ABSTRACT

Title of Document: GENOME-WIDE ANALYSIS OF HISTONE MODIFICATION ENRICHMENTS INDUCED BY MAREK'S DISEASE VIRUS IN INBRED CHICKEN LINES

Apratim Kumar Mitra,

Doctor of Philosophy, 2013

Directed By: Associate Professor Dr. Jiuzhou Song,

Department of Animal and Avian Sciences

Covalent histone modifications constitute a complex network of transcriptional regulation involved in diverse biological processes ranging from stem cell differentiation to immune response. The advent of modern sequencing technologies enables one to query the locations of histone modifications across the genome in an efficient manner. However, inherent biases in the technology and diverse enrichment patterns complicate data analysis. Marek's disease (MD) is an acute, lymphoma-inducing disease of chickens with disease outcomes affected by multiple host and environmental factors. Inbred chicken lines 6_3 and 7_2 share the same major

histocompatibility complex haplotype, but have contrasting responses to MD. This dissertation presents novel methods for analysis of genome-wide histone modification data and application of new and existing methods to the investigation of epigenetic effects of MD on these lines. First, we present WaveSeq, a novel algorithm for detection of significant enrichments in ChIP-Seq data. WaveSeq implements a distribution-free approach by combining the continuous wavelet transform with Monte Carlo sampling techniques for effective peak detection. WaveSeq outperformed existing tools particularly for diffuse histone modification peaks demonstrating that restrictive distributional assumptions are not necessary for accurate ChIP-Seq peak detection. Second, we investigated latent MD in thymus tissues by profiling H3K4me3 and H3K27me3 in infected and control birds from lines 6_3 and 7_2. Several genes associated with MD, e.g. *MX1* and *CTLA-4*, along with those linked with human cancers, showed line-specific and condition-specific enrichments. One of the first studies of histone modifications in chickens, our work demonstrated that MD induced widespread epigenetic variations. Finally, we analyzed the temporal evolution of histone modifications at distinct phases of MD progression in the bursa of Fabricius. Genes involved in several important pathways, e.g. apoptosis and MAPK signaling, and various immune-related miRNAs showed differential histone modifications in the promoter region. Our results indicated heightened inflammation in the susceptible line during early cytolytic MD, while resistant birds showed recuperative symptoms during early MD and epigenetic silencing during latent infection. Thus, although further elucidation of underlying

mechanisms is necessary, this work provided the first definitive evidence of the epigenetic effects of MD.

GENOME-WIDE ANALYSIS OF HISTONE MODIFICATION ENRICHMENTS INDUCED BY MAREK'S DISEASE VIRUS IN INBRED CHICKEN LINES

By

Apratim Kumar Mitra

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:

Associate Professor Dr. Jiuzhou Song, Chair

Professor Dr. Richard Kohn

Associate Professor Dr. Sridhar Hannenhalli

Research Geneticist Dr. Curtis P. Van Tassell

Professor Dr. Yang Tao

Preface

The three chapters in this thesis have either appeared in peer-reviewed publications or are currently in preparation for publication. I am grateful to each of my co-authors on these studies for their respective contributions, which resulted in significant improvement in the quality of research and writing. At the time of this writing, Chapters 2 and 3 have appeared in print and have been edited slightly, while Chapter 4 is in preparation.

Chapter 2:

Mitra A. and J. Song (2012) ***WaveSeq: A Novel Data-Driven Method of Detecting Histone Modification Enrichments Using Wavelets*****. PLoS ONE, 7(9): e45486. doi:10.1371/journal.pone.0045486.**

We would like to acknowledge Dr. Juan Luo who performed the ChIP experiments in chicken bursal tissues and prepared the H3K4me3 sequencing libraries.

Chapter 3:

Mitra A., Luo J., Zhang H., Cui K., Zhao K. and J. Song (2012) ***Marek's Disease Virus Induces Widespread Differential Chromatin Marks in Inbred Chicken Lines.*** **BMC Genomics, 13:557. doi:10.1186/1471-2164-13-557.**

J.S. was supported by National Research Initiative Competitive Grant no. USDA-NRI/NIFA 2008-35204-04660 and USDA-NRI/NIFA 2010-65205-20588 from the USDA National Institute of Food and Agriculture.

Chapter 4:

Mitra A., Luo J., Gu Y., Zhang H. and J. Song, ***Temporal Chromatin Signatures Induced by Marek's Disease Virus Infection in Bursa of Fabricius*****. In preparation.**

Dedication

I dedicate this work to my parents,

Without whose unswerving support,

I would not be here today.

Acknowledgements

I would first like to acknowledge the support of my advisor, Dr. Jiuzhou Song, whose constant encouragement and guidance enabled me to accomplish this work. The creative freedom fostered by him has greatly shaped my thinking and can go a long way in helping me in the future as an independent researcher. The healthy interchange of ideas between us was invaluable in boosting my confidence in my scientific abilities and is greatly appreciated. Thank you.

I would like to express my gratitude to my advisory committee, Dr. Curt Van Tassell, Dr. Rick Kohn, Dr. Sridhar Hannenhalli and Dr. Yang Tao, whose kind words of encouragement and critical scientific input helped me at various points along the way. I would also like to thank Dr. Mihai Pop and Dr. Carl Kingsford, who taught me the important concepts of sequence alignment and graph algorithms, in the process helping me bridge the gap between my applied mathematics training and the current needs of computational biology. I would like to thank Dr. George Liu, who helped get me started on this journey and whose ever-smiling demeanor was a constant reassurance. Thank you.

I would not be here if it were not for the support of my current and former colleagues, Dr. Fei Tian, Dr. Juan Luo, Jose Carrillo, Fei Zhan, Yanghua He, Ding Yi, Yulan Gu, Chunping Zhao, Dr. Ying Yu, Dr. Yali Hou and Ping Yuan. Fei, you were my ‘partner-in-crime’, friend and confidante, and you have led the way for me in so many different ways. Juan, we shared many memorable times working together and your help is much appreciated. Thank You.

During my graduate study I have been fortunate to meet many wonderful people many friends who have all helped and inspired me. Andy, Jason, Ashley, Lindsey, Laura, Katy, Xuan, Kara and all those I have not mentioned, you have all contributed towards making this a success and I hope to remain in touch. Dyan, you and I have shared a lot over the years and your successful defense made me believe that this was indeed possible. Thank you.

Finally, I would like to thank my family. Dad, Mom, Bhaiti and Didu, you have been my rock through all these years, and you continue to be. You lent a patient ear when times were tough and your encouragement, love and appreciation spurred me on. Pratyusha, we shared the highlights and frustrations of my graduate study together. You have been so patient and understanding through the whole process, and been a great part of my successful completion of this work. Thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

1. Literature Review

Introduction

The term ‘epigenetics’ can be loosely defined as the study of changes in the phenotype of an individual caused by mechanisms other than underlying DNA sequence. One of the first indications that there was more to gene regulation than DNA sequence was the discovery of histone modifications and their possible effects on transcriptional regulation [1]. The involvement of DNA methylation in various regulatory functions [2, 3] further confirmed the presence of significant epigenetic mechanisms in transcriptional control. Subsequent studies have shown that epigenetic mechanisms are associated with a multitude of critical biological processes, such as, X chromosome inactivation, stem cell differentiation and immune response. The advent of next-generation sequencing technology has revolutionized the field, making it possible to investigate histone modification profiles in a genome-wide manner. However, the enormity of associated data sets has posed new challenges in data analysis and interpretation that are far from being solved.

Epigenetic processes play major roles in various human diseases. Cancer cells demonstrate major variations in DNA methylation, e.g. large-scale demethylation in tumor cells is concurrent with hypermethylation at specific promoters [2, 3]. Histone modification changes are observed in conjunction with aberrant DNA methylation in various cancers [4, 5]. However, further study has suggested that variations in histone modifications are important prognostic markers for cancer [6-8]. Recent studies have also shown that histone modifications can interact with and regulate viral processes

[9]. Herpesviruses, in particular, appear to be affected by cellular chromatin machinery. For instance, a transcriptional activator HCF-1 (host cell factor 1), which is associated with several chromatin-modifying enzymes [10], controls the early transcriptional program of the herpes simplex virus [11]. Kaposi's sarcoma-associated herpesvirus (KSHV) exhibits increased activating and repressive histone marks during latent infection [12]. Thus, histone modifications are epigenetic indicators of the adverse effect of various diseases, and further study is necessary to delineate their particular roles in the process.

Histone Modifications

DNA is packaged in the form of chromatin, with the DNA double helix wound around an octamer of four core histone proteins (H3, H4, H2A and H2B). A 147 nucleotide-long fragment of DNA, together with the histone proteins it is wrapped around, constitutes the nucleosome, the fundamental unit of chromatin. Eukaryotic nucleosomes also contain lower levels of histone variants with specialized functions, e.g. the histone variant H2A.Z, which occurs within nucleosomes adjacent to the transcription start site (TSS) of genes [13]. Chromatin can be structurally and functionally separated into two forms: euchromatin and heterochromatin. Euchromatin is conformationally open, relatively rich in genes, and conducive to active transcription, while the highly-condensed heterochromatin is relatively inaccessible to transcription factors and hence, constitutively silent [14].

DNA exists primarily in the form of heterochromatin during certain cellular processes such as mitosis and meiosis which lack DNA regulatory activity [15]. On the other hand, the loose conformation of euchromatin allows the dynamic control of

transcription, with various activating and silencing mechanisms at play. However, in spite of the relatively low density of euchromatin, it is still refractory to essential cellular processes and must be relaxed for easier access by the transcriptional machinery. This need has resulted in the evolution of a wide array of chromatin-modifying mechanisms, including chromatin remodeling, an ATP-dependent process which alters the structure, composition and position of nucleosomes, and covalent post-translational modification of histones by particular enzymes.

Histone modifications occur primarily on the unstructured N-terminal tails of histone proteins, which contain several residues that are subject to various modifications, such as, methylation, acetylation and phosphorylation. Eight classes of histone modifications occurring on over 60 different residues have been discovered, with histone methylation and acetylation the two most common and well-studied modifications. Several histone marks have been associated with regulatory roles such as transcription, replication and DNA repair [16]. For instance, the trimethylation of the lysine residue at the fourth position of the histone H3 (H3K4me3) is associated with the TSS of active genes [17], while the trimethylation of lysine 36 (H3K36me3) is found on exons and introns of actively transcribing genes [18]. On the other hand, certain modifications are associated with gene silencing. These changes include H3K9 trimethylation, which is highly associated with heterochromatin, and H3K27me3, which is associated with the chromatin-modifying Polycomb repressive complexes (PRCs) [19]. Multiple histone modifications with seemingly contrasting functions have been observed on the same gene. For example, certain key developmental genes display both the active H3K4me3 and repressive H3K27me3

marks in embryonic stem cells (ESCs) suggesting a possible 'bivalence' depending upon lineage-determination [20]. In some cases, different modifications of the same histone residue perform contrasting functions. For instance, in T helper cells, H3K9 trimethylation and acetylation mark the promoters of repressed and active genes, respectively [21]. Certain lysine and arginine residues can also display varying levels of methylation (mono-, di- or tri- in the case of lysines and mono- or di- in the case of arginines), which, in turn, could be associated with different functions. For example, H3K4me3, as mentioned above, is associated with the promoters of active genes while H3K4me1 is highly enriched on promoter-distal enhancers [22]. Thus, histone modifications encode tremendous diversity into the genome and their dynamic nature plays major roles in a wide range of biological processes ranging from development to disease response. Also, owing to the diversity of function encompassed by histone methylation (activation, repression, transcription elongation and enhancers), we discuss this class of modifications in greater detail throughout this review.

Mechanisms of Formation

The majority of histone modifications are dynamic. A class of enzymes, called the histone-modifying enzymes, catalyzes the addition or removal of specific modifications from histone proteins. A host of such enzymes have been identified recently [23-31]. For instance, histone methylation can occur on lysines and arginines and is carried out by three classes of enzymes:

1) Histone methyltransferases (HMTs), which contain the lysine-specific SET (Su(var)3-9, Enhancer of Zeste [E(Z)], and Trithorax) domain, and can

methylate lysines 4, 9, 27 and 36 of histone H3 and lysine 20 of histone H4 [23, 24].

2) Non-SET domain-containing HMTs methylate the lysine 79 of histone H3 and consist of the evolutionarily conserved protein Dot1 (disrupter of telomeric silencing, also known as Kmt4) [25].
3) Protein arginine methyltransferases (PRMTs) methylate arginines 2, 17 and 26 of histone H3 and also arginine 3 of histone H4 [26].

Similarly, enzymes that remove methyl groups from lysine residues of histone proteins have also been the subjects of great interest. An amine oxidase, lysine-specific histone demethylase 1 (LSD1), was the first protein found to possess histone demethylase activity. LSD1 primarily demethylates H3K4 [27], but can also target H3K9 when complexed with the androgen receptor [28]. However, the enzymatic action of LSD1 requires the presence of a protonated methyl-ammonium group, and therefore, it can only demethylate mono- and dimethylated lysines. Two years after the discovery of LSD1, a class of proteins containing the Jumonji C (JmjC) catalytic domain was discovered and shown to demethylate trimethylated lysines [29]. Indeed, the demethyase activity of JmjC-containing enzymes is amenable to mono-, di- and trimethylated lysines but appears to favour trimethylated residues [30]. Moreover, JmjC proteins have also been shown to demethylate arginine residues [31]. Out of 27 known members of the Jumonji family about 15 possess demethylase activity, further emphasizing the importance of this family of chromatin-modifying enzymes.

Other major histone-modifying enzymes include the activating histone acetyltransferases (HATs) and repressive histone deacetylases (HDACs). Similar to

HMTs, HATs and HDACs have been the subject of intense study leading to the discovery of a large number of members of each class. The addition of the ubiquitin moiety (ubiquitylation) is carried out by members of an enzymatic pathway including ubiquitin activating (E1), conjugating (E2) and ligase (E3) enzymes [32]. E2 and E3 enzymes largely determine the specificity of the modification [33]. SUMOylation consists of the addition of a small ubiquitin-related modifier (SUMO) protein by the E1-E2-E3 enzymes which can also be removed by specific proteases [34].

Biological Functions

Covalent histone modifications can function in two major ways. First, they can affect the high-level interaction between neighboring nucleosomes or between the DNA and chromatin, leading to the 'unraveling' of nucleosomes. For instance, histone acetylation has a strong activating effect as it neutralizes the basic charge on the lysine residues producing electrostatic repulsive forces between the histone protein and the negatively charged DNA. The second and better characterized mode of function for histone modifications is the recruitment of non-histone proteins in what is believed to be a highly ordered and coordinated manner. For example, the addition of a methyl group does not affect the charge on the histone protein and hence, has no effect on chromatin-DNA interactions. However, methylation of specific residues exhibits affinity towards particular proteins (which can act as either activators or repressors), thereby influencing the transcriptional regulation of underlying DNA.

Ubiquitylation and SUMOylation involve the addition of large covalent groups to the chromatin, which can affect chromatin structure via steric effects. SUMOylation is primarily repressive, interfering with activating marks like acetylation by the

recruitment of HDACs [35]. However, histone ubiquitylation, like methylation, can have diverse outcomes. For example, addition of multiple ubiquitin groups marks a protein for proteasomal degradation, while monoubiquitylation alters protein function. However, the latter modification can produce different effects on histones, e.g. lysine residues in the C-termini of H2A and H2B correlate with activation and repression, respectively.

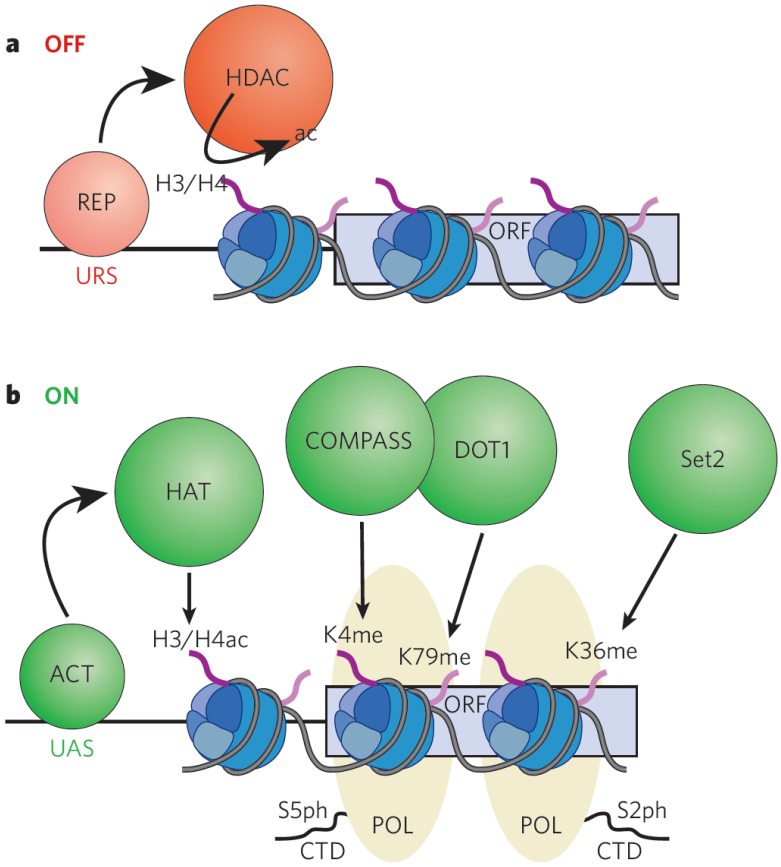


Figure 1.1 Chromatin model of transcriptional regulation (from [36]).
(a) Silenced state: Repressor factors (REP) bound at an upstream repressor site (URS) recruit negative modifiers like histone deacetylase (HDAC) which removes the acetyl group from histone H3/H4. (b) Active state: Activating transcription factors (ACT) bound at an upstream activation site (UAS), induces H3/H4 acetylation by HATs in the promoter region, while RNA polymerase (POL) induces methylation at lysine 4 by SET1 (part of the COMPASS complex) and lysine 79 by DOT1. Later, the POL recruits SET2, which induces methylation of lysine 36 during elongation.

