Medicine* 2, e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Izquierdo-Torres, E., Hernández-Oliveras, A., Meneses-Morales, I., Rodríguez, G., Fuentes-García, G., Zarain-Herzberg, Á., 2019. Resveratrol up-regulates ATP2A3 gene expression in breast cancer cell lines through epigenetic mechanisms. *The International Journal of Biochemistry & Cell Biology* 113, 37–47. <https://doi.org/10.1016/j.biocel.2019.05.020>

6 General discussion

- Jang, H.S., Shin, W.J., Lee, J.E., Do, J.T., 2017. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes* 8, 148. <https://doi.org/10.3390/genes8060148>
- Juan, D., Santpere, G., Kelley, J.L., Cornejo, O.E., Marques-Bonet, T., 2023. Current advances in primate genomics: novel approaches for understanding evolution and disease. *Nat Rev Genet* 24, 314–331. <https://doi.org/10.1038/s41576-022-00554-w>
- Kang, J.G., Park, J.S., Ko, J.-H., Kim, Y.-S., 2019. Regulation of gene expression by altered promoter methylation using a CRISPR/Cas9-mediated epigenetic editing system. *Sci Rep* 9, 11960. <https://doi.org/10.1038/s41598-019-48130-3>
- Kar, S.K., Wells, J.M., Ellen, E.D., te Pas, M.F.W., Madsen, O., Groenen, M.A.M., Woelders, H., 2021. Organoids: a promising new in vitro platform in livestock and veterinary research. *Veterinary Research* 52, 43. <https://doi.org/10.1186/s13567-021-00904-2>
- Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G., Saelao, P., Waters, S., Xiang, R., Chamberlain, A., Korf, I., Delany, M.E., Cheng, H.H., Medrano, J.F., Van Eenennaam, A.L., Tuggle, C.K., Ernst, C., Flicek, P., Quon, G., Ross, P., Zhou, H., 2021. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun* 12, 1821. <https://doi.org/10.1038/s41467-021-22100-8>
- Kiecker, C., Bates, T., Bell, E., 2016. Molecular specification of germ layers in vertebrate embryos. *Cell. Mol. Life Sci.* 73, 923–947. <https://doi.org/10.1007/s00018-015-2092-y>
- Kircher, M., Witten, D.M., Jain, P., O’Roak, B.J., Cooper, G.M., Shendure, J., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310–315. <https://doi.org/10.1038/ng.2892>
- Knouse, K.A., Wu, J., Whittaker, C.A., Amon, A., 2014. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proceedings of the National Academy of Sciences* 111, 13409–13414. <https://doi.org/10.1073/pnas.1415287111>
- Kolarz, B., Majdan, M., 2017. Epigenetic aspects of rheumatoid arthritis: contribution of non-coding RNAs. *Seminars in Arthritis and Rheumatism* 46, 724–731. <https://doi.org/10.1016/j.semarthrit.2017.01.003>
- Kungulovski, G., Jeltsch, A., 2016. Epigenome Editing: State of the Art, Concepts, and Perspectives. *Trends in Genetics* 32, 101–113. <https://doi.org/10.1016/j.tig.2015.12.001>
- Laine, V.N., Gossmann, T.I., Schachtschneider, K.M., Garroway, C.J., Madsen, O., Verhoeven, K.J.F., de Jager, V., Megens, H.-J., Warren, W.C., Minx, P., Crooijmans, R.P.M.A., Corcoran, P., Sheldon, B.C., Slate, J., Zeng, K., van Oers, K., Visser, M.E., Groenen, M.A.M., 2016. Evolutionary signals of selection on

- cognition from the great tit genome and methylome. *Nat Commun* 7, 10474. <https://doi.org/10.1038/ncomms10474>
- Lakens, D., 2022. Sample Size Justification. *Collabra: Psychology* 8, 33267. <https://doi.org/10.1525/collabra.33267>
- Lallès, J.-P., 2016. Microbiota-host interplay at the gut epithelial level, health and nutrition. *J Animal Sci Biotechnol* 7, 66. <https://doi.org/10.1186/s40104-016-0123-7>
- Lancaster, M.A., Knoblich, J.A., 2014. Organogenesis in a dish: Modeling development and disease using organoid technologies. *Science* 345, 1247125. <https://doi.org/10.1126/science.1247125>
- Laude, H., Gelfi, J., 1979. Properties of Border disease virus as studied in a sheep cell line. *Arch Virol* 62, 341–346. <https://doi.org/10.1007/BF01318108>
- Lee, J.-H., Park, S.-J., Nakai, K., 2017. Differential landscape of non-CpG methylation in embryonic stem cells and neurons caused by DNMT3s. *Sci Rep* 7, 11295. <https://doi.org/10.1038/s41598-017-11800-1>
- Lee, J.-H., Saito, Y., Park, S.-J., Nakai, K., 2020. Existence and possible roles of independent non-CpG methylation in the mammalian brain. *DNA Research* 27, dsaa020. <https://doi.org/10.1093/dnares/dsaa020>
- Lee, K., Uh, K., Farrell, K., 2020. Current progress of genome editing in livestock. *Theriogenology, Proceedings of 19th International Congress on Animal Reproduction (ICAR)* 150, 229–235. <https://doi.org/10.1016/j.theriogenology.2020.01.036>
- Leighton, P.A., Schusser, B., Yi, H., Glanville, J., Harriman, W., 2015. A Diverse Repertoire of Human Immunoglobulin Variable Genes in a Chicken B Cell Line is Generated by Both Gene Conversion and Somatic Hypermutation. *Frontiers in Immunology* 6.
- Lewis, K., Yoshimoto, M., Takebe, T., 2021. Fetal liver hematopoiesis: from development to delivery. *Stem Cell Research & Therapy* 12, 139. <https://doi.org/10.1186/s13287-021-02189-w>
- Li, Q., Li, N., Hu, X., Li, J., Du, Z., Chen, L., Yin, G., Duan, J., Zhang, H., Zhao, Y., Wang, J., Li, N., 2011. Genome-Wide Mapping of DNA Methylation in Chicken. *PLoS One* 6, e19428. <https://doi.org/10.1371/journal.pone.0019428>
- Li, R., Zhu, J., 2022. Effects of aneuploidy on cell behaviour and function. *Nat Rev Mol Cell Biol* 23, 250–265. <https://doi.org/10.1038/s41580-021-00436-9>
- Lindner, M., Verhagen, I., Viitaniemi, H.M., Laine, V.N., Visser, M.E., Husby, A., van Oers, K., 2021. Temporal changes in DNA methylation and RNA expression in a small song bird: within- and between-tissue comparisons. *BMC Genomics* 22, 36. <https://doi.org/10.1186/s12864-020-07329-9>

6 General discussion

- Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghghi, F.G., Sejnowski, T.J., Behrens, M.M., Ecker, J.R., 2013. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* 341, 1237905. <https://doi.org/10.1126/science.1237905>
- Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R., Chen, H., Rivkin, A., Castanon, R.G., Clock, B., Li, Y.E., Hou, X., Poirion, O.B., Preissl, S., Pinto-Duarte, A., O'Connor, C., Boggeman, L., Fitzpatrick, C., Nunn, M., Mukamel, E.A., Zhang, Z., Callaway, E.M., Ren, B., Dixon, J.R., Behrens, M.M., Ecker, J.R., 2021. DNA methylation atlas of the mouse brain at single-cell resolution. *Nature* 598, 120–128. <https://doi.org/10.1038/s41586-020-03182-8>
- Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., Li, S., 2021. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biology* 22, 295. <https://doi.org/10.1186/s13059-021-02510-z>
- Lopez-Suarez, L., Awabdh, S.A., Coumoul, X., Chauvet, C., 2022. The SH-SY5Y human neuroblastoma cell line, a relevant in vitro cell model for investigating neurotoxicology in human: Focus on organic pollutants. *NeuroToxicology* 92, 131–155. <https://doi.org/10.1016/j.neuro.2022.07.008>
- Madhani, H.D., Francis, N.J., Kingston, R.E., Kornberg, R.D., Moazed, D., Narlikar, G.J., Panning, B., Struhl, K., 2008. Epigenomics: A Roadmap, But to Where? *Science* 322, 43–44. <https://doi.org/10.1126/science.322.5898.43b>
- Mangul, S., Mosqueiro, T., Abdill, R.J., Duong, D., Mitchell, K., Sarwal, V., Hill, B., Brito, J., Littman, R.J., Statz, B., Lam, A.K.-M., Dayama, G., Grieneisen, L., Martin, L.S., Flint, J., Eskin, E., Blekhman, R., 2019. Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLOS Biology* 17, e3000333. <https://doi.org/10.1371/journal.pbio.3000333>
- Mantri, M., Scuderi, G.J., Abedini-Nassab, R., Wang, M.F.Z., McKellar, D., Shi, H., Grodner, B., Butcher, J.T., De Vlaminc, I., 2021. Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. *Nat Commun* 12, 1771. <https://doi.org/10.1038/s41467-021-21892-z>
- Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S.K., Bignell, A., Boddu, S., Branco Lins, P.R., Brooks, L., Ramaraju, S.B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., Davidson, C., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Fatima, R., Giron, C.G., Genez, T., Ghattaoraya, G.S., Martinez, J.G., Guijarro, C., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Marques-Coelho, D., Marugán, J.C., Merino, G.A., Mirabueno, L.P., Mushtaq, A., Hossain, S.N., Ogeh, D.N., Sakthivel, M.P., Parker, A., Perry, M., Piližota, I., Prosovetskaia, I., Pérez-Silva, J.G., Salam,

- A.I.A., Saraiva-Agostinho, N., Schuilenburg, H., Sheppard, D., Sinha, S., Sipos, B., Stark, W., Steed, E., Sukumaran, R., Sumathipala, D., Suner, M.-M., Surapaneni, L., Sutinen, K., Szpak, M., Tricomi, F.F., Urbina-Gómez, D., Veidenberg, A., Walsh, T.A., Walts, B., Wass, E., Willhoft, N., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G.R., Loveland, J.E., Moore, B., Mudge, J.M., Tate, J., Thybert, D., Trevanion, S.J., Winterbottom, A., Frankish, A., Hunt, S.E., Ruffier, M., Cunningham, F., Dyer, S., Finn, R.D., Howe, K.L., Harrison, P.W., Yates, A.D., Flicek, P., 2023. *Ensembl 2023*. *Nucleic Acids Research* 51, D933–D941. <https://doi.org/10.1093/nar/gkac958>
- Marx, V., 2020. When computational pipelines go ‘clank.’ *Nat Methods* 17, 659–662. <https://doi.org/10.1038/s41592-020-0886-9>
- Meekins, D.A., Morozov, I., Trujillo, J.D., Gaudreault, N.N., Bold, D., Carossino, M., Artiaga, B.L., Indran, S.V., Kwon, T., Balaraman, V., Madden, D.W., Feldmann, H., Henningson, J., Ma, W., Balasuriya, U.B.R., Richt, J.A., 2020. Susceptibility of swine cells and domestic pigs to SARS-CoV-2. *Emerging Microbes & Infections* 9, 2278–2288. <https://doi.org/10.1080/22221751.2020.1831405>
- Meuwissen, T., Hayes, B., Goddard, M., 2016. Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers* 6, 6–14. <https://doi.org/10.2527/af.2016-0002>
- Mitalipov, S.M., 2006. Genomic imprinting in primate embryos and embryonic stem cells. *Reprod. Fertil. Dev.* 18, 817–821. <https://doi.org/10.1071/RD06112>
- Mollandin, F., 2022. Integration of functional annotations in Bayesian genomic prediction models (Doctoral thesis). université Paris-Saclay.
- Moreira, G. c. m., Dupont, S., Becker, D., Salavati, M., Clark, R., Clark, E. I., Plastow, G., Kühn, C., Charlier, C., 2022. 545. Multi-dimensional functional annotation of bovine genome for the BovReg project, in: *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*. Wageningen Academic Publishers, pp. 2261–2264. https://doi.org/10.3920/978-90-8686-940-4_545
- Muhr, J., Ackerman, K.M., 2023. *Embryology, Gastrulation*, in: *StatPearls*. StatPearls Publishing, Treasure Island (FL).
- Nakafuku, M., del Águila, Á., 2020. Developmental dynamics of neurogenesis and gliogenesis in the postnatal mammalian brain in health and disease: Historical and future perspectives. *WIREs Developmental Biology* 9, e369. <https://doi.org/10.1002/wdev.369>
- Nakamura, M., Gao, Y., Dominguez, A.A., Qi, L.S., 2021. CRISPR technologies for precise epigenome editing. *Nat Cell Biol* 23, 11–22. <https://doi.org/10.1038/s41556-020-00620-7>
- Nakano, K., Whitaker, J.W., Boyle, D.L., Wang, W., Firestein, G.S., 2013. DNA methylome signature in rheumatoid arthritis. *Ann Rheum Dis* 72, 110–117. <https://doi.org/10.1136/annrheumdis-2012-201526>

6 General discussion

- Nawy, T., 2014. Single-cell sequencing. *Nat Methods* 11, 18–18. <https://doi.org/10.1038/nmeth.2771>
- OECD/FAO, 2022. OECD-FAO Agricultural Outlook 2022-2031. OECD Publishing, Meat. <https://doi.org/10.1787/f1b0b29c-en>
- Orozco, G., 2022. Fine mapping with epigenetic information and 3D structure. *Semin Immunopathol* 44, 115–125. <https://doi.org/10.1007/s00281-021-00906-4>
- Orozco, L.D., Morselli, M., Rubbi, L., Guo, W., Go, J., Shi, H., Lopez, D., Furlotte, N.A., Bennett, B.J., Farber, C.R., Ghazalpour, A., Zhang, M.Q., Bahous, R., Rozen, R., Lusic, A.J., Pellegrini, M., 2015. Epigenome-Wide Association of Liver Methylation Patterns and Complex Metabolic Traits in Mice. *Cell Metabolism* 21, 905–917. <https://doi.org/10.1016/j.cmet.2015.04.025>
- Pan, Z., Wang, Ying, Wang, M., Wang, Yuzhe, Zhu, X., Gu, S., Zhong, C., An, L., Shan, M., Damas, J., Halstead, M.M., Guan, D., Trakooljul, N., Wimmers, K., Bi, Y., Wu, S., Delany, M.E., Bai, X., Cheng, H.H., Sun, C., Yang, N., Hu, X., Lewin, H.A., Fang, L., Zhou, H., 2023. An atlas of regulatory elements in chicken: A resource for chicken genetics and genomics. *Science Advances* 9, eade1204. <https://doi.org/10.1126/sciadv.ade1204>
- Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L., Kern, C., Halstead, M., Chanthavixay, G., Trakooljul, N., Wimmers, K., Sahana, G., Su, G., Lund, M.S., Fredholm, M., Karlskov-Mortensen, P., Ernst, C.W., Ross, P., Tuggle, C.K., Fang, L., Zhou, H., 2021. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat Commun* 12, 5848. <https://doi.org/10.1038/s41467-021-26153-7>
- Pfister, S.X., Ashworth, A., 2017. Marked for death: targeting epigenetic changes in cancer. *Nat Rev Drug Discov* 16, 241–263. <https://doi.org/10.1038/nrd.2016.256>
- Pinho, R.M., Maga, E.A., 2021. DNA methylation as a regulator of intestinal gene expression. *Br J Nutr* 126, 1611–1625. <https://doi.org/10.1017/S0007114521000556>
- Pizzorusso, T., Tognini, P., 2020. Interplay between Metabolism, Nutrition and Epigenetics in Shaping Brain DNA Methylation, Neural Function and Behavior. *Genes* 11, 742. <https://doi.org/10.3390/genes11070742>
- Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., Sandhu, M.S., 2018. Long reads: their purpose and place. *Human Molecular Genetics* 27, R234–R241. <https://doi.org/10.1093/hmg/ddy177>
- Rakoff-Nahoum, S., Kong, Y., Kleinstein, S.H., Subramanian, S., Ahern, P.P., Gordon, J.I., Medzhitov, R., 2015. Analysis of gene–environment interactions in postnatal development of the mammalian intestine. *Proceedings of the National Academy of Sciences* 112, 1929–1936. <https://doi.org/10.1073/pnas.1424886112>