Non-histone proteins bind to specific histone residues with the help of particular protein domains, e.g. methylation is bound by proteins containing domains of the Royal family similar to chromo-domains, and distinct PHD domains, while acetylation is bound by bromodomains. A schematic model of chromatin regulation of transcriptional is shown in Figure 1.1 [24]. Briefly, activating histone marks appear on gene promoters and transcription start sites in response to cellular stimulus, through recruitment of enzymes by activating transcription factors and RNA polymerase. In contrast, repressive marks are established through the action of DNA-bound repressors or heterochromatic regions. Since a major focus of our studies has been histone methylations, in particular trimethylations of H3K4 and H3K27, it is worthwhile to examine their mechanisms of action in some detail.

H3K4 and H3K27 Methylation

The positive correlation between H3K4 methylation and active genes suggested that this histone modification attracts activating factors for binding. This was proved to be true by the discovery of several such proteins including chromatin-remodeling enzyme CHD1 [37], nucleosome remodeling factor (NURF) [38] and PHD domain-containing Yng1 protein in the NuA3 (nucleosomal acetyltransferase of histone H3) [39]. The latter two are specific to H3K4me3, while CHD1 recognizes either di- or trimethylated H3K4. Surprisingly, H3K4 methylation also associates with repressive protein complexes. The Sin3-Hdac1 complex, which functions as a deacetylase, binds to H3K4me3, thereby stabilizing its recruitment to target genes and leading to the repression of proliferation-inducing genes in response to DNA damage [40]. Also,

H3K4me3 is believed to recruit the lysine demethylase JMJD2A, which demethylates H3K9me3 and H3K36me3 and causes gene repression [41]. Thus, H3K4 methylation is apparently context-specific and can lead to varying outcomes in terms of transcriptional control.

The methylation of H3K27, however, is undoubtedly repressive in nature, and was found to be associated with Polycomb-group (PcG) silencing [19]. Polycomb group proteins, discovered in *Drosophila melanogaster,* are repressors of homeobox (Hox) genes, transcription factors crucial to the determination of cell fate during embryonic development. These proteins are essential for maintenance of the transcriptional status of Hox genes after initial developmental cues, and bind to regulatory elements called Polycomb repressive elements (PREs) [42]. Subsequent studies identified the roles of PcG proteins in diverse biological contexts, such as, X chromosome inactivation [43], cell proliferation [44] and cancer [45], in vertebrates, plants and mammals. One of the Polycomb repressive complexes, PRC2, methylates H3K27, and subsequently, this histone mark is recognized by PRC1, which results in gene silencing. However, the mechanism of H3K27me3 and PRC1-mediated silencing is still unclear, as PRC1 is not found in several organisms, such as plants [46]. H3K27me3 is also found on broad swathes of the genome, which is believed to be the key to epigenetic inheritance of PcG silencing.

Another intriguing subplot to the functional consequences of H3K4 and H3K27 methylation is the interplay of the respective methylating protein complexes. H3K4 methylation is carried out by proteins belonging to the Trithorax group (TxG), which appear to act in an opposing manner to the PcG complexes and hence, contribute to

the determination of transcriptional fate [47, 48]. In summary, these two histone marks encompass a remarkable diversity of function in a variety of biological contexts, but further study is necessary for a clearer understanding of the associated regulatory mechanisms.

Detection using Chromatin Immunoprecipitation (ChIP)

The functional importance of histone modifications made it important to develop assays that could pinpoint the genomic locations of particular histone marks. Chromatin immunoprecipitation (ChIP) is a robust technique of studying DNA-protein interactions [49]. Originally developed to study the association of RNA polymerase and active genes in bacteria [50], this method has been subsequently used across a wide range of organisms, including *Drosophila* [51] and humans [52]. In brief, ChIP involves the use of a crosslinking agent, to preserve protein-DNA interactions, either irreversibly by ultraviolet radiation [53] or reversibly by formaldehyde [54]. Shearing via sonication or restriction enzyme digestion follows crosslinking and subsequently, antibodies specific to the protein of interest, e.g. modified histones, are used to immunoprecipitate the cross-linked protein-DNA complexes. The precipitated products are purified, the crosslinks reversed and DNA fragments analyzed using Southern blot or polymerase chain reaction (PCR). If so desired, the ChIP experiment could also be performed without a crosslinking step (native ChIP) to assay stable DNA-protein interactions. While highly specific and robust, this technique is only suitable for analyzing known regions of interest at a limited number of loci. Therefore, efforts were made to extend this method to genome-wide analyses.

The development of microarray technology provided a significant advance [55]. Microarrays consist of several thousand oligonucleotide sequences or 'probes', chosen to complement specific genomic regions, attached to a solid surface (Affymetrix or Agilent) or microscopic beads (Illumina). The test sample is fluorescently labeled before hybridization to the microarray and laser scanning. The fluorescent intensity at each spot of the microarray is assumed to be proportional to the number of molecules hybridizing to the probe specific to the spot, and provides a measure of the representation of the associated genomic region in the test sample. Thus, microarrays simultaneously query several thousand loci across the genome and, when combined with ChIP (ChIP-on-chip), vastly increases throughput [52]. However, this technique suffers from certain drawbacks. Microarrays and other fluorescence-based detection systems have a fixed dynamic range, with reduced sensitivity at upper and lower extremes of detectable signal amplitudes [56]. ChIP-on-chip depends on the availability of a suitable microarray for performing the experiment, e.g. a high-density tiling array, which consists of overlapping probes placed at a fixed distance from each other. While the possible resolution is high, so is the cost, due to the necessity of biological replicates and multiple arrays for large genome sizes. The design of the tiling array depends on a high-quality genome assembly and hence, has reduced accuracy and limited applicability for non-traditional model organisms. Also, repetitive regions are usually not represented on ChIP-on-chip arrays, and a relatively high amount of ChIP DNA (~several μg) is required.

Combination with Next-Generation Sequencing

The advent of modern sequencing technologies was the next big step forward. Next-generation sequencing (NGS) enables one to obtain the DNA sequence of millions of short fragments or reads from across the genome in a massively parallel manner. There are multiple such sequencers currently available, but since we used the Illumina sequencers for our experiments, I will discuss their experimental workflow in some detail (Figure 1.2).

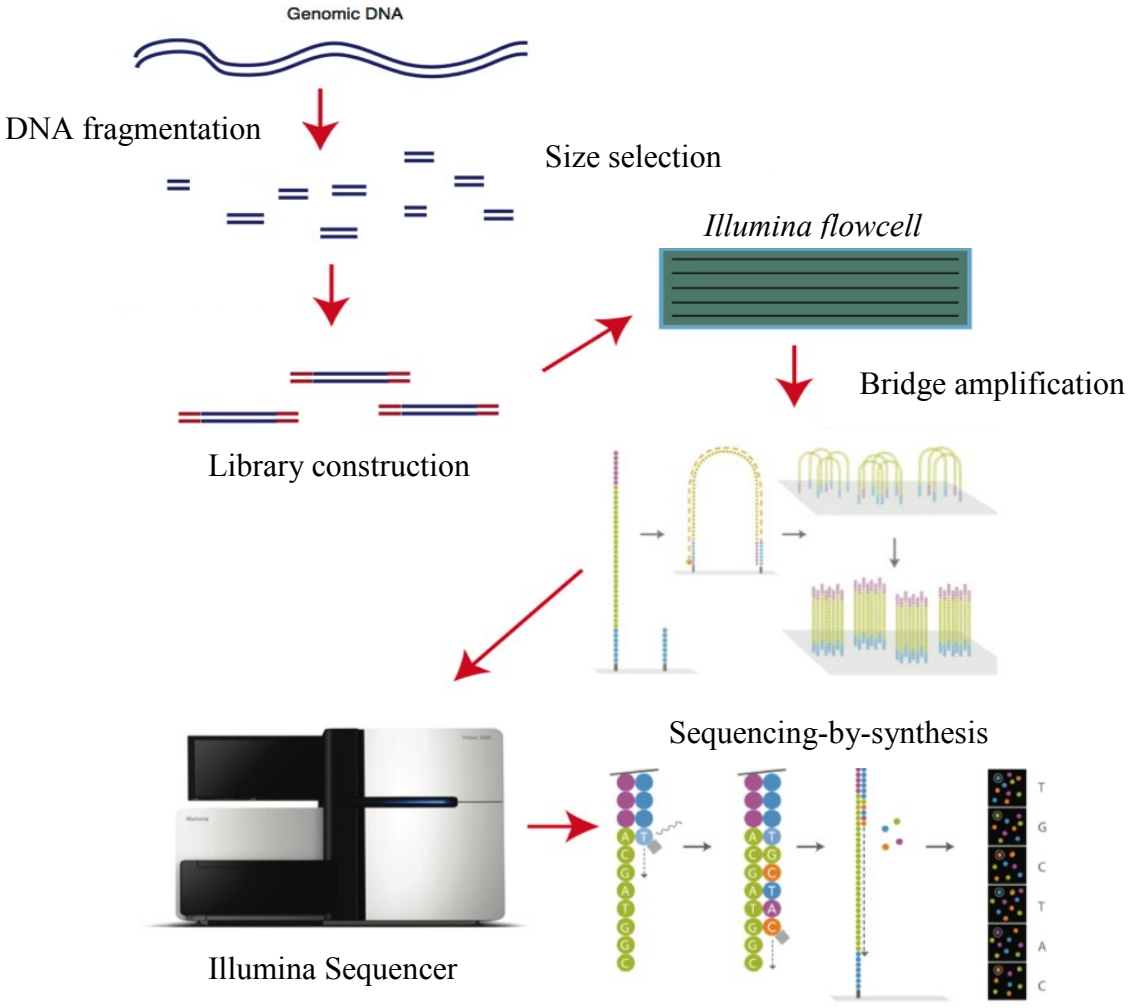


Figure 1.2 Overview of Illumina Sequencing protocol (modified from [57]).
The major steps include fragmentation, size-selection, library construction, bridge amplification and sequencing-by-synthesis.

The Illumina sequencing protocol utilizes reversible DNA terminators in a sequencing-by-synthesis procedure. Initial library preparation involves the repair of DNA fragment-pairs, attachment of an Adenine overhang and ligation of Illumina adaptors. The DNA sample is amplified using PCR to ensure enough starting material and fragments of suitable size are selected from the amplified sample. The next step, called cluster generation, involves use of a 'flowcell', a glass surface with eight channels, each containing adhered adaptors, complementary to those attached to the DNA fragments during library preparation. The size-selected DNA fragments are hybridized to the flowcell and extended by polymerases. Subsequently, the double-stranded DNA is denatured and the original template washed away. The free ends of the DNA molecules randomly attach to neighboring complementary adaptors, forming a 'bridge', and the extension procedure is repeated. This so-called 'bridge-PCR' step, thus, effectively amplifies several million DNA fragments in parallel, and the iteration of hybridization and bridge-PCR results in 'clusters' containing forward and reverse DNA fragments. The reverse strands are cleaved and washed away, leaving several million clusters spread across the surface of the flowcell, each containing approximately 1000 identical DNA fragments. In the final step, sequencing primers are attached to the free ends of the clustered DNA strands, and four fluorescently-labeled NTP terminators and polymerases are added to the reaction mixture. Each cluster incorporates a fluorescent NTP terminator, which represents the corresponding complementary nucleotide of its constituent DNA strands on the detected image. The terminator and fluorescent groups are subsequently cleaved and the sequencing reaction repeated a desired number of cycles.

The above process is highly efficient, dramatically reducing the overall cost of sequencing experiments, leading to a wide array of applications including whole-genome sequencing [58], transcriptomics [59, 60], structural variant detection [61], epigenomics [62-64] and metagenomics [65, 66]. The combination of ChIP and next-generation sequencing, called ChIP-Seq, resulted in a powerful new experimental technique of detecting genome-wide histone modification profiles [67-70]. ChIP-Seq offers several advantages over the microarray-based ChIP-on-chip. ChIP-Seq protocols typically require lower amounts of starting DNA (~ng range) and amplification. Several repetitive regions can be assayed, particularly with longer reads and paired-end sequencing. ChIP-Seq does not suffer from a fixed dynamic range and has single nucleotide resolution. Moreover, with the improvement of associated technologies, sequencing yield has dramatically increased, allowing the use of multiplexing, a technique whereby multiple 'bar-coded' samples can be sequenced simultaneously. For instance, the Illumina Genome Analyzer II.x yields a maximum of 40 million reads per lane, while the newer HiSeq 2000 can generate up to 187 million reads per lane at a comparable cost. Thus, four lanes of the earlier platform could be replaced by just one lane of the latter reducing the sequencing costs by 1/4.

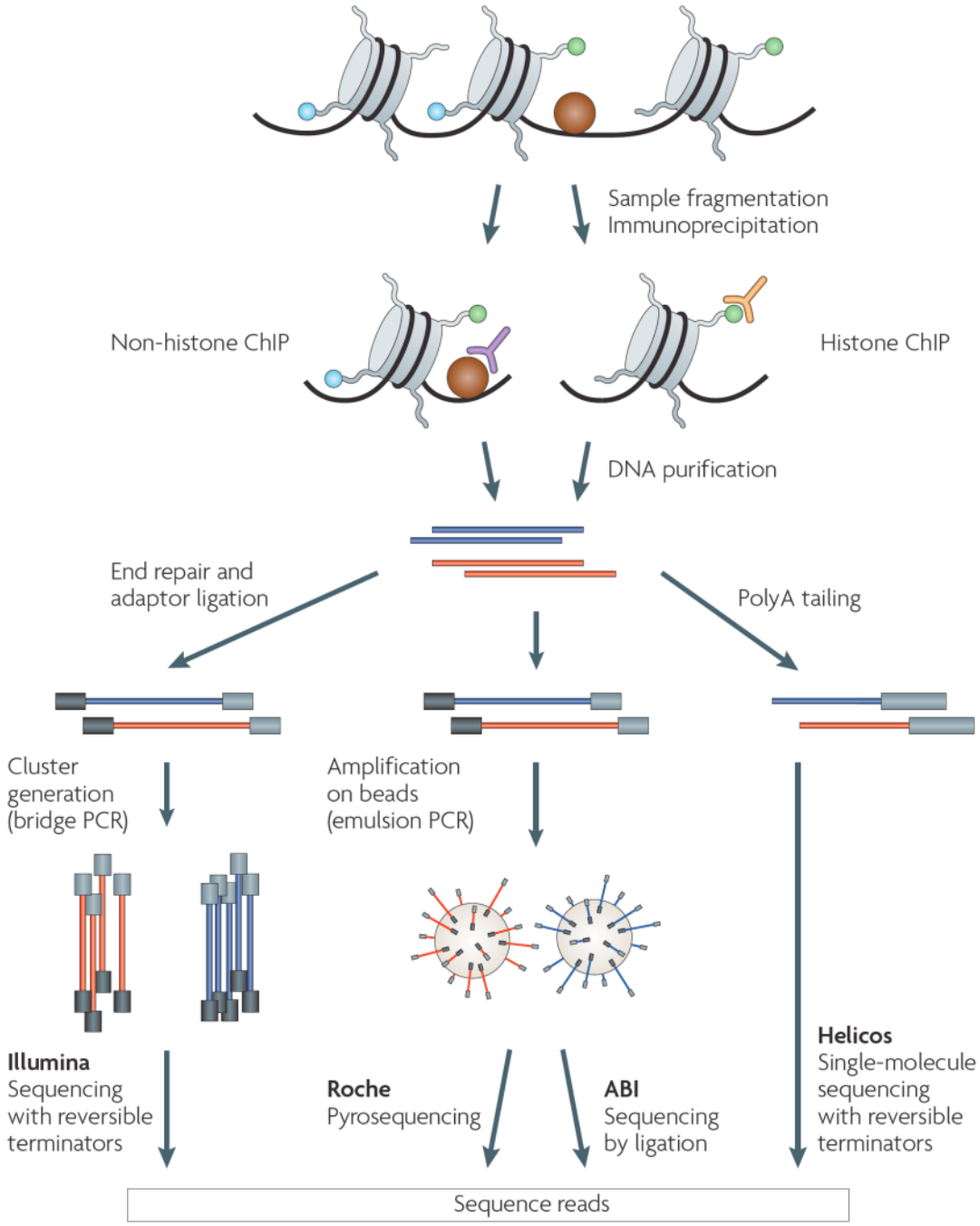


Figure 1.3. Overview of a ChIP-Seq experiment [71]. Using this technique, DNA associated with histones displaying covalent modifications or non-histone proteins, e.g. transcription factors, are obtained using specific antibodies. After subsequent purification the enriched DNA fragments are subjected to next-generation sequencing using one of various platforms, such as, Illumina. The short reads are further analyzed to investigate biological relevance.

ChIP Followed by Next-Generation Sequencing (ChIP-Seq)

At the time of the writing of this introduction, a major limitation of NGS technology is the prohibitive cost of sequencing. However, recent improvements, such as, increases in sequencing yield and development of multiplexing protocols, have resulted in significantly reduced costs. The bigger issue at present is data analysis, as improvements in computing power cannot keep pace with the exponential increase in sequence data. Therefore, a growing need exists for efficient analysis strategies. ChIP-Seq experiments generate millions of short DNA sequence reads representing the locations of proteins of interest, such as, histone modifications or transcription factors, distributed across the genome. The key steps of ChIP-Seq data analysis are outlined here.