- Ran, J., Li, J., Yin, L., Zhang, D., Yu, C., Du, H., Jiang, X., Yang, C., Liu, Y., 2021. Comparative Analysis of Skeletal Muscle DNA Methylation and Transcriptome of the Chicken Embryo at Different Developmental Stages. *Frontiers in Physiology* 12.
- Raza, S.H.A., Hassanin, A.A., Pant, S.D., Bing, S., Sitohy, M.Z., Abdelnour, S.A., Alotaibi, M.A., Al-Hazani, T.M., Abd El-Aziz, A.H., Cheng, G., Zan, L., 2022. Potentials, prospects and applications of genome editing technologies in livestock production. *Saudi Journal of Biological Sciences* 29, 1928–1935. <https://doi.org/10.1016/j.sjbs.2021.11.037>
- Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., Kircher, M., 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* 47, D886–D894. <https://doi.org/10.1093/nar/gky1016>
- Rezaei, S., Uffendorde, J., Gimm, O., Hosseinpour Feizi, M.A., Miemczyk, S., Coutinho, L.L., Jensen, P., Guerrero-Bosagna, C., Pértille, F., 2022. GBS-MeDIP: A protocol for parallel identification of genetic and epigenetic variation in the same reduced fraction of genomes across individuals. *STAR Protocols* 3, 101202. <https://doi.org/10.1016/j.xpro.2022.101202>
- Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.-C., Pfening, A.R., Wang, X., Claussnitzer, M., Yaping Liu, Coarfa, C., Alan Harris, R., Shores, N., Epstein, C.B., Gjonneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Scott Hansen, R., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthal, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J.M., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.-H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoiyannopoulos, J.A., Wang, T., Kellis, M., 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. <https://doi.org/10.1038/nature14248>
- Rodriguez-Polo, I., Behr, R., 2022. Non-human primate pluripotent stem cells for the preclinical testing of regenerative therapies. *Neural Regen Res* 17, 1867–1874. <https://doi.org/10.4103/1673-5374.335689>
- Rosenkrantz, J.L., Carbone, L., 2017. Investigating somatic aneuploidy in the brain: why we need a new model. *Chromosoma* 126, 337–350. <https://doi.org/10.1007/s00412-016-0615-4>
- Ross, C., Boroviak, T.E., 2020. Origin and function of the yolk sac in primate embryogenesis. *Nat Commun* 11, 3760. <https://doi.org/10.1038/s41467-020-17575-w>

- Roura, E., Koopmans, S.-J., Lallès, J.-P., Le Huerou-Luron, I., De Jager, N., Schuurman, T., Val-Laillet, D., 2016. Critical review evaluating the pig as a model for human nutritional physiology. *Nutr. Res. Rev.* 29, 60–90. <https://doi.org/10.1017/S0954422416000020>
- Rubenstein, A.B., Smith, G.R., Raue, U., Begue, G., Minchev, K., Ruf-Zamojski, F., Nair, V.D., Wang, X., Zhou, L., Zaslavsky, E., Trappe, T.A., Trappe, S., Sealfon, S.C., 2020. Single-cell transcriptional profiles in human skeletal muscle. *Sci Rep* 10, 229. <https://doi.org/10.1038/s41598-019-57110-6>
- Ruoß, M., Damm, G., Vosough, M., Ehret, L., Grom-Baumgarten, C., Petkov, M., Naddalin, S., Ladurner, R., Seehofer, D., Nussler, A., Sajadian, S., 2019. Epigenetic Modifications of the Liver Tumor Cell Line HepG2 Increase Their Drug Metabolic Capacity. *International Journal of Molecular Sciences* 20, 347. <https://doi.org/10.3390/ijms20020347>
- Salehin, N., Knowles, H., Masamsetti, V.P., Tam, P.P.L., 2022. Mammalian gastrulation: signalling activity and transcriptional regulation of cell lineage differentiation and germ layer formation. *Biochemical Society Transactions* 50, 1619–1631. <https://doi.org/10.1042/BST20220256>
- Sánchez, O.F., Lin, L., Bryan, C.J., Xie, J., Freeman, J.L., Yuan, C., 2020. Profiling epigenetic changes in human cell line induced by atrazine exposure. *Environmental Pollution* 258, 113712. <https://doi.org/10.1016/j.envpol.2019.113712>
- Satterlee, J.S., Chadwick, L.H., Tyson, F.L., McAllister, K., Beaver, J., Birnbaum, L., Volkow, N.D., Wilder, E.L., Anderson, J.M., Roy, A.L., 2019. The NIH Common Fund/Roadmap Epigenomics Program: Successes of a comprehensive consortium. *Science Advances* 5, eaaw6507. <https://doi.org/10.1126/sciadv.aaw6507>
- Schachtschneider, K.M., Madsen, O., Park, C., Rund, L.A., Groenen, M.A.M., Schook, L.B., 2015. Adult porcine genome-wide DNA methylation patterns support pigs as a biomedical model. *BMC Genomics* 16, 743. <https://doi.org/10.1186/s12864-015-1938-x>
- Schaid, D.J., Chen, W., Larson, N.B., 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19, 491–504. <https://doi.org/10.1038/s41576-018-0016-z>
- Schultz, M.D., He, Y., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A.D., Selvaraj, S., Ren, B., Sejnowski, T.J., Wang, W., Ecker, J.R., 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 523, 212–216. <https://doi.org/10.1038/nature14465>
- Schwartzman, O., Tanay, A., 2015. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 16, 716–726. <https://doi.org/10.1038/nrg3980>

- Sharma, A., Lee, J.S., Dang, C.G., Sudrajad, P., Kim, H.C., Yeon, S.H., Kang, H.S., Lee, S.-H., 2015. Stories and Challenges of Genome Wide Association Studies in Livestock — A Review. *Asian-Australas J Anim Sci* 28, 1371–1379. <https://doi.org/10.5713/ajas.14.0715>
- Sheltzer, J.M., Amon, A., 2011. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends in Genetics* 27, 446–453. <https://doi.org/10.1016/j.tig.2011.07.003>
- Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., Levin, A.M., Eng, C., Yazdanbakhsh, M., Wilson, J.G., Marrugo, J., Lange, L.A., Williams, L.K., Watson, H., Ware, L.B., Olopade, C.O., Olopade, O., Oliveira, R.R., Ober, C., Nicolae, D.L., Meyers, D.A., Mayorga, A., Knight-Madden, J., Hartert, T., Hansel, N.N., Foreman, M.G., Ford, J.G., Faruque, M.U., Dunston, G.M., Caraballo, L., Burchard, E.G., Bleecker, E.R., Araujo, M.I., Herrera-Paz, E.F., Campbell, M., Foster, C., Taub, M.A., Beaty, T.H., Ruczinski, I., Mathias, R.A., Barnes, K.C., Salzberg, S.L., 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* 51, 30–35. <https://doi.org/10.1038/s41588-018-0273-y>
- Sherman, R.M., Salzberg, S.L., 2020. Pan-genomics in the human genome era. *Nat Rev Genet* 21, 243–254. <https://doi.org/10.1038/s41576-020-0210-7>
- Shiffrin, R.M., Börner, K., Stigler, S.M., 2018. Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences* 115, 2632–2639. <https://doi.org/10.1073/pnas.1711786114>
- Silva-Pedrosa, R., Salgado, A.J., Ferreira, P.E., 2023. Revolutionizing Disease Modeling: The Emergence of Organoids in Cellular Systems. *Cells* 12, 930. <https://doi.org/10.3390/cells12060930>
- Silver, A., 2017. Software simplified. *Nature* 546, 173–174. <https://doi.org/10.1038/546173a>
- Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., Timp, W., 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14, 407–410. <https://doi.org/10.1038/nmeth.4184>
- Stodden, V., Seiler, J., Ma, Z., 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* 115, 2584–2589. <https://doi.org/10.1073/pnas.1708290115>
- Su, F., Liu, X., Liu, G., Yu, Y., Wang, Y., Jin, Y., Hu, G., Hua, S., Zhang, Y., 2013. Establishment and Evaluation of a Stable Cattle Type II Alveolar Epithelial Cell Line. *PLOS ONE* 8, e76036. <https://doi.org/10.1371/journal.pone.0076036>
- Tian, Q., He, J.-P., Zhu, C., Zhu, Q.-Y., Li, Y.-G., Liu, J.-L., 2022. Revisiting the Transcriptome Landscape of Pig Embryo Implantation Site at Single-Cell Resolution. *Frontiers in Cell and Developmental Biology* 10.

6 General discussion

- Tomaselli, D., Lucidi, A., Rotili, D., Mai, A., 2020. Epigenetic polypharmacology: A new frontier for epi-drug discovery. *Medicinal Research Reviews* 40, 190–244. <https://doi.org/10.1002/med.21600>
- Tonzani, S., Fiorani, S., 2021. The STAR Methods way towards reproducibility and open science. *iScience* 24, 102137. <https://doi.org/10.1016/j.isci.2021.102137>
- Triantaphyllopoulos, K.A., Ikonomopoulos, I., Bannister, A.J., 2016. Epigenetics and inheritance of phenotype variation in livestock. *Epigenetics & Chromatin* 9, 31. <https://doi.org/10.1186/s13072-016-0081-5>
- van der Hee, B., Madsen, O., Vervoort, J., Smidt, H., Wells, J.M., 2020. Congruence of Transcription Programs in Adult Stem Cell-Derived Jejunum Organoids and Original Tissue During Long-Term Culture. *Front Cell Dev Biol* 8, 375. <https://doi.org/10.3389/fcell.2020.00375>
- van der Velde, A., Fan, K., Tsuji, J., Moore, J.E., Purcaro, M.J., Pratt, H.E., Weng, Z., 2021. Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol* 4, 1–15. <https://doi.org/10.1038/s42003-021-01756-4>
- Vcelar, S., Jadhav, V., Melcher, M., Auer, N., Hrdina, A., Sagmeister, R., Heffner, K., Puklowski, A., Betenbaugh, M., Wenger, T., Leisch, F., Baumann, M., Borth, N., 2018. Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting. *Biotechnology and Bioengineering* 115, 165–173. <https://doi.org/10.1002/bit.26453>
- Verma, A., 2014. Chapter 12 - Animal Tissue Culture: Principles and Applications, in: Verma, A.S., Singh, A. (Eds.), *Animal Biotechnology*. Academic Press, San Diego, pp. 211–231. <https://doi.org/10.1016/B978-0-12-416002-6.00012-2>
- Verma, A., Verma, M., Singh, A., 2020. Chapter 14 - Animal tissue culture principles and applications, in: Verma, A.S., Singh, A. (Eds.), *Animal Biotechnology (Second Edition)*. Academic Press, Boston, pp. 269–293. <https://doi.org/10.1016/B978-0-12-811710-1.00012-4>
- Wang, T., Song, J., Qu, M., Gao, X., Zhang, W., Wang, Z., Zhao, L., Wang, Y., Li, B., Li, J., Yang, J., 2021. Integrative Epigenome Map of the Normal Human Prostate Provides Insights Into Prostate Cancer Predisposition. *Frontiers in Cell and Developmental Biology* 9.
- Waterland, R.A., Kellermayer, R., Rached, M.-T., Tatevian, N., Gomes, M.V., Zhang, J., Zhang, L., Chakravarty, A., Zhu, W., Laritsky, E., Zhang, W., Wang, X., Shen, L., 2009. Epigenomic profiling indicates a role for DNA methylation in early postnatal liver development. *Human Molecular Genetics* 18, 3026–3038. <https://doi.org/10.1093/hmg/ddp241>
- Watson, H., Salmón, P., Isaksson, C., 2019. Dynamic changes in DNA methylation during embryonic and postnatal development of an altricial wild bird. *Ecology and Evolution* 9, 9580–9585. <https://doi.org/10.1002/ece3.5480>

- Weikard, R., Demasius, W., Kuehn, C., 2017. Mining long noncoding RNA in livestock. *Anim Genet* 48, 3–18. <https://doi.org/10.1111/age.12493>
- Weström, B., Arévalo Sureda, E., Pierzynowska, K., Pierzynowski, S.G., Pérez-Cano, F.-J., 2020. The Immature Gut Barrier and Its Importance in Establishing Immunity in Newborn Mammals. *Frontiers in Immunology* 11.
- Wong, E.A., Uni, Z., 2021. Centennial Review: The chicken yolk sac is a multifunctional organ. *Poultry Science* 100, 100821. <https://doi.org/10.1016/j.psj.2020.11.004>
- Wratten, L., Wilm, A., Göke, J., 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* 18, 1161–1168. <https://doi.org/10.1038/s41592-021-01254-9>
- Xiang, R., Berg, I. van den, MacLeod, I.M., Hayes, B.J., Prowse-Wilkins, C.P., Wang, M., Bolormaa, S., Liu, Z., Rochfort, S.J., Reich, C.M., Mason, B.A., Vander Jagt, C.J., Daetwyler, H.D., Lund, M.S., Chamberlain, A.J., Goddard, M.E., 2019. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences* 116, 19398–19408. <https://doi.org/10.1073/pnas.1904159116>
- Xiang, Y., Zhang, Y., Xu, Q., Zhou, C., Liu, B., Du, Z., Zhang, K., Zhang, B., Wang, X., Gayen, S., Liu, L., Wang, Y., Li, Y., Wang, Q., Kalantry, S., Li, L., Xie, W., 2020. Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency. *Nat Genet* 52, 95–105. <https://doi.org/10.1038/s41588-019-0545-1>
- Yang, Y., Fan, X., Yan, J., Chen, M., Zhu, M., Tang, Y., Liu, S., Tang, Z., 2021. A comprehensive epigenome atlas reveals DNA methylation regulating skeletal muscle development. *Nucleic Acids Research* 49, 1313–1329. <https://doi.org/10.1093/nar/gkaa1203>
- Yin, Y., Liu, P.-Y., Shi, Y., Li, P., 2021. Single-Cell Sequencing and Organoids: A Powerful Combination for Modelling Organ Development and Diseases, in: Pedersen, S.H.F. (Ed.), *Reviews of Physiology, Biochemistry and Pharmacology, Reviews of Physiology, Biochemistry and Pharmacology*. Springer International Publishing, Cham, pp. 189–210. https://doi.org/10.1007/112_2020_47
- Yizhar-Barnea, O., Valensisi, C., Jayavelu, N.D., Kishore, K., Andrus, C., Koffler-Brill, T., Ushakov, K., Perl, K., Noy, Y., Bhonker, Y., Pelizzola, M., Hawkins, R.D., Avraham, K.B., 2018. DNA methylation dynamics during embryonic development and postnatal maturation of the mouse auditory sensory epithelium. *Sci Rep* 8, 17348. <https://doi.org/10.1038/s41598-018-35587-x>
- Yu, D.-H., Gadkari, M., Zhou, Q., Yu, S., Gao, N., Guan, Y., Schady, D., Roshan, T.N., Chen, M.-H., Laritsky, E., Ge, Z., Wang, H., Chen, R., Westwater, C., Bry, L., Waterland, R.A., Moriarty, C., Hwang, C., Swennes, A.G., Moore, S.R., Shen, L., 2015. Postnatal epigenetic regulation of intestinal stem cells

requires DNA methylation and is guided by the microbiome. *Genome Biology* 16, 211. <https://doi.org/10.1186/s13059-015-0763-5>

Zhang, H., Wong, E.A., 2019. Expression of avian β -defensin mRNA in the chicken yolk sac. *Developmental & Comparative Immunology* 95, 89–95. <https://doi.org/10.1016/j.dci.2019.02.006>

Zhang, Y., Rohde, C., Reinhardt, R., Voelcker-Rehage, C., Jeltsch, A., 2009. Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biology* 10, R138. <https://doi.org/10.1186/gb-2009-10-12-r138>

Zhao, Z., Chen, X., Dowbaj, A.M., Sljukic, A., Bratlie, K., Lin, L., Fong, E.L.S., Balachander, G.M., Chen, Z., Soragni, A., Huch, M., Zeng, Y.A., Wang, Q., Yu, H., 2022. Organoids. *Nat Rev Methods Primers* 2, 1–21. <https://doi.org/10.1038/s43586-022-00174-y>

Zhu, J., Chen, F., Luo, L., Wu, W., Dai, J., Zhong, J., Lin, X., Chai, C., Ding, P., Liang, L., Wang, S., Ding, X., Chen, Y., Wang, H., Qiu, J., Wang, F., Sun, C., Zeng, Y., Fang, J., Jiang, X., Liu, P., Tang, G., Qiu, X., Zhang, X., Ruan, Y., Jiang, S., Li, J., Zhu, S., Xu, X., Li, F., Liu, Z., Cao, G., Chen, D., 2021. Single-cell atlas of domestic pig cerebral cortex and hypothalamus. *Science Bulletin* 66, 1448–1461. <https://doi.org/10.1016/j.scib.2021.04.002>

Zorn, A.M., Wells, J.M., 2007. Molecular Basis of Vertebrate Endoderm Development, in: *International Review of Cytology*. Academic Press, pp. 49–111. [https://doi.org/10.1016/S0074-7696\(06\)59002-3](https://doi.org/10.1016/S0074-7696(06)59002-3)

Appendix A:

WCGALP conference paper

Allele specific expression as an indication of ploidy in pig IPECJ2 and chicken SL-29 cell lines

Jani de Vos^{1*}, Richard P.M.A. Crooijmans¹, Martijn F.L. Derks¹, Susan L. Kloet², Martien
A.M. Groenen¹, Ole Madsen¹

¹ Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands;

² Leids Universitair Medisch Centrum, The Netherlands * jani.devos@wur.nl

Abstract

Cell lines are useful for investigating traits of interest e.g. intestinal absorption, feed efficiency and immunity in farm animals. We earlier investigated an intestinal cell line in pig and a fibroblast cell line in chicken and found chromosomal abnormalities by whole genome sequence (WGS) data analysis. Results from RNA-seq allele-specific expression (ASE) analysis in 4 cell lines showed aneuploidy in some chromosomes. In this paper we show that RNA-seq can be used to detect whole/partial chromosomal abnormalities based on ASE analysis.