First, the reads are mapped to a reference genome of the organism in question. Once considered a bottleneck in NGS analysis, recent advances have led to the development of several efficient and accurate mapping tools that have greatly sped up this process. The initial mapping step is followed by the detection of peaks signifying enrichments of histone modifications, a process known as 'peak-calling'. To investigate the biological function of observed peaks, the flanking regions of called peaks are often searched for coding or non-coding transcripts. The resulting lists of genes analyzed for evidence of enriched functional terms or pathways using various databases, such as, gene ontology (GO) or Kyoto Encyclopedia for Genes and Genomes (KEGG). Genes associated with biologically relevant pathways can be further examined, e.g. ChIP-Seq profiles in the promoter region can be compared between different experimental treatments to uncover its epigenetic effects. Limiting

factors of such analyses include large sample size – each ChIP-Seq profile may have several thousand loci of interest; incomplete genome annotation that may require the integration of information from several existing databases, and relative lack of existing statistical literature. Before examining the various steps of ChIP-Seq analysis in detail, let us first look at some of the accompanying issues.

Challenges Associated with Analysis

NGS technology is subject to biases dependent on technical aspects of the experiments and varied genomic context [71-73]. For instance, the 'mappability' of a genomic region measures the likelihood of sequences from this region being uniquely mapped. Mappability depends on read length, because longer reads have a greater likelihood of mapping uniquely to most regions of the genome. A highly mappable genomic region could, therefore, have higher read counts purely as a result of the sequencing process. Nucleotide composition could be another source of sequencing bias, because Illumina sequencers, for example, favor guanine-cytosine (GC)-rich regions [74]. Copy-number variations can lead to fluctuations in the expected numbers of reads from a genomic region that may not be observed in the ChIP sample. These sources of variation contribute to the 'background' signal that is non-stochastic [75, 76] and present significant modeling difficulties. The ChIP-Seq assay is also prone to amplification bias; PCR amplification is a part of the standard ChIP-Seq protocol to ensure enough starting material, but can lead to preferential amplification of abundant species. The more serious problem, however, is the variation in sampling rates due to differences in chromatin accessibility. In other words, regions having a loose chromatin conformation are more accessible to

restriction enzyme or micrococcal nuclease digestion, compared to regions having compact chromatin. While the former scales with fragment abundance, the latter depends on the particular library and is unpredictable. It is impossible to distinguish between the two at a sequence level and thus, it is considered prudent to discard redundant reads as a pre-processing step.

Negative controls can be used to partially account for the above factors. Examples include input DNA (normal sample preparation but no ChIP), non-specific antibodies, such as, immunoglobulin G (IgG), or ChIP without antibodies (mock IP). The suitability of such controls is a topic of continuing debate. Of the above three, input DNA is used most often and can correct biases in shearing and amplification. However, since input DNA fragments are spread across the genome, increased depth of sequencing may be necessary for improved coverage. A region of non-specific IgG binding can be a true binding site for a particular transcription factor and the rejection of such a site constitutes a false negative. Mock IP results in very low pull down and corresponding results are difficult to replicate [71]. Thus, there are obvious drawbacks of each method, which underlines the importance of accurate estimation of background variation.

A major difficulty of ChIP-Seq analysis is the diversity of patterns observed in enrichment regions. The detection of such enrichments, termed peak-calling, is not a trivial problem, as ChIP-Seq profiles demonstrate remarkable diversity, ranging from the sharp, punctate peaks of transcription factor data to broad, diffuse enrichments characteristic of certain histone modifications (Figure 1.4). This variability in the definition of peaks translates to modeling complexity, which is perhaps why there is

still no undisputed *numero uno* when it comes to peak-calling algorithms, in spite of much recent interest. A favored approach is to model ChIP-Seq data using a fitted discrete distribution (see below), but such approaches have their shortcomings [77]. Thus, there is a need for accurate methods of ChIP-Seq peak-calling free of limiting assumptions that is robust to diversity in binding profiles.

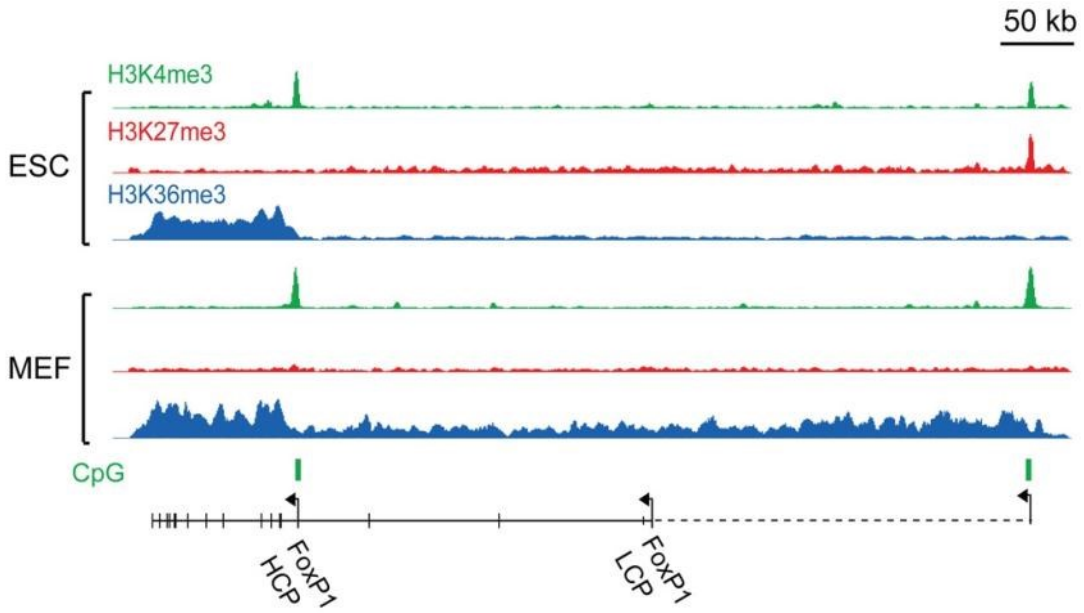


Figure 1.4. Diverse histone modification profiles observed on *FoxP1* in murine embryonic stem cells (ESCs) and embryonic fibroblasts (MEFs) (from [69]). H3K4me3 exhibits punctate peaks at TSS regions while H3K36me3 enrichment is broad and diffuse. *FoxP1* has one high CpG (HCP) and one low CpG promoter (LCP). An additional promoter 500 kb upstream of the HCP appears to be bivalent as it shows peaks of H3K4me3 and H3K27me3.

Genomic Mapping of Sequence Reads

As mentioned above, one of the first steps of NGS analysis is the mapping of sequence reads to the reference genome. A hot topic of research in the past few years, read mapping tools have made great strides in recent years, and as a result, a large number of mapping softwares are currently available, e.g. MAQ [78], RMAP [79], Bowtie [80], BWA [81], BFAST [82] and BLAT [83]. In spite of the bewildering

profusion of sequence aligners, a majority of available tools can be broadly divided into two groups based on underlying principles – hash tables and suffix arrays (reviewed in [84]). In both cases, the key to computational efficiency is the creation of an index either for the sequence reads or the reference genome, to enable fast matching.

Methods Based on Hash Tables

Hash tables are computational data structures that utilize index-key pairs to enable fast searching of lists. In other words, given a list of sequences, a hash table can be used to store the location (index) of each unique sequence (key) in the list. This idea can be easily extended to the genomic mapping of short sequence reads where the reference genome or sequence reads represent a searchable list, while unique *k*-mers of nucleotides and their locations represent key-index pairs. The iconic BLAST [85, 86] tool utilizes this approach; the query is first hashed into its constituent *k*-mers (keys), following which database lookup for matches is performed for each key. Exact matches (seeds) are joined before being refined and extended to produce the final alignment result, in an approach termed seed-and-extend.

BLAST requires *k* consecutive exact matches (default = 11) which represents a seed of '11111111111'. However, allowing for mismatches in the seed was found to increase sensitivity [87], thus lending credence to the use of spaced seeds. For example, a seed of '110110111101101', which looks for matches of length 15 while allowing for 4 mismatches, will find alignments with up to three mismatches in the first 11 positions, none of which can be detected using an unspaced seed.

MAQ [78], which stands for mapping and alignment with quality, was one of the first widely-used short-read alignment programs that employed spaced seeds. The mapping algorithm is applicable for k mismatches, but to avoid prohibitive memory requirements the default policy of MAQ ensures maximum sensitivity for up to two mismatches in the first 28 bp of Illumina reads. Two mismatches can be divided between four sections of a read in ${}^{4}C_{2}$ ways; thus, full sensitivity for at most two mismatches is achieved using six spaced seeds, MAQ provided various other advances such as the concept of mapping quality, an estimate of the error probability of an alignment based on sequencing qualities at mismatched bases, and also output a consensus sequence which could be used for variant detection and genotyping. Moreover, the gapped alignment used by MAQ is robust to indels (insertion-deletions). However, the memory requirements of holding a hash table in memory are large. Also, increasing sequencing yields and read lengths are likely to impact processing time as MAQ hashes the sequence reads.

Methods Based on Suffix Arrays

A suffix array is a data structure that consists of a sorted list of all suffixes of a string. A close relative of the suffix tree, the suffix array has lower space constraints and is easier to construct and implement [88]. Moreover, when combined with additional enhancements, such as, the full-text minute-space (FM) index [89], which enables efficient string matching in an array compressed using the Burrows-Wheeler Transform (BWT), the suffix array provides a significant improvement in speed. Identical substrings of the search space are collapsed and hence, alignment to such

regions need only be performed once, while in the case of hash table-based methods, explicit matching would be necessary for each occurrence.

The most commonly used tool in this category is Bowtie [80]. Default parameters of this program are similar to MAQ, with at most two mismatches allowed in an acceptable alignment. However, in contrast to MAQ, Bowtie indexes the reference genome and has a low memory footprint. Also, multiple CPU cores are utilized, if available, to further accelerate the alignment. At the time of its release, Bowtie was several orders of magnitude faster than MAQ or SOAP under similar conditions and, thus, represented a sizeable step forward. Other implementations of BWT and FM-index, such as, BWA [81] and SOAP2 [90], have since been released. However, despite obvious speed advantages, the sensitivity of this approach is reduced in the presence of indels. Recent tools, such as, Bowtie 2 [91] seeks to overcome this weakness by combining the efficiency of suffix arrays with the sensitivity of spaced seeds.

Peak Detection in ChIP-Seq Data

Following the accurate mapping of sequence reads to the reference genome, there needs to be a quantification step to determine regions that exhibit marked enrichment of reads or peaks. The challenges associated with peak calling, as mentioned above, have led to great interest in recent years to develop efficient, accurate and sensitive peak callers. As a result, a large number of peak-calling algorithms encompassing a great variety of techniques are currently available [92-103]. Although efforts to benchmark these algorithms have been carried out, there are no clear winners [104-106]. A majority of the methods showed comparable sensitivity and specificity when

tested on a limited number of qPCR-validated sites, although a great deal of variation in the number of called peaks was observed. However, comparisons of peak sites revealed significant overlap – smaller peak sets called by more conservative algorithms were usually contained within larger sets output by less stringent methods [105]. Often, the default parameters of a program are tuned to specific training data sets and therefore, results from different methods diverge considerably in general usage. Peak lengths for different methods on the same data set also display marked differences.

Despite major differences in algorithm design and performance, the primary workflow of most ChIP-Seq peak callers involves an initial modeling or training step, followed by peak detection either in the presence or absence of negative control data. Each peak is, then, assigned a significance score or p-value, an estimate of the likelihood of it being a 'true' enrichment versus an artifact.

Data Preprocessing

Raw sequence data needs to be preprocessed before being subjected to peak calling. As sequencing occurs in the 5'-3' direction, sequence reads represent the 5' end of the sequenced DNA fragment. For a more representative view, reads can be 'shifted' towards the 3' end to represent the middle of the fragment [92, 98, 101, 102], or extended to the length of the entire fragment [93, 95]. Some methods model the fragment length empirically. For instance, as sequencing of a DNA fragment is independent of the original strand, clusters (peaks) of sense and anti-sense reads usually flank *bona fide* transcription factor-binding sites (TFBSs). The distance between the sense and anti-sense peaks offers an empirical estimate of the fragment

length [92, 93], which can also be determined experimentally. However, sequenced fragment lengths usually consist of a spread of values and a point estimate is only a rough approximation.

After adjusting for fragment length, most algorithms produce an estimate of the read density across the genome, which is then analyzed for peak detection. Some tools partition the genome into bins or windows and calculate the distribution of reads. For instance, the number of 'shifted' reads falling within each window can be used to produce a read count histogram. This mode ensures the unambiguous representation of each read, but can suffer from edge effects depending on the size of the window. Counting the number of overlaps between extended reads or reads within a predetermined distance (sliding window) produces a smoother profile. The latter approach, while less susceptible to edge effects, contains some redundancy as a read can overlap more than one window. Smoothing techniques, such as, kernel density estimation (see below), can be used to produce a continuous probabilistic estimate of read density after one of the above steps. However, the degree of smoothing needs to be monitored closely to avoid removal of low intensity peaks.

Background Correction

As discussed above, NGS data is subject to multiple sources of background variation and bias, and accurate discrimination of ChIP-Seq peaks against the background is one of the foremost challenges of analysis. Negative controls have been used for the purposes of background correction, but, it is difficult to decide the appropriate control for a particular study, each of which has its own weaknesses. For instance, the increased depth of sequencing required for input DNA controls make it infeasible to

have a matching input library for each sample in a large sequencing experiment. Thus, even though it is recommended, many studies do not include negative controls. Consequently, accurate modeling of the background from the ChIP-Seq data is extremely important. A common assumption for this step is that genomic regions with lower read counts are likely to be part of the background signal. Thus, several methods approximate the background as a random variable that follows a discrete distribution, such as, Poisson [64, 101, 107] or negative binomial (NB) [96, 103], fitted to genomic regions with low read densities. The Poisson distribution can be used to model the probability of observing counts y, and has a probability density function (pdf) defined as follows:

$$\Pr(Y = y) = \frac{\lambda^y e^{-y}}{y!}$$

where, λ is both the mean and variance of the distribution. This assumption is inadequate for NGS data as the observed variance (also called dispersion) can be much higher than the mean [108]. A variation of the Poisson model to allow for greater dispersion include the generalized Poisson model which has the pdf [109],

$$\Pr(Y = y) = \frac{1}{1+\alpha y}\frac{\tilde{\lambda}^y e^{-\tilde{\lambda}}}{y!}$$

where,

$$\tilde{\lambda} = \frac{\lambda(1+\alpha y)}{(1+\alpha\lambda)}$$

The mean of the generalized Poisson distribution is λ and variance is $\lambda(1 + \alpha\lambda)^2$. The parameter α controls dispersion with $\alpha > 0$ modeling overdispersion and $\alpha = 0$ reducing it to the Poisson model. A better fit to count data is provided by the NB

distribution which includes a dispersion parameter ϕ, and can be represented as a mixture of the Poisson and Gamma distributions [110]. If observed counts are distributed as $y \sim$ Poisson(λ), but λ is itself a random variable with a Gamma distribution,

$$y|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}}$$

where, α is called the shape parameter and β represents the scale parameter. The mean of the above distribution is $\alpha\beta$ and variance is $\alpha\beta^2$. Then, the probability mass function of y is negative binomial as,

$$\Pr(Y = y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha) y!} \left(\frac{\beta}{1+\beta}\right)^{y} \left(\frac{1}{1+\beta}\right)^{\alpha}$$

The mean of the above distribution is $\alpha\beta$ and variance is $\alpha\beta^2$. For statistical modeling, an alternate parameterization is used: $y \sim$ NB(μ,ϕ), where $\mu = \alpha\beta$ and $\phi = 1/\beta$, so that the mean of the distribution is μ and variance is $\mu + \mu^2\phi$. This model also reduces to the Poisson model when $\phi = 0$. The dispersion parameter can be estimated from the data using maximum likelihood and allows greater dispersion than allowed by the Poisson model.

However, given that the background signal has been shown to be non-random the above models are often inadequate [77] and as a result the associated methods tend to call more false positives [111]. Thus, it is clear that the existing approaches for background correction are not perfect and there is definite room for improvement.

Methods of Peak Detection

Peak-calling algorithms, as mentioned above, are extremely diverse. A majority of methods are aimed at detecting TFBSs [64, 92, 93, 98], while a relative few focus on histone modification data [100-102]. However, underlying modeling philosophies share some similarities and can be loosely grouped under the following headings:

1. Simple threshold
2. Local measures of enrichment
3. Kernel density estimation
4. Hidden Markov models
5. Incorporation of additional covariates

I will now discuss the characteristics of each category and give a brief overview of some tools within each class.

Methods Based on a Simple Threshold

The number of reads within a putative enrichment region is often used as an estimate of significance and thus, early peak calling methods utilized a simple read-height threshold *T* to call peaks [68]. However, this simplistic approach can be difficult to apply as peak heights observed in a ChIP-Seq sample are subject to sequencing depth, antibody quality and data characteristics. ChIP-Seq profiles for transcription factor binding are usually sharp and well defined, and thus, the choice of a suitable threshold may be evident from the data. However, the same cannot be said of most histone modifications and even certain proteins, such as, growth-associated binding protein (GABP), which produce more diffuse peaks. Secondly, an antibody that exhibits some non-specific binding is especially vulnerable to threshold effects. Low

affinity binding sites can be mistaken for background, although, in this case, a majority of peak callers would have difficulty in distinguishing these peaks from true peaks. Finally, the sequencing depth varies from one sample to another and thus, appropriate thresholds have to be different for different samples in a single sequencing experiment.

FindPeaks [93], a widely used tool, is based on this approach. Briefly, sequence reads are extended to represent their estimated fragment length and the peaks of overlapped read profiles are used for peak detection. This initial step can be followed by peak refinement via 'trimming' and segmentation into sub-peaks. An empirical estimate of false discovery rate (FDR) is obtained using Monte Carlo simulations. FindPeaks boasts a modular design with various user options to tweak performance. However, the lack of a user guide to choose a suitable threshold is a major drawback.