Introduction

Pig and chicken are the primary sources for meat production worldwide and are also important for use as biomedical models and to study embryonic development. Cell lines are a valuable tool for gaining insight into genomic architecture and regulatory regions of genomes. Cell lines are derived from specific tissues of a species and can either continue to divide endlessly or perish after a specific number of divisions (passages) (Verma et al., 2020). Immortalized cell lines (i.e., that can be grown indefinitely) often show aneuploidy (presence of an abnormal number of chromosomes in a cell) or heteroploidy which is most pronounced in cancer cell lines (Verma et al., 2020; Molina et al., 2021). Animal cell lines from pig or chicken have proven useful to obtain insight in e.g. intestinal transport and immune function (Nossol et al., 2015).

The phenotype of an individual is determined by gene expression which is a process where information from genes, encoded within DNA, is translated into proteins through RNA (Hartwell et al., 2015). The phenomenon of unequal expression between alleles in diploid cells caused by either genetic variation or epigenetic regulation (cis-regulated gene expression) is known as ASE. Furthermore, non-haploid chromosomes show imbalances in gene transcriptional activity, implying that chromosomal abnormalities can influence gene expression of different alleles, resulting in ASE. Our aim was to investigate the usability of RNA-seq data to identify whole chromosomal abnormalities, with the pig IPEC-J2 and chicken SL-29 cell lines as model.

Materials & Methods

Cell lines. The pig IPECJ67 cell line (IPECJ2 cells grown for 67 passages) and chicken SL-29 cell line (grown for 4 passages) used were obtained from the cell repositories at DSMZ (<https://www.dsmz.de/collection/catalogue/details/culture/ACC-701>) and ATCC (https://www.lgcstandards-atcc.org/products/all/CRL-1590.aspx?geo_country=nl#generalinformation) respectively. The pig cell line is derived from intestinal epithelial cells while the chicken cell line is derived from embryonic fibroblast cells.

Appendix A – Conference paper

Data analysis. Pig (*Sus Scrofa* 11.1) and chicken (*Gallus gallus* GRCg6a) reference genomes, together with ENSEMBL annotations (*Sus Scrofa* 11.1 - release 103 & *Gallus gallus* GRCg6a - release 94) were utilized for all data analyses of our study.

Whole genome sequencing. Whole genome sequences were trimmed using Sickle v1.33 (<https://github.com/najoshi/sickle>) in paired end mode, followed by alignment (bwa mem v0.7.15 (Li., 2013)) of the trimmed reads, and removal of duplicates (Samblaster v0.1.26 (Faust & Hall, 2014)). Mate coordinates were added using samtools (v1.9, Li *et al.*, 2009). Single nucleotide variants (SNVs) were called using FreeBayes (v1.3.1), thereafter the read support ratio was assessed for heterozygous variants within the VCF file.

RNA sequencing. RNA-seq data (pig IPECJ67 and chicken SL-29) were trimmed for adapters, and minimum length using TrimGalore v0.6.4 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). FastQC v0.11.9 (Andrews *et al.*, 2010) was implemented to evaluate the quality of the data. RSEM (Li and Dewey, 2011) was used for alignment of the trimmed reads, as well as gene quantification (STAR v2.7.3a as aligner) (Dobin *et al.*, 2013). Average gene expression level per chromosome was calculated and plotted using a custom python script with the Seaborn package. Additional raw RNA-seq data from the PRJNA610529 project was downloaded from ENA. This data comprises a pig Jejunum organoid sample (12 weeks (SAMN14300021), a 5-week-old pig Jejunum tissue sample (SAMN14300018), cell lines IPECJ87, an IPECJ2 cell line grown for 87 passages (SAMN14300016), and IPECJ91 an IPECJ2 cell line grown for 91 passages (SAMN14299997). We trimmed, aligned, and completed gene quantification of this data following the same procedure as the above for the IPECJ67 cell line. These samples were used to compare the average gene expression levels per chromosome in the tissue, organoid, and cell lines to the IPECJ67 cell line. Initial ASE analysis was completed with variant calls from WGS data (FreeBayes, Garrison and Marth, 2012) and aligned reads from RNA-seq using GATK ASEReadr (McKenna *et al.*, 2010). Variant calling from RNA-seq alignments for SL-29, IPECJ67, IPECJ87, and IPECJ91 was completed using FreeBayes and thereafter allele specific expression analysis was completed using GATK ASEReadr implementing the bam and VCF files. Results from the ASE analysis was plotted using Seaborn package in python.

Results

Chromosomal abnormalities within the cell-line genomes. In a previous study of the cell lines IPECJ67 and SL-29 we analysed the structure of the genome in multiple aspects using whole genome sequence data (**Chapter 2**). In pig IPECJ67 we observed that chromosome 16 is diploid and chromosome 17 is triploid based on the frequency of the alleles (data not shown).

Additionally, aneuploidy and structural variations were observed in many other chromosomes in the IPECJ67 cell line as well as chicken SL-29 (e.g. chromosome 20 is tetraploid).

Allele specific expression. RNA-seq data provides insight into gene expression levels at a chromosome-wide level. We investigated the expression of genes and transcripts in the pig IPECJ67, IPECJ87, IPECJ91 cell lines, jejunum intestinal tissue and jejunum derived organoid sample (data not shown). We found elevated gene expression on chromosome 17 in the cell lines compared to organoid and tissue whereas for the diploid chromosomes 15 and 16 similar gene expression levels across tissue, organoids, and cell lines, with slightly higher expression levels in cell lines were observed. Investigating the allele specific expression of the IPECJ67 cell line (figure 1) shows that the frequency of the allele expression displays a pattern that confirms chromosome 17 as triploid and chromosomes 15 and 16 as diploid.

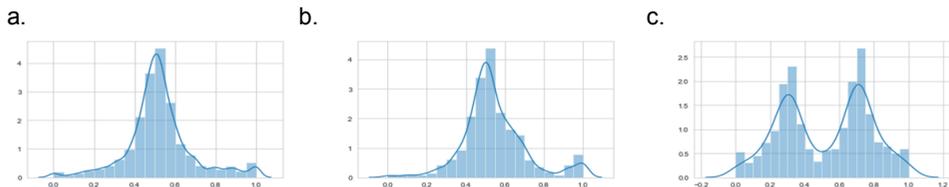


Figure 1. Allele specific expression of chromosomes 15(a), 16(b), 17(c) for the IPECJ67 cell line supports the ploidy level of the three chromosomes - allele expression of ~ 0.5 for diploids and allele expression of ~ 0.33 and ~ 0.7 for triploids, respectively.

Further investigation of the ASE and chromosomal abnormalities of the IPECJ2 derived cell lines cultured for a longer time also showed triploidy for chromosome 17 (figure 2 c,f) in both cell cultures. Interestingly, chromosome 16 was observed as triploid (figure 2 b,e) for these cell cultures, which is different from IPECJ67 which was diploid for chromosome 16 (figure 1b).

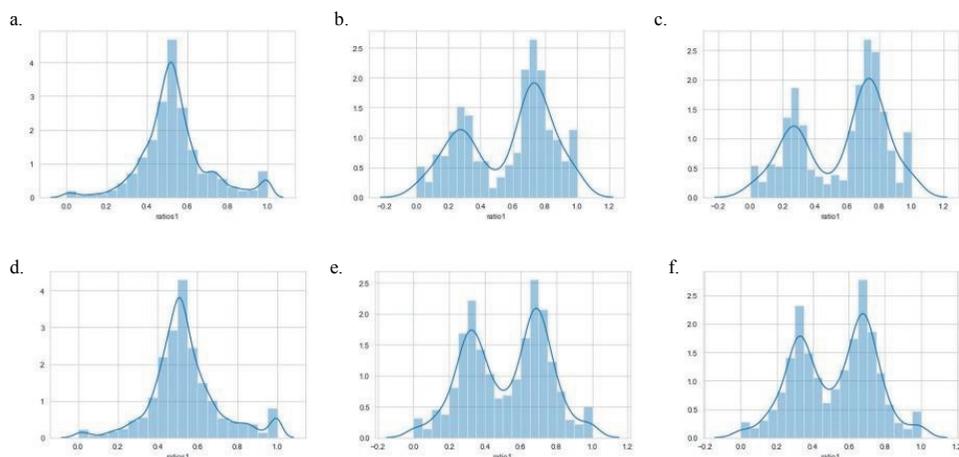


Figure 2 ASE for IPECJ2 derived cell lines cultured for a longer time (passages). IPECJ87 a. chromosome 15, b. chromosome 16, c. chromosome 17. IPECJ91 d. chromosome 15, e. chromosome 16, f. chromosome 17.

Similar patterns of allele distribution for aneuploid chromosomes detected using ASE are observed on other chromosomes in pig IPECJ67 and in chicken SL-29 (**Chapter 2**). Lastly, we called the genome variation from the RNA-seq data and repeated the ASE analysis based on the RNA-seq called genome variation. This resulted in the same ASE results as using the variation called from the genome sequencing data (**Chapter 2**). Thus, genome sequencing is not required to detect genome aberrations based on ASE.

Discussion

Detecting chromosome abnormalities in e.g. cell lines is important as such abnormalities may influence the results obtained in cell line assays, especially if genes involved in the process studied show copy number variation as this will result in deviating expression (Figure 1c). Chromosome ploidy can be investigated by e.g. cytogenetic and whole genome sequencing methods, but these methods are generally not applied in standard cell-line research. As RNA-seq is (becoming) the general method used for gene expression detection, also in cell-lines, we investigated the potential of implementing RNA-seq as a tool for detecting chromosomal abnormalities through variant calling, ASE analysis and plotting the allele ratios. We observed similar patterns of allele support ratios from WGS variant calling and RNA-seq allelic

expression variants. This suggests we do not need WGS data for detection of chromosomal abnormalities. We show that a larger number of passages of the IPECJ2 cell line results in an increased level of aneuploidy. Thus, our approach, using RNA-seq, is a sufficient and cost-effective tool to detect aneuploidy, which we suggest to be regularly applied in cell line experiments.

References

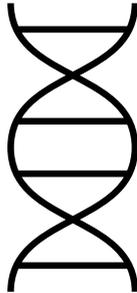
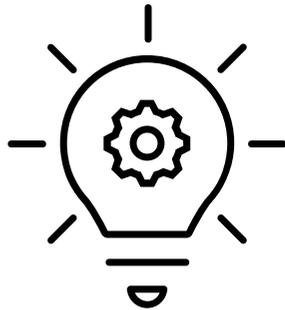
- Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Berlin: ScienceOpen, Inc.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), pp.15-21.
- Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17), 2503-2505.
- Garrison E, Marth G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.
- Hartwell, L., Goldberg, M.L., Fischer, J.A., Hood, L.E., Aquadro, C.F., Bejcek, B. (2015). *Genetics: from genes to genomes* (pp.). New York: McGraw-Hill.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20:1297-303. DOI: 10.1101/gr.107524.110.
- Molina, O., Abad, M.A., Solé, F. and Menéndez, P. (2020). Aneuploidy in Cancer: Lessons from Acute Lymphoblastic Leukemia. *Trends in Cancer*.
- Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1), 1-16.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.
- Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].
- Nossol, C., Barta-Böszörményi, A., Kahlert, S., Zuschratter, W., Faber-Zuschratter, H., Reinhardt, N., Ponsuksili, S., Wimmers, K., Diesing, A.K. and Rothkötter, H.J., (2015). Comparing two intestinal porcine epithelial cell lines (IPECs): morphological differentiation, function and metabolism. *PLoS one*, 10(7), p.e0132323.

Appendix A – Conference paper

Verma, A., Verma, M. and Singh, A., (2020). Animal tissue culture principles and applications. In *Animal Biotechnology* (pp. 269-293). Academic Press.

Acknowledgements: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998. Due to the restriction of the page number for this paper we were unable to include figures for all the results. The results not included can be obtained by contacting the corresponding author(s).

Summary



*'Hope is being able to see that there
is light despite all of the darkness'*

Desmond Tutu

The genome provides the entire set of DNA instructions of an organism, while the epigenome involves modifications that do not alter the DNA sequence. Gene expression is regulated by the intricate interplay between the genome and epigenome. We can explore the epigenome and regulatory elements through functional assays such as bisulfite sequencing, ChIP-seq, and ATAC-seq. Cell lines provide a valuable model for studying the genomic architecture and the regulatory genome of species. In vivo experiments are often complex and cell lines offer ethical alternatives.

Moreover, development is a dynamic and complex process which is regulated by gene expression. One epigenetic modification is DNA methylation, which plays an essential role in regulating development. The general dynamics of DNA methylation in the growing embryo and fetus, however, are still poorly understood. The aim of the research described in this thesis was two-fold, starting with understanding the (epi)genetic makeup of a pig and chicken cell line. The second aim was to disentangle the dynamic changes occurring within the epigenome, regulating gene expression during embryonic and fetal development in pig and chicken.

To do so, we first investigated the molecular characteristics of a pig intestine epithelial cell line, and chicken fibroblast cell line, using an integrative omics approach (**Chapter 2**). Functional assays such as ChIP-seq of histone modifications associated with regulatory elements, ATAC-seq indicative of open chromatin, reduced representation- and whole genome bisulfite sequencing (RRBS and WGBS) for DNA methylation, RNA-seq for transcriptome profiling and whole genome sequencing, were profiled both on an individual level as well as by integrating the various epigenetic modifications. This provided insights into the complex interactions between the genome and epigenetic modifications regulating the gene expression of these cell lines. Chromosomal abnormalities, copy-number variations, and aneuploidy, typical for a cell line, were identified for several chromosomes in both cell lines. Furthermore, higher gene expression for genes located on aneuploid chromosomes compared to diploid chromosomes was observed. Although these cell lines are described as nontumorigenic and untransformed, aneuploidy occurs not only due to the number of passages, but with each passage, the

Summary

probability of its occurrence within a cell increases. This can lead to the emergence of cells with a growth advantage, ultimately causing them to become the predominant cell type in the culture over time.

It is important for future studies to take note of these cell lines' features and it is recommended that researchers proceed cautiously when interpreting findings. Understanding the characteristics and (epi)genetic composition of these cell lines, help to improve our knowledge about their limitations and potential use as an *in vivo* research model.

A DNA methylation pipeline, specifically focussed on processing bisulfite sequencing data is described in **Chapter 3**, which was developed to include methylation calling, and visualisation of various methylation statistics. To gain further insights into methylation levels across various genomic characteristics, the processed data produced by this pipeline may be easily imported into visualization tools. The aim of this pipeline was to ensure reproducibility of the results generated within this thesis, specifically for the GENE-SWitCH project. This pipeline was implemented and used to analyse both the RRBS and WGBS data described in **Chapters 4 and 5** to investigate the dynamics of the DNA methylome that regulate gene expression during development.

Investigating this role of the epigenome and the functional regulatory genome was facilitated by identification of the methylome dynamics in seven tissues during embryonic and fetal development for both pig (**Chapter 4**) and chicken (**Chapter 5**). In both species we identified differentially methylated sites and regions, together with categories of methylation, which were combined for insight into the developmental methylation dynamics (**Chapter 4 and Chapter 5**). By integrating transcriptomic data with methylation data, we gained further insight into the dynamic nature of the methylome during fetal development. We identified tissue specific changes, and in the pig at early development (30 dpf) germ layer specific methylation characteristics are evident in lung, muscle, kidney, skin, and small intestine, which then transition to tissue specific methylation from 70 dpf until newborn stages (**Chapter 4**). In the

chicken we found enrichment of processes indicating development of brain and skin function from an early developmental stage (E8) (**Chapter 5**).

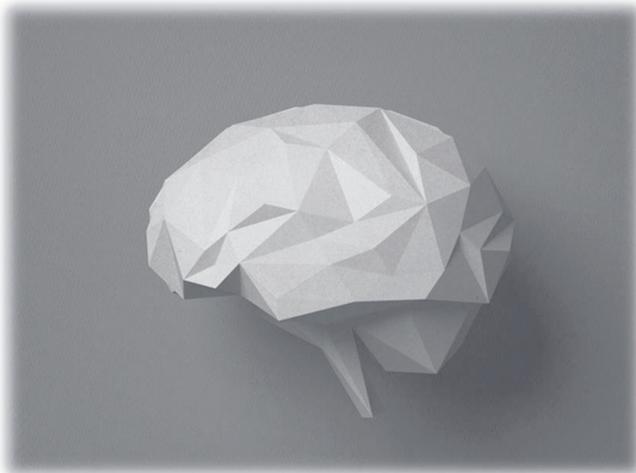
Remarkable differences in the methylome patterns between chicken and pig were observed, with the most notable differences found in the liver tissue during development. Liver is fully developed at an early stage of development in pig in comparison to chicken, which is due to the hematopoietic function of the liver in mammals during development in comparison to birds. This research provides fundamental insights into the methylome dynamics that regulate development and highlights key distinctions between avian and mammalian systems.