Another method that utilizes a height threshold in peak detection is cisGenome [96], a tool capable of analyzing both ChIP-on-chip and ChIP-Seq data. CisGenome implements a two-pass peak detection procedure. In the first pass, genome-wide read counts are obtained using a sliding window and those above a user-defined threshold are called as putative peaks. High confidence peaks obtained from the first pass are used to estimate DNA fragment size; subsequently, sequence reads are shifted to represent the center of the fragment and the peak detection process is repeated. In the absence of negative control data, cisGenome estimates FDR by fitting a NB distribution to low read-count windows in the ChIP sample. If a negative control is present, cisGenome calculates binomial p-values as a measure of significant enrichment. If there are T ChIP reads and C control reads in a putative peak, the

proportion of successes $p = T/(C+T)$. Then the binomial probability of observing at least T successes in $t = C + T$ trials under the null hypothesis H_0: $p = 0.5$, is,

$$\Pr(\#\ of\ successes\ in\ t\ trials \geq T) = \sum_{i=T}^{t} \binom{t}{i} p^i (1-p)^{t-i}$$

In contrast to FindPeaks, the user-defined cutoff used in cisGenome is associated with an FDR level making it easier for the user to choose a suitable value. Other innovations include the application of the NB distribution, shown to be a better fit to the background than the Poisson distribution, and a graphical user interface (GUI) for clickable data analysis and visualization.

Local Measures of Enrichment

The gradual increase in sophistication of peak calling algorithms saw tools utilizing local features of the data to detect peaks. The so-called 'directional methods' leveraged the distance between nearby peaks of sense and anti-sense reads to serve as an indication of the existence of a TFBS. Kharchenko et al.'s spp package [111], contains a collection of measures most of which depend on the strand-specific read density. For instance, window tag density (WTD) scores each window based on sense and anti-sense read counts within a user-specified distance. Peaks are called based on local maxima of score profiles and FDR calculated as a ratio of the number of peaks detected in the test sample versus that in a negative control. SiSSRs (Site Identification from Short Sequence Reads) [64, 94], uses a similar idea. First, strand specific read count profiles are calculated with a sliding window approach. Sense and anti-sense read counts are assigned positive and negative scores, respectively, and a composite count is calculated for each window. Putative binding sites are predicted at the points where the composite count transitions from positive to negative. These

TFBS predictions are further filtered for total read counts and FDR estimated as a ratio of the number of peaks with the same number of reads in the background (Poisson distribution or negative control), to that observed in the ChIP sample. Directional models are simple and thus, efficient and easy to implement. However, the assumption of proximal sense and anti-sense peaks flanking a binding site is less applicable to broad enrichment regions and results in lower sensitivity.

The widely used tool, MACS [92], also uses local modeling of the data to detect peaks. Similar to cisGenome, MACS uses a two-pass approach to peak calling. Sequence reads are shifted to represent the center of the fragment and read count profiles are calculated based on a sliding window scan. In the first pass, MACS fits a global Poisson model (λ_{global}) to the ChIP data and calls putative peaks based on a specified p-value cutoff. The second pass is used to capture local biases by fitting Poisson models to regions of varying length (λ_{1k}, λ_{5k}, λ_{10k}) flanking the putative peaks. For each peak, a dynamic Poisson parameter λ_{local} is defined as,

$$\lambda_{local} = \begin{cases} \max(\lambda_{global}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}), & if\ control\ data\ is\ present \\ \max(\lambda_{global}, \lambda_{5k}, \lambda_{10k}), & otherwise \end{cases}$$

and used to assign a p-value. The latter process is designed to model the background and is performed in a control data set or in the ChIP data in the absence of a control. Note that λ_{1k} is not used in the absence of control data ensuring that local variations in the ChIP sample In the presence of a control data set, MACS also calculates the FDR as follows: the peak calling procedure is performed in the ChIP sample versus control, and again in control vs ChIP. The FDR is then estimated as the ratio of control peaks to the ChIP peaks. MACS has performed well for several different

ChIP-Seq data sets and has thus, been widely adopted, particularly for TFBS prediction [112].

Kernel Density Estimation

Kernel density estimation (KDE) is a non-parametric procedure to estimate the pdf of a data set. Widely used for data smoothing, KDE involves sampling a set of points within a specified distance that are weighted based on a predefined function referred to as the 'kernel'. If y represents the observed counts, a kernel density estimator of the 'true' ChIP-Seq profile F at i could be represented by,

$$\hat{F}(i) = \frac{1}{2nh} \sum_{j=i-n}^{i+n} K\left(\frac{y(j) - y(i)}{h} \right)$$

where, n points on either side of i are sampled to produce the estimate, K is the kernel function and h is the bandwidth. The kernel defines the shape of the smoothed data, while the 'bandwidth' determines the degree of smoothing, with larger values resulting in greater smoothing. Since the shearing of DNA is a random process, the shape of ChIP-Seq peaks resembles a Gaussian distribution, making a Gaussian kernel suitable for ChIP-Seq data analysis.

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

KDE produces a continuous smoothed estimate of the data, and enables easier visualization of various genomic patterns. However, it is important to choose a suitable bandwidth, since too much smoothing could markedly reduce the signal from smaller peaks, thus reducing sensitivity.

Several methods employed the above statistical framework to detect peaks in NGS data, such as, GeneTrack [113] and F-Seq [97], but the first one to be used widely for

ChIP-Seq data was QuEST (Quantitative Enrichment of Short Tags) [98]. This tool first generates a KDE profile for each strand, before combining them into a single profile for detecting local maxima. QuEST enforces a fold-change requirement for peak calls, i.e. ChIP reads have to be at least a certain fold (specified by the user) greater than a control sample. Peaks satisfying the above criteria are marked as putative enrichments and an FDR estimate is calculated based on the negative control. Briefly, the control sample is randomly divided into two parts, and the peak-calling procedure is performed on one part of the control sample with the other serving as the background. The ratio between the number of control and ChIP peaks serves as an estimate of FDR. QuEST implements some stringent restrictions, which limit its applicability. The FDR estimation step is only performed if the control sample contains at least twice as many reads as the ChIP sample. Also, experiments without a negative control are not supported.

Another method that employs KDE is SICER (Spatial clustering approach for the Identification of ChIP-Enriched Regions) [101], although at a different stage of peak calling. First, shifted reads are partitioned into non-overlapping windows and a Poisson model is used to mark windows with significantly elevated read counts (eligible). Windows containing non-significant read counts are called ‘gaps’ and the maximum number of allowable gaps in a peak (g) can be set by the user. Putative peaks are scored with the negative logarithm of the product of window p-values with gaps contributing 0. The likelihood of observing the score distribution is modeled using KDE with a kernel that assumes exponential decay,

$$K(x) = \alpha e^{-\beta x}$$

where, the fitting coefficient α is modeled from the data. The expected number of peaks in a random background model (E-value) is used to control for false positives. In the presence of negative control data, the number of reads within each peak is compared to that in the control sample using a binomial test. SICER was one of the first peak callers aimed at histone modifications and the 'gap' parameter allowed the detection of broader enrichment regions.

KDE is the most commonly used technique for ChIP-Seq peak detection due in part to its flexibility and statistical properties. However, the choice of bandwidth is critical and should match the DNA fragment length from the sequencing experiment.

Hidden Markov Models

A random process is said to have the Markov property if the state of the process at any time t only depends on its state at the immediately preceding time point $t - 1$. A hidden Markov model (HMM) is a probabilistic statistical framework used for modeling a random Markov process with unobserved (hidden) states. The most important parameters associated with an HMM are the number of states of the process N, state transition probabilities A and emission probabilities E (the probability of observing an output value given a specific state). The HMM framework has seen wide application in pattern recognition across a variety of fields [114].

Given its properties some methods have applied HMMs to ChIP-Seq data by likening the analysis to a classification problem: by observing a sequence of counts we want to infer the (hidden) state of the system, whether enriched or not. BayesPeak [103], uses the above framework, employing an HMM that can assume one of two states: a true binding site or background. Read density profiles are generated for sense and anti-

sense strands using the 5' ends of sequence reads. The emission probabilities are modeled using the NB distribution and parameters are estimated in a fully Bayesian manner using Markov chain Monte Carlo (MCMC) sampling techniques. HPeak [99] also models the data with a two-state HMM, but emission probabilities are based on generalized Poisson (enrichment) or zero-inflated Poisson distributions (background). Parameters for the HMM are estimated using the Viterbi algorithm [115] and the read counts in predicted peaks are compared with that in a control sample using a χ^2 test.

A more recent method, RSEG [100], uses HMMs to detect broad enrichment regions characteristic of several histone modifications, such as, H3K27me3 and H3K36me3. Although RSEG also applies a two-state HMM to the problem of ChIP-Seq peak detection, it focuses on the detection of boundaries between regions of significant enrichment and background. Like BayesPeak, the emission probabilities of the two states are modeled using the NB distribution. The empirical distribution of transition probabilities is used to find windows with a high likelihood of being points of transition from an enrichment region to background and vice-versa. Other innovations include the development of a novel distribution for the difference between two independent random variables that follow the NB distribution (NBDiff), which is used for comparisons with a control. In the latter case, the HMM has three states corresponding to no difference between ChIP and control, greater enrichment and lower enrichment in the ChIP sample, respectively.

The greater algorithmic complexity of HMMs makes the implementation of the above methods more difficult. BayesPeak, owing to its generalized design, requires extensive simulations for parameter estimation and training, which, in turn, results in

high computational demands [116]. HPeak is simpler in comparison and hence more efficient, although the Poisson models implemented in the algorithm might not provide a very good fit for the background. RSEG, by design, is well suited to the detection of broad peaks, but this may limit its applicability to TFBS prediction.

Methods Incorporating Additional Covariates

As mentioned above, accounting for sources of technical variation in NGS data is necessary for the accurate discrimination of peaks from background. Therefore, it is reasonable to expect that the incorporation of additional covariates, e.g. mappability and G/C content, into the peak-calling procedure can improve accuracy. PeakSeq [95] was one of the first methods to adopt this approach by accounting for mappability in its peak-calling procedure. First, read density maps are created by calculating overlaps between reads extended to their fragment length. Each chromosome is divided into segments and a random distribution of reads mapping to these segments is generated using the Poisson distribution, taking into account the number of mappable bases in the segment. This procedure is similar to the dynamic Poisson model used in MACS, although the recommended segment is much larger (1 Mb) and the effective segment size accounts for mappability. In the first pass, a set of putative peaks are generated using a read count threshold, which is calculated separately for individual segments using the fitted Poisson model. This threshold when applied to negative control data also provides an estimate of the FDR and can be adjusted to user specifications. Linear regression is used to normalize the ChIP sample to a control and binomial p-values calculated to denote statistical significance of a peak.

A recent method, ZINBA (Zero-Inflated Negative Binomial Algorithm), further generalizes the process of including covariates in the peak detection procedure [102]. ZINBA uses a mixture regression approach to classify windowed read-counts into one of three components – enrichment, background and zero. The third component is introduced due to the presence of large numbers of zero-count windows in sparse ChIP-Seq data sets, either due to inherent characteristics of the data or low sequencing depth. ZINBA employs an expectation-maximization (EM) algorithm to estimate the probability of component membership of each window, with enrichment and background read counts modeled using the NB distribution. Moreover, the relative contributions of covariates (including interactions) can be estimated and the important factors chosen using a model selection procedure based on the Bayesian information criterion (BIC). Windows having probability of enrichment greater than a specified threshold (default 0.95) are marked as putative peaks and adjacent enriched windows are merged. A further setting enables the concatenation of peaks within a fixed distance for broad enrichment regions. ZINBA provides an important advance in the field of ChIP-Seq peak-calling by jointly modeling ChIP-Seq data with genomic covariates. However, increased complexity, e.g. in the computationally intensive model-selection step, results in high computational demands.

Detection of Differential Binding

The accurate detection of enrichment regions in ChIP-Seq data may be sufficient for most exploratory analyses. However, at times it may be informative to compare ChIP-Seq profiles across different experimental conditions. Possible questions include, for example, the variation of transcription factor binding in response to a disease or the

dynamics of histone modifications during stem cell differentiation. The prohibitive cost of next generation sequencing at the time of its introduction, served as a deterrent for such extensive epigenetic studies. However, with the gradual reduction of sequencing cost, the use of intricate experimental designs for epigenetics assays has become more prevalent, with the detection of 'differential' binding or enrichment a topic of interest.

ChipDiff [117] was an early attempt at the detection of differential histone modification enrichments. This method uses a three-state HMM to compare two sequencing libraries L_1 and L_2 using fold-changes of normalized read counts. Similar to the two-sample analysis in RSEG, the three states correspond to no difference between the two samples, higher enrichment in L_1 and higher enrichment in L_2, respectively. Windows with a high probability of being in one of the latter two states are marked as putative points of differential histone marks with adjacent sites being merged. The fold-change approach used by ChipDiff can be prone to large variations particularly at low signal strength while a fixed window size of 1 kb causes lowered resolution. Moreover, ChipDiff only supports two ChIP-Seq libraries and thus, has limited applicability to more complex experimental designs.

Rapid advances in the field of mRNA sequencing (RNA-Seq) analysis saw the development of suitable statistical methodology and the development of two popular tools, edgeR [118] and DESeq [119]. Both these methods utilize the NB distribution to model read counts although the respective implementations are somewhat different. EdgeR adopts the classical NB distribution wherein the mean μ and variance σ^2 are related as $\sigma^2 = \mu + \alpha\mu^2$. The dispersion parameter α is estimated for

each gene using conditional maximum likelihood, while an empirical Bayes procedure is employed to enable the shrinkage of dispersions. Since α is the only parameter to be estimated per gene, edgeR can be applied to experiments with smaller numbers of replicates, as is often the case in sequence-based assays.

DESeq extends the model used by edgeR with a more flexible data-driven approach. The 'true' fragment count for each gene is assumed to be proportional to the observed count scaled by a normalization factor dependent on the library size. The raw variance of each gene is assumed to be a smooth function of the read count of each gene, which is estimated using local regression. The local regression approach uses genes of similar expression level to predict gene-wise variances and as a result DESeq is applicable to experiments having small numbers of replicates.

Although both edgeR and DESeq were developed for gene expression assays, they can be extended to other NGS applications, such as, ChIP-Seq. However, due to major differences in the applied protocols, relevant results may need to be treated with some caution. For instance, the RNA-Seq is less susceptible to amplification bias and thus, raw reads can be used directly with the above tools for differential expression analysis, while redundant reads need to be removed from ChIP-Seq data.

In summary, ChIP-Seq analysis is a complicated process comprising several important steps. The maturation of next-generation sequencing technologies and development of efficient software has meant that the computationally intensive read-mapping step is no longer the bottleneck of the analysis. Also, statistical methodology suited to analysis of count data has made it easier to perform differential analyses. However, in spite of the availability of several peak-callers, virtually every algorithm

makes distributional assumptions for computational efficiency that have been shown to be inadequate. The added complexity of diverse enrichment patterns observed in ChIP-Seq data means that there is a continuing need for accurate peak-calling algorithms, robust to background variations and sensitive to diverse binding patterns.

Marek's Disease

Marek's disease (MD) is a highly contagious, lymphoproliferative disease of chickens caused by an α–herpesvirus, Marek's disease virus (MDV). MD was initially described and characterized in 1907 by eminent Polish veterinarian, József Marek, as a 'polyneuritis', but was later found to also cause lymphomas. The discovery of the causative agent, MDV, in the 1960s proved to be the next major step forward, occurring soon after the economic boom of the poultry industry [120]. The ubiquitous nature of MDV results in exposure for virtually all chickens from birth, and the acute forms of the disease became a particular cause for concern to the industry during expansion and increased production of that decade. The introduction of a successful vaccine in 1969 [121] temporarily allayed fears, but also led to increased virulence of the virus. Further development of vaccines followed [122], but resulted in even greater levels of virulence [123], as it became clear that alternative sustainable methods were necessary for controlling MD in the long term.

Genetic resistance to MD provides such an alternative. Natural resistance to MD was observed in commercial flocks of chicken as early as 1932 [124] and the breeding of chicken lines selected for resistance or susceptibility to MD had been shown to be possible in 1947 [125], even before the discovery of MDV as its causative agent. Subsequently, two independent research groups selected and bred MD-resistant and

susceptible lines – lines N and P selected by the above researchers at Cornell University and lines 6 and 7 selected by Stone at East Lansing. The above two unrelated groups of inbred lines have since been the center of extensive study and are the primary source of the current understanding of MD-resistance and susceptibility.

Marek's disease has several interesting features. It is the only known lymphomatous disease that has been successfully controlled by a vaccine. Three closely related serotypes of MDV exist – MDV-1, MDV-2 and herpesvirus of turkeys (HVT). MDV-2 and HVT are usually non-pathogenic, but MDV-1 causes acute lymphomas in susceptible birds. Neoplastically transformed cells in MD tumors have been found to overexpress CD30 antigen [126] and thus, MD is a natural model for Hodgkin's lymphoma in humans [127]. Also, the outcome of MDV infection depends on various host, viral and environmental factors; non-oncogenic strains of MDV can become oncogenic under certain conditions, such as, stress. Thus, it is a great animal model for the study of host-pathogen interactions, in general, and virus-induced lymphoma formation, in particular. Also, the populations of inbred lines can help understand the genetic basis of resistance and susceptibility to a cancer-causing agent.

Marek's Disease Pathogenesis

MDV exhibits a complex life cycle in host cells involving an initial cytolytic phase, a latent phase, a late cytolytic phase and transformation. Initial infection is believed to occur when the birds inhale the virus particles. Once in the respiratory tract, the virus is phagocytosed by macrophages or dendritic cells (DCs) that, as a result, become infected. The phagocytosis might occur directly or after replication in epithelial cells. The MDV-infected macrophages or DCs enter circulation and carry the virus to the

major lymphatic organs of the bird, and within 24 hours of infection the virus is detectable in the spleen, thymus and bursa of Fabricius.