Curriculum Vitae

About the author

Contributions to conferences

Training and Education



'When one dream burns to ash, you don't crumble beneath it. You get on your hands and knees, and you sift through those ashes until you find the very last ember, the very last spark. Then you breathe. You breathe, and you make a new fire'

The Summer of chasing mermaids by Sarah Ockler

About the author

Jani de Vos's life journey began on 26 August 1994 in Roodepoort, South Africa. As a child of Africa, she was surrounded by animals from a very young age as she grew up with different pets and farm animals around her. She spent many hours on the family dairy farm with her dad, playing around and helping with the cattle (big and small) visiting auctions of livestock and generally being outdoors as much as possible. Growing up in SA, also meant being surrounded by wildlife in the bush, when visiting National Parks and enjoying an abundant sea life when visiting beaches.

While attending an academic high school, she selected Agricultural Science as an extra subject, and excelled in this, achieving top student in her province position in her final school year.

Because of her keen interest in animals, Jani enrolled in the Animal Science Bachelor program in 2013 and obtained her BSc (Animal Science) degree in 2016 at the University of Pretoria (South Africa). She then continued with her MSc, on the study of Nguni cattle, where she wrote her thesis on "A Genome wide association study of carcass traits based on Real Time Ultrasound in South African Nguni cattle".

After graduating in 2018, Jani left South Africa at the end of July 2019, to embark on an exciting new journey to continue with her PhD at Wageningen University. Here her research was also part of the GENE-SWitCH Consortium. She researched the functional regulatory genome regulating foetal development of chicken and pig. As a part of her PhD research, Jani visited the European Bioinformatics Institute (United Kingdom) for 3 months to establish a framework for doing integrative data analysis on functional genomics data, as well as comparative analysis. Since middle September she started work at Hendrix Genetics as Genomics coordinator.

Publications & Contributions to Conferences

- **de Vos J.**, Theron, H., & Van Marle-Köster, E., *RTU measurements as early indicator for carcass characteristics in Nguni cattle*. South African Society of Animal Science conference (Port Elizabeth, South Africa), 18-21 September 2017. Oral Presentation
- **de Vos, J.**, Theron H., & Van Marle-Köster E., *Selection index for prediction of dressing percentage in Nguni cattle using RTU measurements*. 69th EAAP annual meeting (Dubrovnik, Croatia), 27-31 August 2018. Poster Presentation
- **de Vos J.**, van Marle-Köster E., & Berry D.P., *Genome-wide association study of carcass quality using real-time ultrasound scans in South African Nguni cattle*. 36th International Society for Animal Genetics conference (ISAG) (Lleida, Spain), 7-12 July 2019. Poster presentation.
- **de Vos, J.**, Madsen O., Groenen M.A.M., *GENE-SWitCH methylation pipeline for expanded methylation analysis results*. Reproducible genomics workflows using Nextflow and nf-core workshop (Online), 2020. Oral Flash presentation
- **de Vos J.**, Crooijmans R.P.M.A., Derks M.F.L., Kloet S.L., Groenen M.A.M, Madsen O., *Resilience through knowledge: Molecular characterization of Pig IPECJ-2 and Chicken CRL cell lines*. Wageningen Institute of Animal Science annual conference (online). 28-29 April 2021. Oral presentation
- **de Vos J.**, Crooijmans R.P.M.A., Derks M.F.L., Kloet S.L., Groenen M.A.M, Madsen O., *Detailed molecular and epigenetic characterization of Pig IPECJ-2 and Chicken SL-29 cell lines*. 37th International Society of Animal Genetics conference (ISAG) (Online), 26-30 July 2021. Oral presentation.
- **de Vos J.**, *GSM-analysis pipeline*. GENE-SWitCH annual meeting (Online). 2021. Oral presentation.
- **de Vos J.**, *A glimpse into integrative analysis for GENE-SWitCH data*. GENE-SWitCH annual meeting (Spain). July **2022**. **Oral presentation**
- **de Vos J.**, Crooijmans R.P.M.A., Derks M.F.L., Kloet S.L., Groenen M.A.M., Madsen O., *Allele specific expression as an indication of ploidy in pig IPECJ2 and chicken SL-29 cell lines*. Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). July 2022 (Rotterdam, The Netherlands). Poster presentation
- Acloque H., Harrison P.W., Lakhal W., Martin F., Archibald A.L., Beinat M., Davey M., Djebali S., Foissac S., Guizard S., Guyomar C., Kuo R., Kurylo C., Madsen O., Miedzinska K., Mongellaz M., Smith J., Smith J., Sokolov A., **de Vos J.**, Giuffra E., Watson M., Extensive functional genomics information from early developmental time

points for pig and chicken. Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). July 2022 (Rotterdam, The Netherlands). Oral presentation

- **de Vos J.**, Groenen M.A.M, Madsen O., Exploring the Dynamic Nature of the Functional Regulatory Genome: Insights from Epigenomic States. **WIAS annual conference** (Ede, The Netherlands). 16 February 2023. Oral presentation
- **de Vos, J.**, Crooijmans, R. P., Derks, M. F., Kloet, S. L., Dibbits, B., Groenen, M. A., & Madsen, O. (2023). Detailed molecular and epigenetic characterization of the pig IPEC-J2 and chicken SL-29 cell lines. *Iscience*, 26(3). [https://www.cell.com/iscience/pdf/S2589-0042\(23\)00329-2.pdf](https://www.cell.com/iscience/pdf/S2589-0042(23)00329-2.pdf)
- **de Vos J.**, Derks M.F.L., Acloque H., Djebali S, Foissac S., Guyomar C, Kurylo C., Giuffra E., Groenen M.A.M., and Madsen O., *DNA methylation dynamics regulating embryonic development in pig*. 39th International Society of Animal Genetics conference (ISAG) (Cape Town, South Africa), 2-7 July 2021. Oral presentation. (Bursary winner)
- Calus M.P.L., Perez B.C., **de Vos J.**, Madsen O., Ayres L., Bovenhuis H., Ballester M., Mercat M.J., Bink M.C.A.M., Using functional annotation and individual omics in genomic prediction. 39th International Society of Animal Genetics conference (ISAG) (Cape Town, South Africa), 2-7 July 2021. Oral presentation.
- Schokker D., **de Vos J.**, Stege P.B., Madsen O., Wijnen H.J., Kar S.K., and Rebel J.M.J., *Elevated temperature induces transcriptional and epigenetic changes in the lungs of broiler chickens*. 74th EAAP annual meeting (Lyon, France), 26 August – September 1st 2023.
- Perez B., **de Vos J.**, Crespo-Piazuello D., Bink M., Ballester M., Mercat MJ., Madsen O., Calus M.. *Using methylation annotation to improve genomic prediction of gene expression phenotypes in pigs*. 74th EAAP annual meeting (Lyon, France), 26 August – September 1st 2023.
- Martín A., Madsen O., Giráldez F.J., **de Vos J.**, Andrés S., *The effect of methionine supplementation on DNA methylation during the suckling period of ewe lambs*. 74th EAAP annual meeting (Lyon, France), 26 August – September 1st 2023.

Training and education



The Basic Package (3.0 credits)	
WIAS Introduction Day	2019
Course on philosophy of science and/or ethics	2023
Course on essential skills	2023

Disciplinary Competences (17 credits)	
Project proposal	2020
FAANG shared workshop: Foundation for the future	2020
EuroFAANG GENE-SWitCH workshop: utilisation of GENE-SWitCH data in Ensembl and beyond	2023
Reproducible genomics workflows using Nextflow and nf-core workshop	2020
PhD proposal review (3 Proposals)	2020-2023
Practical computing for biologists	2019
Genomics	2029

Professional Competences (1.0 credits)	
The Final Touch: Writing the General Introduction and Discussion	2023

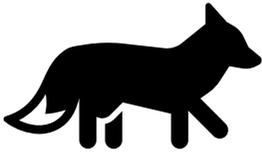
Societal relevance (2.0 credits)	
GENE-SWitCH young researcher newsletter	2021
GENE-SWitCH YouTube video - Young researcher interview	2021

Presentation Skills (4.0 credits)	
WIAS science day (oral)	2021 & 2023
ISAG (oral)	2021 & 2023
WCGALP (poster & oral)	2022
Reproducible genomics workflows using Nextflow and nf-core workshop (oral)	2020

Teaching competences (4.0 credits)	
Genomics course (ABG-30306)	2020-2022
Supervising MSc student	2022-2023

Total credits	31
----------------------	-----------

Acknowledgements



'If you can't fly, then run. If you can't run, then walk. If you can't walk, then crawl, but whatever you do, you have to keep moving'

Martin Luther King

The amount of people that enter your life during four years of a PhD and play even a small role, is too many to count. During this PhD journey, there have been some people who really stand out and I would like to take a moment to acknowledge the special part they have played.