During the early cytolytic infection that follows, the virus first targets B lymphocytes which likely surround infected ellipsoid-associated reticular cells (EARCs) in the spleen [128]. This phenomenon is also the reason why B lymphocytes are the primary targets of MDV at the cytolytic stage of infection. Subsequently, the infection spreads to other lymphoid tissues, such as, bursa and thymus, that lag behind the spleen by a day. In each of these organs, B lymphocytes form the largest proportion of infected cells, along with smaller numbers of CD4+ and CD8+ T lymphocytes [129]. Cytolytic infection can cause major atrophy of bursa and thymus, accompanied by immunosuppression, in contrast to the spleen which shows slightly increased weight and greater virus load. Interestingly, T lymphocytes activated as a consequence of MDV infection of B cells renders them susceptible to infection, while naïve T lymphocytes are relatively immune [130]. This has led to the suggestion of an MDV receptor expressed on the surface of CD4+ T lymphocytes, but in the absence of further evidence this remains a matter of conjecture.

At 6-7 days post infection (dpi), the infection enters latency during which the viral genome is present in host cells but no viral antigens are expressed in lymphoid tissue and no viral replication observed. By this time, most cytolytically infected B cells are dead and CD4+ T lymphocytes form the bulk of the infected cell population. Latently infected T lymphocytes may be transformed in latter stages of the disease and go on to form lymphomas, and the relationship between these two stages is poorly understood. Latency is a hallmark of many herpesvirus infections and the switch from

cytolytic to latent infection is believed to be influenced by host factors. The time of incidence of latency coincides with the establishment of the host immune response. Also, the reemergence of cytolytic infection in susceptible genotypes is likely concurrent with immunosuppression in the host. Various host cytokines such as interleukin (IL)-6, IL-18 and interferon (IFN)-γ and other cell signaling molecules such as nitric oxide (NO) are believed to play major roles in the establishment and maintenance of latency [131]. Certain virus genes such as a group of latency associated transcripts (LATs) and *Meq* are important players in latency. MDV LATs include three RNAs that interfere with MDV immediate-early gene ICP4 and inhibit translation of the ICP4 protein resulting in abrogation of lytic infection and onset of latency [132]. *Meq* blocks apoptosis of latently infected CD4+ T lymphocytes and transactivates latent gene expression [133], thereby helping maintain latency. In resistant chickens, latent infection persists at low levels in circulating lymphocytes without reactivation, while inflammatory changes in lymphoid tissues gradually recede.

In susceptible chickens, latency is followed by a second phase of cytolytic infection 2-3 weeks after initial infection [134]. This late phase of infection affects immune organs of thymus and bursa, along with epithelial tissues, such as, kidney. It appears that latently infected lymphocytes circulate the virus to different parts of the body such as, brain, nerves and skin before reactivating as a result of immunosuppression [134]. Following reactivation of the virus there is heightened inflammation, necrosis of infected lymphocytes, infiltration of mononuclear cells into infected tissue and major atrophy of bursa and thymus. Virus particles carried to the skin result in

infection of the feather follicle epithelium, which is fully productive, i.e. there is widespread release of infectious, cell-free virus particles and apoptosis of infected follicular cells. The feather follicle epithelium is the site of continued expression of MDV antigens and persistence of virus particles in resistant and susceptible birds alike.

The final stage of MDV infection is the transformation and proliferation of latently infected cells into lymphomas. The major site of proliferation appears to be the spleen, although it is not believed to be essential for the formation of lymphomas [135]. About 21 dpi, large increases in T cells that are possible precursors of transformed cells, are observed in the spleen [136]. Cells expressing high levels of CD30 antigen are detected in blood and spleen of both resistant and susceptible birds at the end of the early cytolytic infection [129]. This marker, encoded by the host and expressed in MD tumors and cell lines, is found only on a small population of MDV-free lymphocytes [126] and not expressed on naïve CD4+ T lymphocytes. Thus, it is likely that the CD30+ T lymphocytes in spleen are precursors of the transforming cell population. Soon after, infected T lymphocytes migrate to visceral organs and peripheral nerves where they proliferate into tumors. Approximately three-quarters of cells found in MD tumors are CD4+ T lymphocytes with the rest being B lymphocytes. However, almost all cells showing non-productive infection are CD4+ T lymphocytes [137], indicating that these cells form the bulk of the neoplastic cell population in lymphomas. In susceptible genotypes, the above CD4+ T lymphocytes undergo major proliferation, going on to form mature lymphomas. However, in

resistant chickens, cytotoxic CD8+ T lymphocytes appear to keep proliferation in check, resulting in apoptosis and regression of MD lesions [138].

Immunity to Marek's Disease

Host responses to Marek's disease are determined by innate and acquired immune responses. The two major components of innate immunity are macrophages and NK cells. Macrophages play an important role in innate immune response and adaptive immunity by functioning as antigen-presenting cells (APCs). As mentioned above, macrophages engulf virus particles in the respiratory tract and transport them to lymphoid tissue where cytolytic infection is initiated. Initial studies *in vitro* suggested that macrophages were resistant to MDV infection [139], but subsequent studies showed that splenic macrophages express MDV antigens, consistent with virus replication [140]. Macrophages also recognize antigens via pattern recognition receptors, release cytokines and soluble factors (e.g. NO) that aid in defense against infections. Recent reports have suggested that NO produced by inducible nitric oxide synthetase (*iNOS*) can inhibit MDV replication in early cytolytic and latent phases of infection [131]. Higher levels of NO are observed in MD-resistant chickens at early stages, and possibly contribute to lowered viral load in these genotypes. Further support for the above was obtained when studies found increased tumor incidence and viral load after treatments to reduce macrophage numbers [141] and vice-versa [142]. Thus, macrophages play an important role in reducing viral load during early cytolytic infection possibly via the secretion of NO.

NK cells constitute another important component of innate defense mechanisms through their ability to respond to the secretion of chemokines and cytokines. This

property of NK cells is shared with T lymphocytes even though they are non-phagocytic and do not express antigen receptors. In normal individuals, NK cells are found only in peripheral blood, spleen and bone marrow, but can move quickly to sites of inflammation upon induction by various chemotactic molecules. Various reports have suggested the possible involvement of NK cells in MDV infection. NK cells isolated from spleen of normal chickens lysed cells from a MD tumor cell line [143]. NK cell activity increased 7 days after MDV infection in both resistant and susceptible chickens [144]. MD-resistant line N chickens exhibited higher and more sustained NK cell activity than susceptible line P chickens [145]. Also, a genomic region strongly associated with MD-resistance was found to be syntenic to human and murine NK cell clusters [146]. All in all, NK cells appear to be involved in protective immunity against MDV and are possibly most active during the early cytolytic phase of infection. However, their mechanism of action remains unclear as the characterization of NK cells is hindered by the lack of available markers [147].

The major components of the acquired immune response are CD8+ cytotoxic T lymphocytes (CTLs) and CD4+ T helper cells that secrete cytokines. CTLs are associated with MD-resistance in line N chickens as they interact with and subsequently remove MDV ICP4 through the action of specific receptors [148]. CTLs help in reducing MDV replication, transmission and persistence. The role of cytokines in MD has been the subject of intense study in the past few years. As mentioned above, several cytokines such as IFN-γ, IL-1β and iNOS were preferentially upregulated in spleen from line N chickens from early stages of infection [131]. More comprehensive studies of cytokine responses found

upregulation of IFN- γ in all infected chickens, consistent with the previous study, along with inflammatory cytokines IL-6 and IL-18, in susceptible birds during early cytolytic infection. In addition to the above host factors, a virus-encoded IL-8 homolog (vIL-8) has also been found [149]. Since IL-8 acts as a chemoattractant for T lymphocytes, the above finding has led to speculation that the viral homolog attracts T cells to sites of infection. However, vIL-8 shares greater homology with a B lymphocyte chemoattractant and can be better categorized as a CXC chemokine. The precise role of vIL-8, therefore, remains unclear.

Marek's Disease Resistance and Susceptibility

The first major step towards understanding the mechanisms behind MD resistance was provided by the observed association between MD resistance and inheritance of the B blood group locus [150]. Since this locus was a known marker for the chicken major histocompatibility complex (MHC), the above observation gave rise to the possibility that genes found within the chicken MHC could be responsible. Several subsequent studies confirmed this finding, although it did not preclude the possibility that other genes might also be involved. Based on this and further experiments, genetic resistance to MD can be subdivided into two categories – MHC-associated and non-MHC associated resistance.

Several known haplotypes of the B locus provide varying levels of resistance, such as, B^{21} confers high resistance irrespective of genetic background, and B^{19} is associated with susceptibility. However, certain other haplotypes, e.g. B^{2}, can have widely varying effects on MD-resistance depending on other factors. It was shown that the differential susceptibility of the lines N and P mentioned above is largely

correlated with their B haplotype [151]. Line N possesses the B^{21} haplotype associated with high resistance while line P contains the B^{19} haplotype which confers high susceptibility. The mechanism behind MHC-associated MD resistance has been elucidated to some degree. MDV infection is believed to induce low levels of CTLs that are specific for certain proteins encoded by MDV. For instance, CTL specific to the viral immediate-early protein ICP4 were found in MD-resistant line N chickens carrying the B^{21} haplotype, but not in line P [152]. It was suggested recently that natural killer (NK) cells may be involved in the process [145]. Also, class I MHC molecules had varying levels of expression on the cell surface of uninfected cells, with B^{19} having the highest expression and B^{21} the lowest [153]. This raised the possibility that NK cells are major effectors in MD, as in mammals they can detect differences in cell surface expression. Alternatively, CTLs and NK cells can both confer some level of protection from infection.

In contrast to lines N and P, lines 6 and 7 both carry the B^2 haplotype and thus, differences in MD resistance observed in these lines depend on factors outside the chicken MHC. This situation is also true of several outbred and commercial flocks of chickens whose resistance or susceptibility to MD cannot be fully explained on the basis of their B haplotypes alone. Many other genes could possibly be involved in this form of MD resistance and several different approaches have been used to investigate the underlying mechanisms.

Early studies reported differences in viral replication between the two lines, with the susceptible line 7 showing higher rates of viral replication and subsequent viral load. MDV-infected lymphocytes from this line contained high numbers of virus particles

from the early stages of infection which was maintained throughout their lifetime. Resistant line 6 chickens, on the other hand, showed a gradual increase in viral load which peaked around 10 dpi before falling to low levels. The early differences between the two lines suggested differences in innate rather than adaptive immune response. The clearance of infection observed in line 6 at later stages of MD suggests an adaptive response, although the inability of line 7 to mount a successful defense could either be due to greater injury to the immune system during early cytolytic infection [136] or inherent differences in immune response [154].

Lymphocyte surface markers were believed to be partly responsible for the differential disease in the two lines. Investigations led to the discovery of three alloantigens designated as Ly-4 [155], Bu-1 and Th-1 [156], each of which showed a certain degree of association with MD resistance. Genomic mapping revealed 14 genomic regions associated with disease resistance [157, 158]. Further attempts to map resistance loci using a backcross population [146] resulted in the discovery of a region on chromosome 1 with a strong association with MD-resistance. This region appeared to control viremia and shared homology with human and mouse NK cell clusters. One putative resistance gene present in this region was identified and designated MDV1. Subsequent studies using microarrays [159] identified several immune-related genes, such as, IFN-γ, that showed significant differences in expression in the two lines, located in genomic regions associated with MD resistance. Recent studies of host responses to MDV infection have further expanded current knowledge [160, 161]. However, given the obvious complexity of the disease,

the focus has gradually shifted to systems analyses to uncover pathways associated with MD [162, 163].

Thus, non-MHC associated MD resistance is influenced by many genetic and environmental factors. Studies attempting to map the MD-resistance observed in these lines to specific genomic regions have met with moderate success, while the investigation of the transcriptional effects of MDV have revealed differential expression of certain important host cytokines and viral genes. However, none of these loci can completely explain the mechanism of MD-resistance and susceptibility. Also, environmental factors can have a major impact on the outcome of infection, which suggests that epigenetic processes play an important role in MD progression. Therefore, this is a great animal model to study epigenetic effects of a lymphomatous virus and the epigenetics of disease resistance. We propose to investigate one aspect of this by studying histone modifications induced by MDV in line 6_3 and line 7_2 chickens at various time points of the disease. Our results could have potentially far-reaching consequences on our understanding of the epigenetics of disease resistance.

Rationale and Significance

ChIP-Seq combines traditional ChIP with next-generation sequencing to form a powerful experimental framework that targets specific histone modifications across the genome. This technique is highly efficient and suited to the study of complex biological phenomena, such as, MD. Methods of ChIP-Seq analysis often make assumptions about the distribution of ChIP-Seq data for computational efficiency. These assumptions have been shown to be inadequate, which limit their sensitivity to diverse enrichment patterns observed in the data. It is necessary to develop an

advanced method or strategy to overcome the unreasonable distribution inference. Thus, the first goal of this project is to develop an efficient and sensitive method of ChIP-Seq analysis that does not make any distributional assumptions. This goal can be achieved with the help of spectral analysis techniques, such as, the wavelet transform and Monte Carlo sampling procedures. Further, our experimental model of MD resistant and susceptible inbred chicken lines provides a unique data set that we can use to validate this method.

Highly inbred chicken lines with drastically different responses to MDV infection, originating from the lines 6 and 7 described above, have been developed in the Avian Disease and Oncology Laboratory (ADOL), USDA, Michigan – Line 6_3 shows high MD-resistance with very few birds (0-3%) developing tumors; Line 7_2 exhibits MD-susceptibility with virtually all individuals (99-100%) developing tumors. The investigation of histone modification profiles in this unique population of chickens can provide an insight into the epigenetic effects of MDV infection and factors influencing disease predisposition. Thus, the second goal of this work is to investigate genome-wide chromatin signatures induced by MDV infection in this population, with a view to a greater understanding of associated epigenetic factors. This goal can be achieved by the application of existing and novel methods to the analysis of histone modification data generated from the population of inbred chicken lines.

The outcomes of this project will further our understanding of histone modifications in poultry, in general, and the effect of MDV infection on host chromatin signatures, in particular. The development of a robust, sensitive and accurate algorithm for ChIP-

Seq analysis will greatly benefit the scientific community and be useful for many future applications. Hence this project consists of the following 3 parts:

1. To develop a novel method of detecting significant peaks in ChIP-Seq data
2. To investigate the epigenetic differences induced by MDV infection in the thymus
3. To apply existing and novel methods to conduct a temporal analysis of chromatin signatures in the bursa of Fabricius

2. WaveSeq: A Novel Data-driven Method of Detecting Histone Modification Enrichments using Wavelets

Abstract

Chromatin immunoprecipitation followed by next-generation sequencing is a genome-wide analysis technique that can be used to detect various epigenetic phenomena such as, transcription factor binding sites and histone modifications. Histone modification profiles can be either punctate or diffuse which makes it difficult to distinguish regions of enrichment from background noise. With the discovery of histone marks having a wide variety of enrichment patterns, there is an urgent need for analysis methods that are robust to various data characteristics and capable of detecting a broad range of enrichment patterns.

To address these challenges we propose WaveSeq, a novel data-driven method of detecting regions of significant enrichment in ChIP-Seq data. Our approach utilizes the wavelet transform, is free of distributional assumptions and robust to diverse data characteristics such as low signal-to-noise ratios and broad enrichment patterns. Using publicly available datasets we showed that WaveSeq compares favorably with other published methods, exhibiting high sensitivity and precision for both punctate and diffuse enrichment regions even in the absence of a control data set. The application of our algorithm to a complex histone modification data set helped make novel functional discoveries which further underlined its utility in such an experimental setup.

WaveSeq is a highly sensitive method capable of accurate identification of enriched regions in a broad range of data sets. WaveSeq can detect both narrow and broad peaks with a high degree of accuracy even in low signal-to-noise ratio data sets. WaveSeq is also suited for application in complex experimental scenarios, helping make biologically relevant functional discoveries.

Introduction

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) is a powerful experimental framework that enables genome-wide detection of epigenetic phenomena such as histone modifications. Histone modification profiles have diverse characteristics ranging from sharp well-defined peaks surrounding transcription start sites of genes to broad diffuse marks on large genomic regions. This inherent variability makes it difficult to distinguish regions of true enrichment from background noise.

There have been several attempts at solving the problem of finding statistically enriched peaks in ChIP-Seq data. One class of methods focuses on transcription factor ChIP-Seq experiments and uses various features of the data to predict binding regions. For instance, FindPeaks [93] adopts a height threshold together with a simulated random background to find significant peaks, while MACS [92] uses a local Poisson p-value to detect chromatin enrichments. Most of these methods have comparable sensitivity in detecting transcription factor binding sites (TFBSs) and are often used in conjunction with motif-finding algorithms.

While the success of the above set of methods in finding transcription factor binding patterns from ChIP-Seq data is undeniable, histone modification data pose new challenges. Utilization of local features to detect histone modification peaks is difficult due to the relative diffuseness of enrichment patterns. Also, common assumptions of such analyses may not hold in this case. For instance, TFBSs cover a small proportion of the genome, but certain histone marks can be present on much larger genomic fractions. Strong TFBSs are flanked by clusters of sense and anti-sense reads and this information can be leveraged to predict the location of the binding site. However, the diffuse nature of most histone modifications renders this impossible. A combination of such factors has led to a relative paucity of methods to analyze histone modification data. A commonly used tool, SICER [101], fits a Poisson distribution before employing kernel density estimation to predict enriched regions, while a recent study employed a negative binomial regression framework and incorporated genomic covariates to improve ChIP-Seq peak detection [102]. However, with the discovery of an ever-increasing number of histone marks that encompass a wide variety of enrichment patterns, there is a continuing need for improved methods robust to a range of data characteristics.

Wavelets belong to a class of spectral analysis techniques that can extract meaningful information from data by decomposing it into its underlying patterns. The versatility of wavelets has seen them being used in a wide variety of disciplines ranging from image processing to medical diagnostics. Recently, we applied this technique to the analysis of comparative genomics hybridization data [164], utilizing the wavelet property of global pattern quantification to find evolutionary relationships between

copy-number profiles in human and bovine populations. However, wavelets also have excellent spatial resolution and comparing data sets one can not only find differences in frequencies of global patterns but also the precise locations of such variations. This property is highly desirable for genome-wide analyses and is the primary motivation for this work.

We present WaveSeq, a novel data-driven method of ChIP-Seq analysis that utilizes the wavelet power spectrum to detect statistically significant peaks in ChIP-Seq data having punctate or broad enrichment patterns. WaveSeq employs Monte Carlo sampling in the wavelet space to predict regions of true enrichment in ChIP-Seq data. In the absence of a control, a randomized algorithm constrained by the length distribution of putative peaks is used to estimate the background read distribution and predict regions of significant enrichment. The non-parametric modeling approach ensures that WaveSeq is robust to variations in data characteristics (e.g. genome coverage) and produces accurate peak calls for a wide variety of data types.

WaveSeq was applied to ChIP-Seq data of Growth-associated binding protein (GABP), Neuron restrictive silencing factor (NRSF) and trimethylations of histone H3 at lysine 4 (H3K4me3), lysine 27 (H3K27me3) and lysine 36 (H3K36me3), which were chosen to capture significant diversity of enrichment patterns and signal-to-noise ratios (SNRs). We demonstrated that WaveSeq peak calls have high sensitivity and precision for narrow and broad regions over a range of SNRs even in the absence of a control data set. We further exhibited the utility of our approach in a complex experimental setting by analyzing H3K4me3 data from genetically similar chicken lines that exhibit divergent responses to a lymphomatous virus. Differentially marked

regions detected by WaveSeq revealed functional differences between the lines that could contribute to differences in disease prognosis. Thus, we conclude that WaveSeq is a highly sensitive algorithm for ChIP-Seq analysis, with applicability for a diverse range of enrichment patterns.

Materials and Methods

H3K4me3 data from chicken bursa

Two specific-pathogen-free inbred lines of White Leghorn chickens either resistant (6_3) or susceptible (7_2) to MD were hatched, reared and maintained in the Avian Disease and Oncology Laboratory (ADOL, Michigan, USDA). The chickens were injected intra-abdominally with a partially attenuated very virulent plus strain of MDV (648A passage 40) at 5 days after hatch with a viral dosage of 500 plaque-forming units (PFU). Chickens were terminated at 5dpi to collect bursa tissues. All procedures followed the standard animal ethics and use guidelines of ADOL.

Chromatin immunoprecipitation (ChIP) was carried out using bursa from MDV infected and controls birds. About 30 mg bursa samples were collected from three individuals, cut into small pieces (1 mm^3) and digested with MNase to obtain mononucleosomes. PNK and Klenow enzymes (NBE, Ipswich, MA, USA) were used to repair the ChIP DNA ends pulled down by the antibody. A 3′ adenine was added using Taq polymerase and Illumina adaptors ligated to the repaired ends. Seventeen cycles of PCR was performed on ChIP DNA using the adaptor primers and fragments with a length of about 190 bp (mononucleosome + adaptors) were isolated from agarose gel. Subsequently, cluster generation and sequencing using the purified DNA

was performed on the Illumina Genome Analyzer IIx following manufacturer protocols. Sequence reads of length 25 bp were aligned to the May 2006 version of the chicken genome (galGal3) using bowtie version 0.12.7 [80]. Default alignment policies of bowtie were enforced. The antibodies used and the total number of reads obtained for all samples are listed in Appendix I.

Published datasets used in this study

We used five ChIP-Seq data sets for benchmarking purposes [69, 98]. The GABP and NRSF (monoclonal) ChIP-Seq data sets were produced from the human Jurkat cell line while a negative control data set was obtained by reverse crosslinking extracted DNA without the subsequent immunoprecipitation step (RX-NoIP). The H3K4me3, H3K27me3 and H3K36me3 data sets were obtained from murine embryonic fibroblast (MEF) cells. We also utilized a previously published synthetic spike-in data set for testing precision and recall [107]. For two-sample ChIP-Seq analyses of GABP and NRSF, we used the RX-NoIP data set as control. The spike-in data consisted of a human input control data set which was randomly divided into three subsets; reads corresponding to the spikes were added to one of the subsets which constituted the mock ChIP sample while a second subset (without the spike-in reads) served as the control. For the MEF histone modification data no control data sets were used to assess algorithm performance in the absence of control.

The GABP and NRSF data from human Jurkat cells were downloaded from:

- http://mendel.stanford.edu/SidowLab/downloads/quest/

H3K4me3, H3K36me3 and H3K27me3 data from MEFs were downloaded from:

- http://www.broadinstitute.org/scientific-community/science/programs/epigenomics/chip-seq-data

The list of qPCR validated sites for GABP and NRSF were obtained from [105]. The synthetic spike-in data were downloaded from:

- http://bioserver.hci.utah.edu/SupplementalPaperInfo/2008/Nix_EmpiricalMethods/

The "JohnsonSpikeDataHg17Low" data set used for specificity benchmarks was generated using human input control data from [68]. All data was downloaded in aligned format with read lengths of 25 bp for the GABP and NRSF data and approximately 32 bp for the H3K4me3 and H3K27me3 data. All analyses were performed on a 2.66 GHz dual core desktop computer running Windows Vista with 3 GB of RAM, a licensed copy of Matlab v7.4 (R2007a) with the Wavelet Toolbox and R version 2.13.0 [165].

Analysis parameters

Downloaded data consisting of aligned sequence reads were converted to the browser extensible data (BED) format. Redundancies were removed before subsequent analysis. Sequence reads were shifted by 95 bp from the 5' end to represent the center of the DNA fragments obtained from the nucleosome and the linker DNA (≈ 190 bp). Summary read counts were calculated using non-overlapping windows of 200 bp for visualization and normalized to per million mapped reads in each sample.

Five methods were chosen for benchmarking: MACS [92] version 1.3.7.1, FindPeaks [93] version 4.0.15, SiSSRs [64] version 1.4, SICER [101] version 1.1 and RSEG [100]. We downloaded and configured the tested algorithms as follows:

1) FindPeaks v 4.0.15 was downloaded as part of the Vancouver Short Read Analysis Toolkit (VSRAT) from http://vancouvershortr.sourceforge.net. The reads in BED format were first separated into chromosomes using SeparateReads.jar. The following parameters were then used for FindPeaks.jar:

-aligner bed

-dist_type 0 190

2) MACS v 1.3.7.1 was downloaded from http://liulab.dfci.harvard.edu/MACS/. The following parameters were used:

--shiftsize=95

--nomodel True

For applying MACS to histone modification data sets, we used the additional parameter --nolambda as recommended by [166].

3) SiSSRs v 1.4 was downloaded from http://sissrs.rajajothi.com/. The following parameters were used:

-F 190

4) SICER v 1.1 was downloaded from http://home.gwu.edu/~wpeng/Software.htm. The following parameters were used:

Gap size = 2 (H3K4me3), 5 (H3K36me3) and 10 (H3K27me3)

E-value = 100

Window size = 200

5) RSEG was downloaded from http://smithlab.usc.edu/histone/rseg/. The following parameters were used:

-i 20

For the transcription factor binding site detection all methods were configured to have p-value < 0.001 in single sample experiments and p < 0.01 in the presence of matched controls. For uniformity, we set genome size = 3,107,000,000 bp for the GABP and NRSF (hg18) data sets and 2,725,000,000 bp for murine embryonic fibroblast (mm8) histone modification data. Recommended values were used for all other parameters.

Gene annotation and functional analysis of differentially marked regions (DMRs)

RefSeq and Ensembl gene annotations for the chicken genome (galGal3) were downloaded from the UCSC genome browser [167]. Gene promoters were searched for overlaps with DMRs and all gene names were converted to their Ensembl IDs using the biomart data retrieval system from Ensembl [168, 169]. This unified list of gene IDs was then analyzed for functional annotation enrichment with DAVID [170]. Default parameters were used for DAVID analyses.

Software implementation

Data pre-processing, Monte Carlo estimation of wavelet coefficient thresholds and peak-calling modules of WaveSeq were implemented in Matlab. FDR estimation in the presence and absence of control was performed in R. We are currently working on a unified R implementation of the software for public release. WaveSeq can be run on a standard desktop computer with at least 3 GB of RAM and a 2 GHz processor. The software can be used on any species with a sequenced genome. WaveSeq has been tested on Windows, UNIX and MAC OSX and available from the authors on request.

Results

Wavelets for ChIP-Seq analysis

The wavelet transform has great utility in data compression and pattern finding, the latter involving the choice of a suitable 'mother' wavelet ψ to best capture underlying patterns in the data. An example of a mother wavelet is the Morlet wavelet, defined as the product of a Gaussian envelope and a cosine wave:

$$\psi_0(t) = \pi^{-1/4} e^{-t^2/2} \cos \omega_0 t$$

where, t is the genomic location and ω_0 is the non-dimensional frequency (Figure 2.1A). The wavelet transform may be either continuous or discrete – the continuous wavelet transform (CWT) is highly redundant and resistant to data loss while the discrete transform is less computationally intensive but more prone to information loss. The peaks observed in ChIP-Seq data are relatively smooth, making it better suited to the application of the CWT.

The CWT consists of the convolution of a translated and scaled mother wavelet $\psi_0(t)$ to the signal x_t at a predefined step-size (δ) as follows:

$$W_t(s) = \sum_{t'=0}^{T-1} x_{t'} \psi * \left[\frac{(t'-t)\delta}{s} \right]$$

where, (*) indicates the complex conjugate, s is the wavelet scale and t' denotes translation along the genome. The wavelet scale s is representative of the size of the scaled wavelet and the mathematical formulation of the transform implies an inverse relationship, i.e. the higher the scale, the smaller the scaled wavelet. The wavelet decomposition produces a series of 'wavelet coefficients', real numbers that indicate

the correlation between the mother wavelet and the data, which may be either positive or negative. This is also a multi-scale decomposition, i.e. the coefficients at different scales represent the correlation of scaled versions of the wavelet to the signal. Therefore, smaller localized patterns are likely to be captured by higher scales of the transform and vice-versa.

A natural way of quantifying the wavelet decomposition is the wavelet power spectrum, defined as the square of the wavelet coefficients, and synonymous with the 'energy density'. A contour plot of the wavelet power spectrum for ChIP-Seq data revealed hot-spots that correlated with peaks (Figures 2.1 B, C). This suggested that wavelets could be used to detect enrichment regions in this type of data and inspired us to use this approach for ChIP-Seq analysis.

WaveSeq overview

We introduce WaveSeq, a novel method of ChIP-Seq peak detection that utilizes the wavelet power spectrum (Figure 2.1 D). Sequence reads are first 'shifted' to represent the center of DNA fragments obtained from the ChIP experiment. The genome is divided into non-overlapping windows and read counts for each window calculated. The summary read counts are the primary input data format used by WaveSeq. Typical analyses can be of two types: (i) single sample experiment – without control, and (ii) two-sample experiment – with matched control samples.

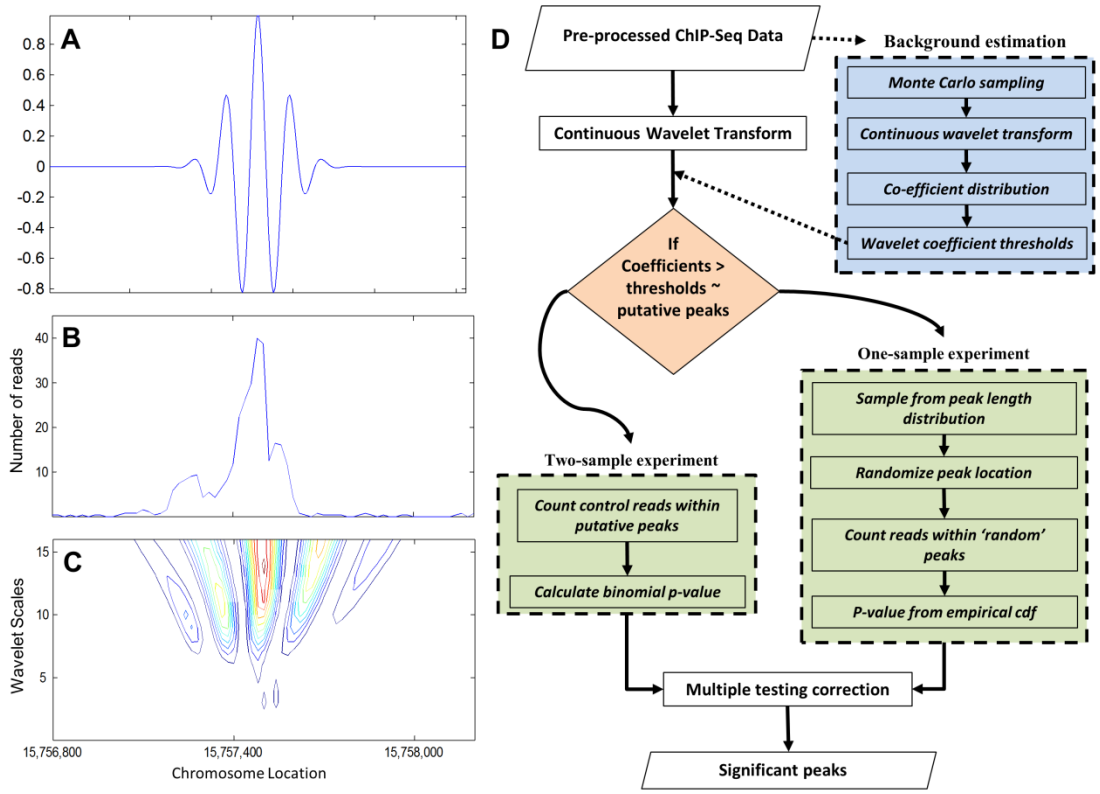


Figure 2.1. WaveSeq utilizes the continuous wavelet power spectrum to detect peaks in ChIP-Seq data.
(a) A scaled representation of the morlet wavelet. (b & c) H3K4me3 data and a contour plot of the associated wavelet power spectrum shows hot spots that correlate with ChIP enrichments. The ChIP-Seq data represents the 15,756,800 – 15,758,200 bp region of the mouse chromosome 1 from the MEF H3K4me3 data set. (d) A schematic of the WaveSeq analysis pipeline. The workflow consists of two major modules: (i) the Monte Carlo background estimation step and (ii) significance estimation from randomized algorithm using the peak length distribution (one-sample experiment) or an exact binomial test (two-sample experiment).

For both analyses, we first employ a Monte Carlo sampling technique for modeling the data [171]. *N* random samples are drawn from the ChIP-Seq data and the wavelet power calculated for each instance. A slice of the power spectrum at a fixed point of each random sample is used to generate an empirical distribution of wavelet powers for each scale. This distribution enables us to obtain a suitable significance threshold, which is applied to the wavelet transform of read count profiles to detect windows

having significant enrichment. Our thresholding procedure is, therefore, dependent on the *wavelet fit* to the data at a particular position and distinct from a simple read-count cutoff.

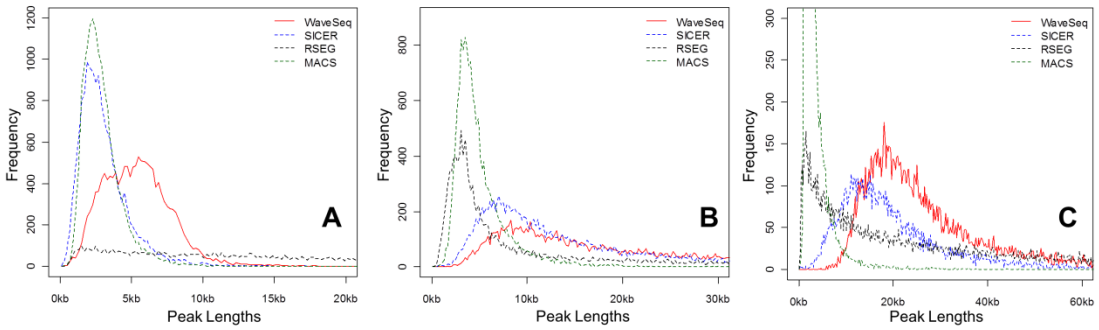


Figure 2.2. Peak length distributions of tested methods when applied to histone modification data.
A comparison of peak length distributions for the top 15000 peaks called from the (a) H3K4me3, (b) H3K36me3 and (c) H3K27me3 data. (a) SICER and MACS have similar peak lengths in the H3K4me3 data, followed by WaveSeq. RSEG peak lengths are almost uniformly distributed between 0 and 20 kb. (b) MACS and RSEG called relatively short peaks for H3K36me3 while SICER and WaveSeq detected greater peak lengths. (c) WaveSeq called the longest peaks when applied to H3K27me3 data followed by SICER and RSEG.

To further account for broad peaks seen in histone modification data, our algorithm implements a 'gap' parameter, *g*. We define a 'gap' as a window having a non-significant wavelet power (non-significant window); for example, if *g* is set to two, peaks separated by at most two non-significant windows are aggregated together. This parameter is necessary for two reasons: (i) chromatin enrichments, especially broad marks, such as, H3K36me3 and H3K27me3, can be discontinuous and (ii) wavelets are very sensitive to boundary events and local fluctuations. A strong enrichment region interspersed with areas of low read counts could, therefore, result in multiple peak calls and the gap parameter of WaveSeq helps to reduce the effect of this scenario. This parameter is similar in principle to that used in SICER, but with

one major distinction. SICER also imposes an upper limit on allowable non-significant windows within a significant peak. While this results in an elegant closed form expression for estimating statistical significance from the score distribution, in practice, this results in smaller peak lengths for the same value of *g* (Figure 2.2).

One-sample experiment

The estimation of statistical significance is crucial to ChIP-Seq analysis approaches to filter the results of genome-wide studies, particularly in the absence of a control. For a single-sample experiment, WaveSeq utilizes the length distribution of putative peaks to estimate the likelihood of observing a peak with a given number of reads.