Firstly, to my supervisory team, you made this PhD journey possible from picking me, up to the last few months (and weeks). I appreciated all the kind words, the motivation and feedback on all the drafts of this thesis. I would like to give a huge thanks to my promotor **Martien**. There are a few key moments that stand out to me: firstly, I will never forget the day you phoned me to offer me the PhD position which would change my life. Two other notable memories: the GENE-SWitCH kick-off meeting in Paris with Ole and Mario. And then ISAG this year (2023) in Cape Town was extremely special for me to share with you and many other colleagues. Naturally, next I want to thank my daily supervisor **Ole**. I already miss our meetings, during which we would discuss work, but also many life topics. We shared many fun meetings for GENE-SWitCH, starting in Paris, Cambridge (at the EBI) and then in Spain.

Everyone that was a part of the **GENE-SWitCH consortium**, I had the best four years and learnt a great deal from every person within this consortium. My fellow WP2 members, I already miss the monthly calls discussing results, deliverables, etc. I remember feeling very lost and unsure during my first few calls. **Sylvain, Sarah, Cyril, Cervin, Fergal, Peter, Herve, Sebastien, and Allan**, your feedback and the stimulating conversations about work are much appreciated and will always be remembered. GENE-SWitCH meetings in Paris, online, Spain and Rome were absolutely unforgettable with WP2, **Elisabetta, Yuliaxis, Maria, Fanny, Andrea, Marco, Camille, Mario and Bruno**.

The special people of the Genomics group, who would always listen and give feedback during my PhD. It went from being in person, to fully online during Covid to hybrid again. I would like to give a special mention to **Richard, Bert, Kimberley, Mirte, Hendrik-Jan, Zhou, Lim, Langqing, Henri, Xiaofei, Annemiek, Yun, Martijn, Marta, Rayner**, and many more. **Richard**, you were always a friendly face and would always have something positive to say, it was wonderful to enjoy (2!) ISAG conferences together.

To my corridor mates! It was lovely to share parts of the journey with a handful of you, in particular, **Marieke, Benan, Shuwen, and Jan-Erik**. Dearest **Fadma, Lisette and Marielle**, thank you for all the support in the years and for nice chats at the coffee corner or in the office!

Friends I always think of in two categories: the OG 2019 (LON) friends and my post-Covid 'new' friends. My OG 2019 friends, WOW! **Maria and Anne**, you two became my two dearest friends so far away from home. Thanks for many hugs, deep conversations, gin & tonics, coffees, gym and parties. **Anne**, I will never forget knowing you only for a short time and then travelling to Ghent together. **Maria**, amor, to name all the memories would fill many pages, but my top memory is visiting you in Switzerland and enjoying hikes and many talks. **Pascal**, thanks for many coffees, beers, parties, dinners, barbeques, and insightful conversations. **Renzo!** Life would be oh so dull without you. We have really known each other through many circumstances, and I will never forget going for a walk one evening near the end of my PhD and you giving some words of support; and of course your amazing Carbonara! There were others of course: **Tom, Joaquin, Zhou, Gibbs. Zhou**, visiting you in Edinburgh was such a fun time, even when you forced me to try Haggis. Gibbs, my friend who has the shared love for South Africa, Mugg & Bean, and Woolworths. You brought me rusks and biscuits from SA when I most needed it.

The post-Covid 'new' friends! **Mirjam, Jan-Erik, Margot, Matias, Larissa. Mirjam**, we became such good friends enjoying tea, discussions on running, and exercise. **Margot**, my dear! What a wonderful friend you are, and our shared love for our pups and clothes has made this a great friendship. On the topic of pups, **Larissa**, we became much greater friends later in the journey and I have enjoyed our chats, your wise words, barbeques and walks. **Jan-Erik**, we have enjoyed many chats at the office, dinners and after-work drinks. **Matias**, it has been fun to be neighbours who can share many dinners and quick after work drinks! **Pedro**, you have become my friend to "spill the tea" with, and I think that says enough about the importance of our friendship.

During this PhD I also had the wonderful opportunity to travel to the UK and complete my research visit at the EBI in Hinxton Cambridge. The friends I made there are so special and the memories will forever be deep in my heart. **Maša**, you were such an incredible mentor during my time there and I will forever be grateful to you for all the lessons learnt. The girls, **Maelle and Martina**, we had such fun in the office, talking so much, and shopping in Cambridge. The best was our brunch the morning I returned to the Netherlands. **Peter and Fergal**, you were both a part of GENE-SWitCH but also worked at the EBI. I think of you both as good friends now; having shared coffee, conferences, GS meetings, especially the GS meeting in Spain!

My two paranymphs, **Marta and Martijn**. You have both played such different and unique roles in my life. **Martijn**, we have known one another from (almost) day one and we have collaborated a lot in work, shared many coffees, drinks after work, your defence, and visits with your sweet kiddos and Liesanne. **Marta**, what a dear friend you have become and we have enjoyed lunch/dinners, coffees at work and most recently the most unforgettable time in South Africa, together with **Manu** (thank you for being a great friend also!!).

Henk en Lorraine, julle het familie so ver van die huis geword. Ek is ewig dankbaar vir so baie adventures, bank bou, etes, en natuurlik braai. Lorraine, jy is n vriendin wat goud werd is en ons koffies, shopping en tulpe is van die beste herinneringe!

There is a South African saying: *It takes a village to raise a child* and just like that, there was a special village who supported me on a very personal level throughout my journey thus far. To all of you I want to thank you for supporting me through all decisions and life stages.

Rulien, wie sou ooit gedink het hierdie dag sal aanbreek?! Jy was deur al die tye saam met my van die oproep dat ek aanvaar is vir die PhD, dan ISAG in Spanje, die heerlike kuier in Nederland, en eintlik te veel tye! Liefste **Em(mouse)**, **Emma**, my sussie, dankie vir wie jy is en veral die kuintjie in 2019 wat net onvergeetlik was. Ons dag in Barcelona saam met Rulien is een van my beste memories saam julle albei. Ons tydjie in Brugge was ook onvergeetlik.

Leo en Oom Heiko, dankie vir al die omgee, en dat julle my nie-biologiese familie geword het. My liewe **Prof**, dankie vir alles; die mentorskap, diep gesprekke, in my glo, my help droom, die beste mens wees. Daar is net te veel herinneringe, maar my gunsteling, is my kuier in Stilbaai en al ons koffies! **Dr. Visser**, dankie vir al die mentorskap, etes, koffies en onlangse heerlike geselses by WCGALP in Rotterdam.

Then my family. **Neil and Tam, Kieran, Caily, and Kels**, during my PhD, it was great that I was able to visit you several times in the UK and you were both so supportive – especially during Covid times and Neil, with your visit in my first week in the Netherlands! **Wend, Pete, Carrie and James**, thanks for all the interest, advice and support. **Ams and Arran (and now Harley)**, thank you for your words of wisdom, for always being there when I need someone and just for being you. **Taryn and Rob**, probably my biggest cheerleaders! Now also with **Harry and Caity**. You have both been there during the best moments and during more difficult times. I have always felt like we are a small crew! **Marina & Pete**, the best God-parents any girl could hope for! Your support from when I first arrived and when we met in the UK, all the way through until the end has been incredible. Thanks for all the catch-ups, coffees (in SA) and just generally always being there for whatever happens!

Mamma en Pappa. Om boeke vol oor julle te skryf, sal nooit genoeg wees nie. Dankie vir ontelbare oproepe, drukkies, virtual koffies, glasies wyn, en net die beste kuiers wanneer ons saam is. Mams dankie dat ek enige tyd kan bel en jy sal luister en n wyse woord gee. Pappa, dankie dat jy altyd reg is om te help en enige probleem net te fix. Sonder julle ondersteuning en liefde sal ek nie hier wees nie.

Rayner, my love. You were there with each happy and difficult moment during my PhD. You saw me feeling all the emotions; from writing my first paper, celebrating publishing, disappointments, and thesis writing. Your support and hugs made the difficult days possible to face. I also love the many adventures we have had during this journey, coffees at the office and at home, dinners and walks. You also helped in finding Ash, my loveliest doggy!

Lastly and most importantly, I could never be here without the most important, my G. I will always remain thankful in all I do.

“Do not judge me by my success, but judge me by how many times I fell and got back up again” – Nelson Mandela

Colophon

This research was funded by the GENE-SWitCH project which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998. The research visit of Jani de Vos to the European Bioinformatics Institute (Hinxton, UK) was financially supported by WIAS.

The cover was designed by Mariska Duenk.

Printed by DigiForce | Proefschriftmaken.nl