A large number of peaks, *P*, are sampled with replacement from the length distribution of putative peaks, and their positions on the genome randomized. The number of reads within each randomized peak is counted, generating the empirical distribution, *F(R)*, for the number of peaks having a given read count *R*. The probability of observing a peak with read count *r* is:

$$\Pr[\#\text{reads in a peak} = r] = \frac{F(r)}{P}$$

and the p-value of observing this peak is,

$$p(r) = \frac{(\#\text{peaks with total reads} \geq r)}{P} = \frac{1}{P}\sum_{R=r}^{\infty} F(r)$$

The p-values are subsequently corrected for multiple-testing using the Benjamini-Hochberg FDR procedure [172].

Most ChIP-Seq experiments produce sparse enrichment regions covering a small fraction of the genome and therefore, only few of the randomized peak locations

would be likely to overlap significantly enriched regions. However, this is not always the case – histone modifications such as, H3K27me3, mark large regions for silencing and could occupy a significantly greater genomic fraction. In the latter case, a higher proportion of randomized peaks would potentially overlap 'true' enrichment regions – but this is a fair reflection of a relatively low SNR data set where the boundaries between true signal and background are blurred.

Thus, it is important to note that in predicting areas of true enrichment in ChIP-Seq data, we do not make any assumptions about the read distribution, instead relying on Monte Carlo sampling techniques – first, to construct the empirical distribution of wavelet coefficients and second, to assign significance scores to predicted enriched regions using a randomized algorithm constrained by the peak length distribution. In addition, the association of statistical significance of a peak with its read count provides a natural and interpretable criterion for thresholding genome-wide analyses where the number of reads mapping to a region is often indicative of the presence of a true biological signal.

Two-sample experiment

If a ChIP-Seq experiment has matched controls, WaveSeq uses the binomial distribution to compare read counts between normalized test and control samples. For each putative peak, reads in the corresponding region of the control data (C) are counted and compared to the test sample (T) using a two-sided exact binomial test. A putative peak can be considered to be a Bernoulli experiment with $t = (C + T)$ trials wherein the number of reads in the test sample T is the number of successes. The proportion of successes, $p = T/(C+T)$ and failures, $q = 1 - p$. In this case, the

probability of observing at least T successes in t trials under the null hypothesis, H_0: $p = 0.5$, is given by the expression,

$$\Pr[\,\#\text{ of successes in } t \text{ trials} \geq T\,] = \sum_{i=T}^{t} \binom{t}{i} p^i (1-p)^{t-i}$$

The p-values for the list of putative peaks are subsequently corrected for multiple testing as above [172].

Choice of parameters

Systematic tuning of the WaveSeq peak-calling algorithm was carried out. We applied several different wavelet mother functions to ChIP-Seq data e.g. Morlet, Coiflets 1 and 2 and Mexican hat, to find the wavelets most suited to the data sets. All wavelets performed comparably when applied to punctate ChIP-Seq data sets but the morlet wavelet outperformed the others in detecting enrichment regions upto ~10kb while the Mexican hat wavelet was the most effective in calling very broad peaks (e.g. H3K27me3). A comparison of the energies at the various scales of the wavelet transform showed a higher density in a smaller band for the morlet wavelet and a more uniform distribution for the Mexican hat wavelet (Figure 2.3). The energy compression characteristic of the morlet wavelet represents a higher discriminative power over a smaller subset of scales and explains its performance for relatively strong enrichment patterns. The diffuse distribution of Mexican hat, however, is a better fit for the dispersed H3K27me3 marks as evidenced by its greater sensitivity for this dataset. Therefore, we used the morlet wavelet for GABP, NRSF, H3K4me3 and H3K36me3 data and the Mexican hat wavelet for the H3K27me3 data.

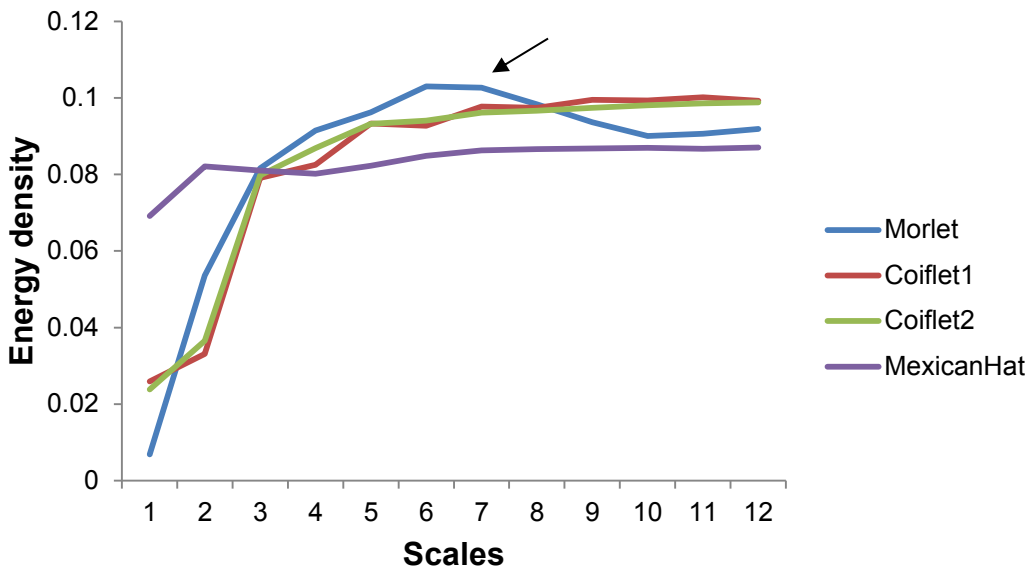


Figure 2.3. Comparison of wavelet energies for different wavelets.
A comparison of wavelet energies shows a higher density in a smaller band of scales for the Morlet wavelet as shown by the arrow-head. Other wavelets have more broadly distributed energy densities. Bursa H3K4me3 data from chromosome 2 of the S.inf group was used to obtain the above.

We assessed the effect of the number of samples (N) in the Monte Carlo threshold estimation step. The wavelet coefficient thresholds quickly reached saturation for all scales (Figure 2.4). Therefore, we chose $N = 5000$ for optimal accuracy and speed. The sampling was performed chromosome-by-chromosome. There was marked variation in wavelet coefficient thresholds for different chromosomes at a specified p-value (Figure 2.5). There are two possible reasons for this: the number of enrichments on a specific chromosome and the chromosome size. The first arises out of the natural variation of different data sets and the latter out of the particular choice of the length and number of samples. In either case, this variation represents important information about the data and we account for it in our algorithm as follows: The mean and standard deviation of wavelet coefficient thresholds for each scale across the chromosomes were calculated and wavelet coefficients from the wavelet transform of

the data were considered significant at the specified p-value if it was greater than the mean + standard deviation. The p-value for a significant wavelet power at a window was chosen to be p_{thres} = 0.2 for punctate data sets (transcription factors and H3K4me3) and p_{thres} = 0.4 for broad marks (H3K36me3 and H3K27me3).

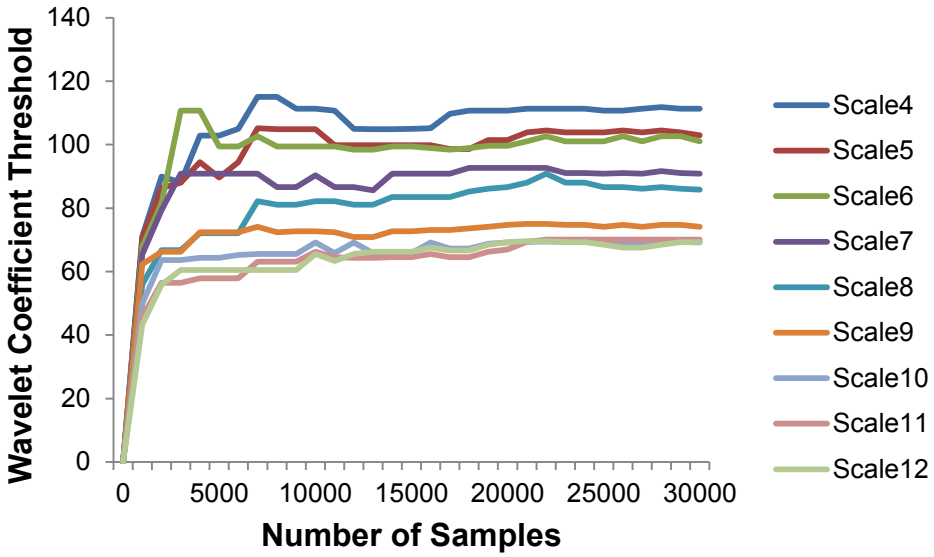


Figure 2.4. Wavelet coefficient thresholds reach saturation quickly. Morlet wavelet thresholds at p < 0.001 of H3K4me3 data from chromosome 1 in the chicken bursal samples (S.inf group, window size = 200 bp).

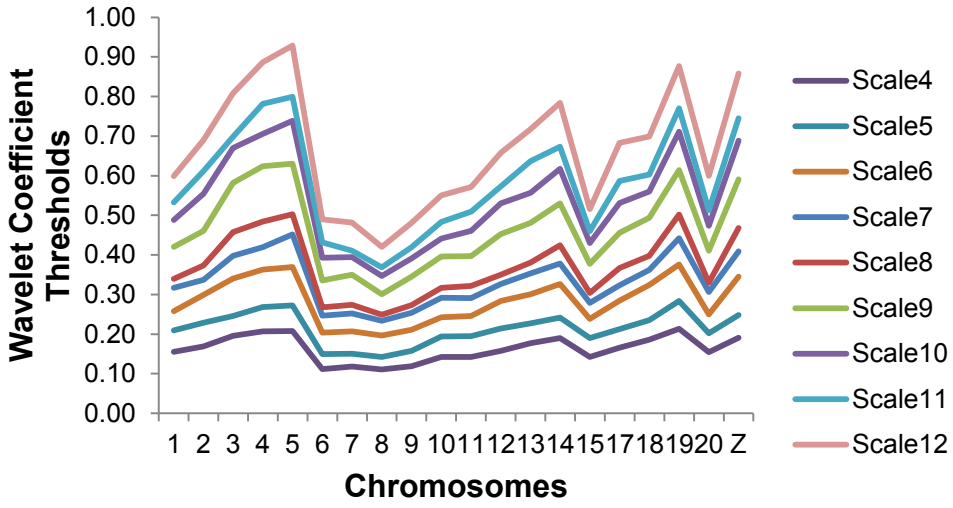


Figure 2.5. Comparison of wavelet coefficient thresholds for different chromosomes (p = 0.2). A variation is observed in the wavelet coefficient thresholds for different chromosomes of the chicken genome. Various factors may be responsible for this observation ranging from different chromosome lengths, the number of enrichment regions and the choices of size and number of samples. Bursa H3K4me3 data from the S.inf group was used for the above plot.

Wavelet coefficient thresholds were larger for greater sample sizes but the effect was more pronounced for smaller chromosomes (Figure 2.6). This was possibly due to oversampling effects as the increase in wavelet coefficients was inversely correlated with chromosome size. We found a strong negative power law correlation between chromosome size and wavelet coefficient thresholds for sample length 2^{15} (R^2 = 0.7765, Figure 2.7) which was absent for smaller samples (2^{12}: R^2 = 0.0299; 2^{10}: R^2 = 0.1198). Greater sample lengths, therefore, are biased by chromosome size that could lead to large variations in coefficient thresholds. A smaller sample, on the other hand, could lead to lower thresholds and possibly more false positives. These two effects appeared to be reduced at a sample size of 2^{12} and hence we chose this for subsequent

experiments. To further minimize the effect of chromosome size, we only considered chromosomes that have at least twice the length of the sample.

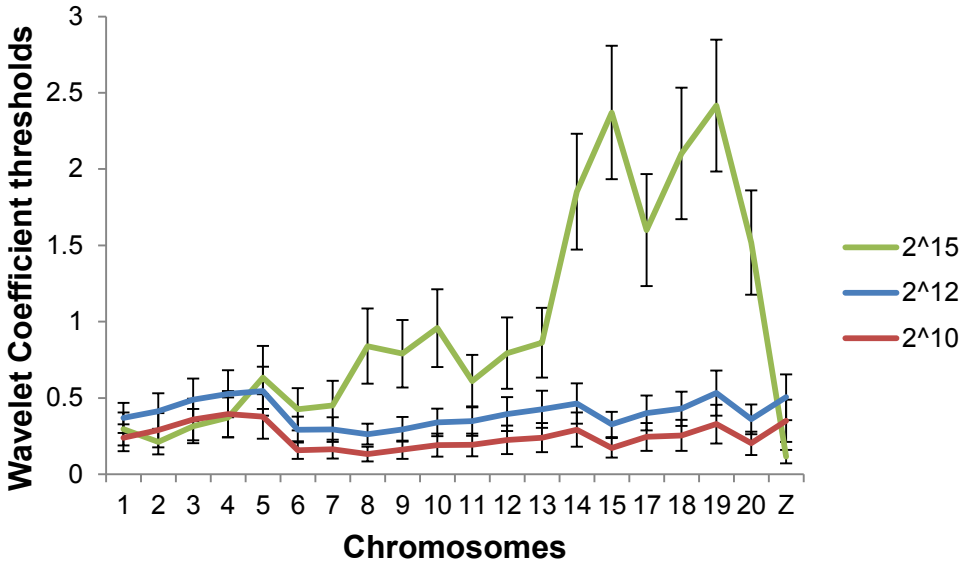


Figure 2.6. Effect of sample length on wavelet coefficient thresholds (p = 0.2). At higher sample sizes, wavelet coefficient thresholds are larger but the effect is only noticeable for the smaller chromosomes (13-20) of the chicken genome and is possibly due to oversampling. There is little difference between sample size of 2^{10} and 2^{12}. The error bars depict the standard errors over 9 scales (4-12) for sample lengths 2^{12} and 2^{15} and 7 scales (4-10) for sample length 2^{10}. The data corresponds to bursa H3K4me3 from the S.inf group.

The minimum scale considered for peak calling was $s = 4$, since lower scales are representative of broader patterns that are more likely to be background noise. We also noticed that a significant ChIP-Seq peak was significant at several scales simultaneously (See Figures 2.1 B, C) while localized peaks had fewer significant scales. Therefore, to further eliminate spurious peak calls due to local fluctuations, a window was considered significant only if there were at least 2 significant scales for the window. For estimating FDR in one-sample analyses, we used number of simulated peaks, $P = 10^6$.

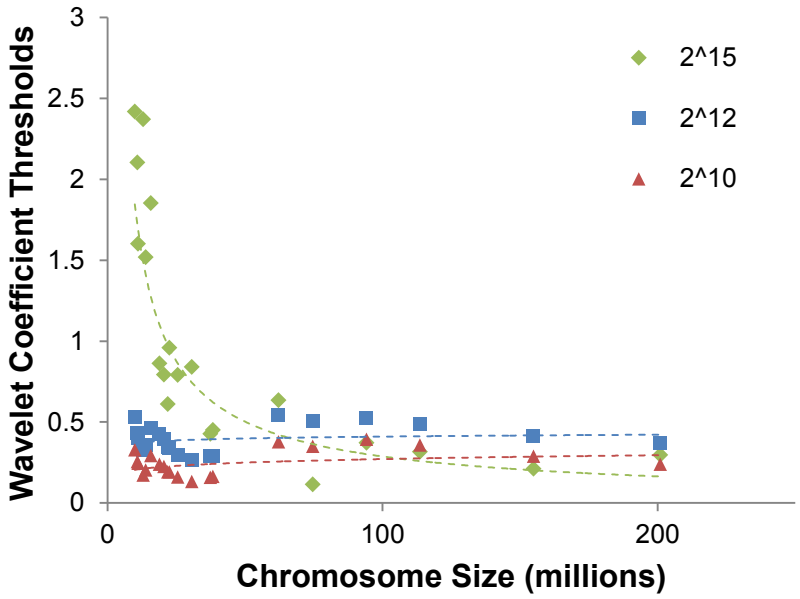


Figure 2.7. Correlation of chromosome size and wavelet coefficient thresholds. Large sample lengths have a strong negative correlation with chromosome size which follows a power law distribution. This correlation is absent for smaller samples. The dotted lines represent power law regression lines for different sample lengths. The data corresponds to bursa H3K4me3 from the S.inf group.

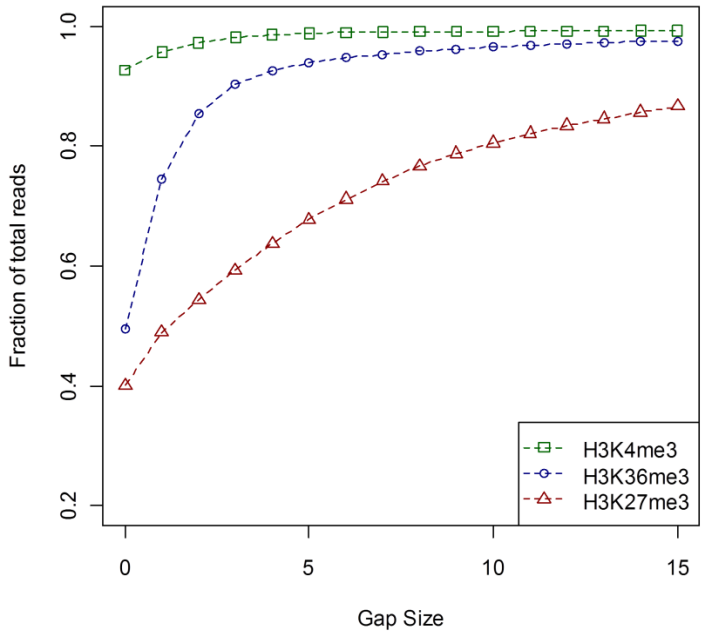


Figure 2.8. The effect of increasing gap sizes on read coverage of top peaks. The fraction of reads covered by the top N peaks saturates at larger gap sizes. This saturation is almost immediate for H3K4me3, intermediate for H3K36me3 and more gradual for H3K27me3. In the case of H3K4me3, N = 20000, while for H3K36me3 and H3K27me3, N = 40000. The window size is 200 bp.

The choice of a suitable gap size is dependent upon multiple factors including histone mark characteristics and sequencing depth. The read coverage fractions for different histone marks appear to saturate with increasing gap sizes (Figure 2.8). However, the saturation rate is highly variable between marks - H3K4me3 shows little change with increasing gap sizes, H3K27me3 exhibits a gradual increase while the pattern for H3K36me3 is intermediate between the two, in keeping with the intermediate characteristics of the mark. The above comparison shows that a gap size of 0 to 400 bp (0-2 200 bp windows) would be suitable for the H3K4me3 data set while larger gap sizes may be more appropriate for the broader histone marks e.g. $g = 5$ for H3K36me3 and $g = 10$ for H3K27me3. A similar comparison of read coverage saturation rates can, therefore, help the user choose a gap size appropriate for a particular data set.

Comparison with other methods using published data

Recent studies have compared the performance of several published ChIP-Seq peak calling algorithms [104, 105]. From the list of methods tested in the above studies, we chose five commonly used tools: FindPeaks, MACS and SiSSRs [64], which were developed primarily for detecting transcription factor binding sites (TF-methods) along with SICER and RSEG [100] which were specifically aimed at chromatin enrichment data (CH-methods). A variety of ChIP-Seq data sets were selected to compare the performance of WaveSeq with the above methods including GABP, NRSF [98], H3K4me3, H3K27me3, H3K36me3 [69] and a synthetic spike-in data set [107].

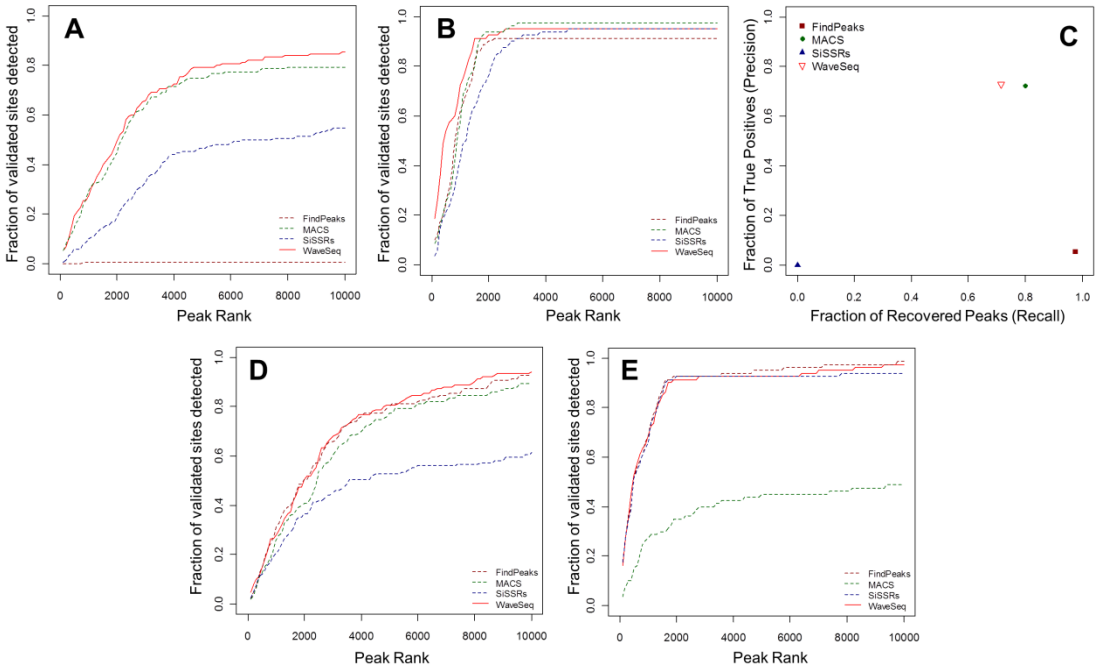


Figure 2.9. WaveSeq has high sensitivity and precision for punctate data sets. (a & b) Plots of peak ranks against the fraction of validated sites detected by WaveSeq, FindPeaks, MACS and SiSSRs for the (a) GABP and (b) NRSF data sets. (c) A plot of the fraction of true positives (precision) against the fraction of recovered peaks (recall) for the synthetic spike-in data set. (d & e) Sensitivity plots for the (d) GABP and (e) NRSF data sets shows that WaveSeq has high sensitivity for these data sets even in the absence of control.

WaveSeq has high sensitivity

Several GABP and NRSF binding sites have been validated with qPCR [105] allowing us to compare the sensitivities of the TF-methods with that of WaveSeq using the corresponding ChIP-Seq data. The peaks called by each TF-method were ranked by significance scores output by the method and tested for overlap with the validated sites. Subsequently, we plotted the peak rank against the fraction of validated sites detected by each algorithm (Figures 2.9 A, B).

WaveSeq had the highest sensitivity among tested methods for both data sets. In the case of GABP, WaveSeq had the best performance closely followed by MACS which

had slightly lower recall. SiSSRs came in third but still significantly outperformed FindPeaks which had low sensitivity for this data set. On the other hand, all the methods had similar performance on the NRSF data. WaveSeq showed marginally higher sensitivity with MACS, FindPeaks and SiSSRs performing comparably. A further comparison of peak lengths showed that MACS, FindPeaks and WaveSeq had similar peak length distributions while a majority of SiSSRs peaks were very small (Figure 2.10).

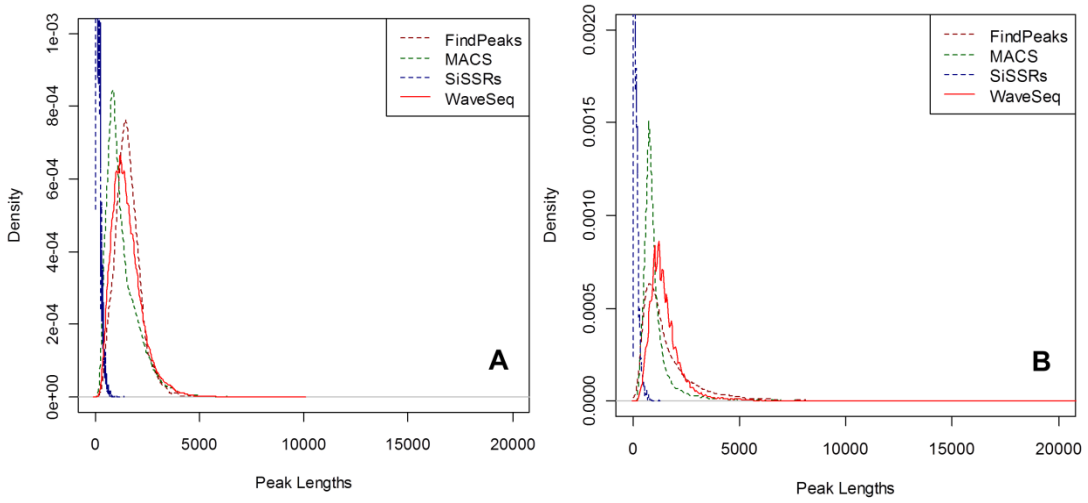


Figure 2.10. WaveSeq has comparable peak lengths to MACS and FindPeaks in punctate data sets.
A comparison of peak length densities of the top 20000 peaks for the (a) GABP and (b) NRSF data sets for WaveSeq, MACS, FindPeaks and SiSSRs.

WaveSeq has good precision

It is difficult to evaluate the specificity of ChIP-Seq peak-calling algorithms due to the unavailability of adequate 'true-negative' binding sites for systematic analysis. However, one can estimate the false positive rates using synthetic data sets which contain simulated binding events. For this analysis we utilized a published synthetic data set generated from human input control data that was 'spiked' with simulated

reads at fixed locations [107]. We applied WaveSeq and the TF-methods to this data set and plotted the proportion of recovered peaks (recall) against the fraction of true positives (precision) (Figure 2.9 C).

MACS had the best combination of precision (0.724) and recall (0.799), closely followed by WaveSeq which had slightly better precision (0.728) but lower recall (0.716). However, FindPeaks had a very high number of false positives (precision = 0.06) in this test while SiSSRs failed to detect any peaks.

WaveSeq performs well even without a control data set

The data from a matched input control sample is considered to improve the power of a ChIP-Seq experiment by reducing systematic biases [77]. However, matching input controls are often not available and negative controls such as IgG that bind in a non-specific manner, can give rise to additional sources of error. Moreover, it is not clear if the use of input alone can offset the effect of various confounding factors such as mappability and G/C content. Therefore, it is important to assess the performance of ChIP-Seq peak callers in the absence of a matched control.

We compared the sensitivity of TF-methods and WaveSeq using the GABP and NRSF data sets as above, but without the use of control data (Figures 2.9 D, E). WaveSeq again had high sensitivity for both data sets, almost identical to FindPeaks which performed much better on these data sets without control. SiSSRs and MACS had mixed results; the former had similar performance to FindPeaks and WaveSeq for the NRSF data set, but lower sensitivity for the GABP data, while the situation was

reversed for MACS. Thus, WaveSeq has high accuracy for punctate peaks and was the only method that performed consistently well for the tested data sets.

WaveSeq improves detection of broad histone modification peaks

A lack of adequate validated sites for histone modification data makes it difficult to assess the performance of analysis methods on these data sets. However, we can argue that if multiple methods of analysis based on different detection algorithms predicted significant enrichment in a particular region, it was more likely that a true region of enrichment existed in that region. Indeed, studies have shown that a smaller number of peaks generated by certain methods were largely contained within larger peak lists called by other methods, indicating a common set of peaks detected by most algorithms [105]. With the above intuition we ran the CH-methods on the MEF histone modification data sets. We included MACS in the latter as it has been used for broad peak calling [166], even though it was originally developed for the analysis of transcription factor ChIP-Seq data. The top peaks (15000 for H3K4me3 and 20000 for H3K36me3 and H3K27me3) called by each of the above programs were compared and regions detected by at least two peak-callers were defined as putative 'true positives'. When calculating putative true positive peaks, we did not enforce any restrictions on the overlap, i.e. if there was even a single bp overlap between two peak calls, these regions were merged together (union) into a putative positive peak. This is because peak-calling algorithms will sometimes call only part of a putative histone modification enrichment as a peak, and merging adjacent peak-calls is likely to produce a better reflection of enrichment patterns.

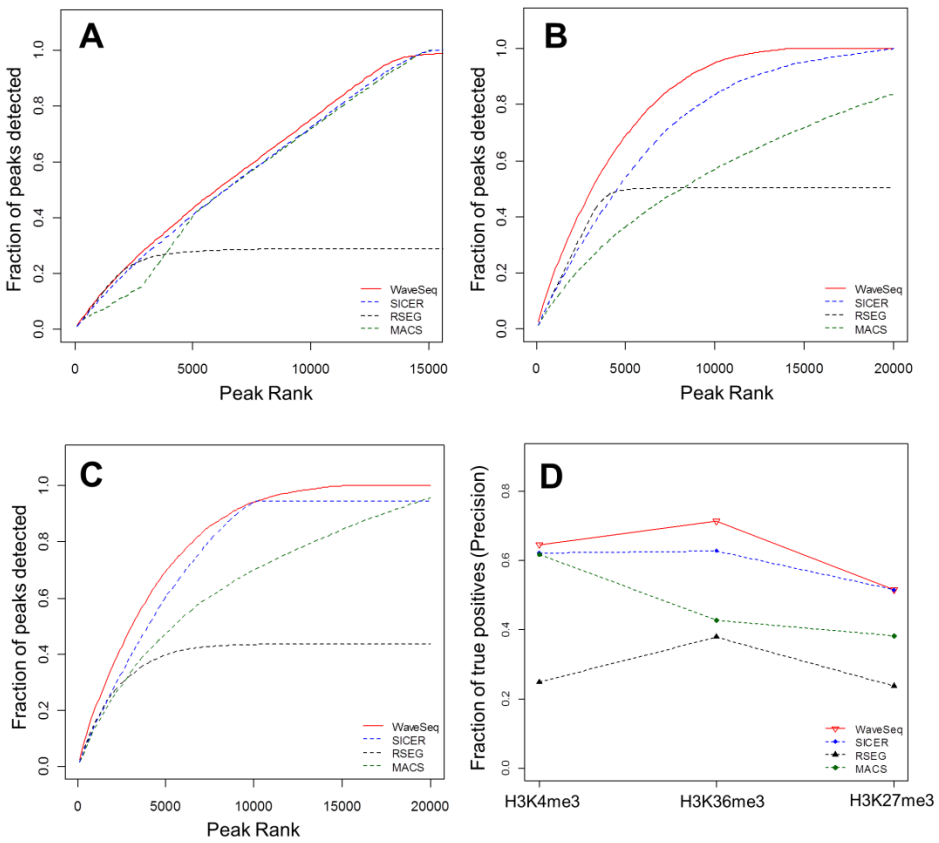


Figure 2.11. WaveSeq improves detection of histone modification peaks.
(a, b & c) Plots of peak ranks against the fraction of putative 'true positive' sites detected by WaveSeq, SICER, RSEG and MACS for the (a) H3K4me3, (b) H3K36me3 and (c) H3K27me3 data sets. (d) A plot of the fraction of true positives (precision) from the top 10000 peaks detected by the above four methods in the MEF histone modification data sets.

The above procedure yielded 8592, 7522 and 5463 peaks for the H3K4me3, H3K36me3 and H3K27me3 data sets, respectively. These peaks were compared with the peak lists from all methods (SICER, RSEG, MACS and WaveSeq) and relative performance was assessed by comparing the fraction of recovered peaks against peak ranks (Figures 2.310 A-C). For punctate H3K4me3 data, all methods apart from RSEG performed well, with near-identical recall rates. WaveSeq had the best sensitivity on the H3K36me3 and H3K27me3 data sets with SICER coming in second. MACS showed lower recall rates for these two data sets while RSEG

detected the top peaks with good accuracy but was unable to detect any peaks in chromosomes 10-19. A fFurther, WaveSeq had the highest precision in all three data sets analysis of precision (Figure 2.310 D) showed that WaveSeq had the highest performance in all three data sets.

Pair-wise comparisons between peaks detected by WaveSeq and those called by SICER and MACS showed a high degree of overlap (98-100%) across all the data sets. In the case of RSEG the overlap was lower (20-68%) but closer examination revealed that a majority of regions not called by WaveSeq, particularly in the H3K4me3 and H3K36me3 data sets, had low average read counts and were possibly false positives (Appendix II). WaveSeq also called larger peaks on average compared to SICER, particularly in the H3K27me3 and H3K4me3 data sets (Figure 2.2). However, RSEG detected very broad regions in both H3K27me3 and H3K4me3 data. Since this algorithm was developed with the express purpose of detecting dispersed chromatin domains, the above behavior is expected, although very long peaks in punctate ChIP-Seq data may not be desirable. Also, somewhat surprisingly, WaveSeq and SICER had greater average peak lengths compared to RSEG for the H3K36me3 data. MACS, on the other hand, detected very small peaks in all the data sets, proving its general unsuitability for broad histone marks.

Thus, WaveSeq once again showed the highest sensitivity of all tested methods across a variety of histone modification data sets. While there was little to choose between the different algorithms for the punctate high SNR H3K4me3 data, WaveSeq outperformed the other tested methods in the analysis of broad enrichment regions characteristic of broad marks such as H3K27me3 and H3K36me3. SICER comes in

second while MACS has low sensitivity for diffuse data. RSEG has good sensitivity for the strongest peaks but has low recall, failing to detect any peaks in chromosomes 10-19.

Analysis of complex histone modification data

The bursa of Fabricius is a specialized immune organ that is the site of haematopoiesis and B cell development in chickens. This tissue is one of the first targets of Marek's disease virus (MDV), a herpesvirus that induces T-cell lymphomas in susceptible birds. Genetically similar lines of chickens that show differential resistance to Marek's disease (MD) have been developed and studied for decades, but the exact causes of the divergent response have not been found, although it is believed that epigenetic factors play an important role in determining the level of resistance of an individual. This is an interesting epigenetic model for human cancers as individuals having high genetic similarity exhibit natural resistance to a cancer-causing agent. Moreover, this is a complex ChIP-Seq experiment representing studies in non-traditional systems that are becoming more prevalent with the plummeting costs of sequencing. To demonstrate the utility of WaveSeq in such an experimental scenario we used it to analyze H3K4me3 profiles in matched infected and control birds from inbred chicken lines having diverse responses to MD.

WaveSeq detects differential H3K4me3 marks induced by virus infection

We generated H3K4me3 ChIP-Seq data from inbred chicken lines – line 6_3 is highly resistant while line 7_2 is highly susceptible to MD – in matched infected and control groups. In the subsequent discussion, we refer to the resistant line 6_3 and susceptible

line 7_2 as R and S groups, respectively. We first analyzed the infected group with the non-infected group as control. The samples were then swapped to account for significant peaks in the control that were absent in the infected group. This is in contrast to traditional ChIP-Seq experiments where peaks detected in an input control represent false positives and are removed from subsequent analyses. Statistical significance for differentially marked regions (DMRs) was defined at a false discovery rate of 5% (FDR < 0.05). DMRs were compared across the control-swapped comparisons and merged into a single non-redundant list.

WaveSeq detected a comparable number of peaks in the two groups, with 25050 and 27169 peaks in the R and S groups, respectively. The resistant line did not show any differential H3K4me3 marks at the predefined significance level. In contrast, there were 310 H3K4me3 DMRs in the susceptible line, all but five of which were more enriched in infected individuals. This confirmed the presence of dramatic differences in the epigenetic effects of MDV on the two lines, with a predominantly activating effect of the virus infection.

Table 2.1. Functional annotation of genes having H3K4me3 DMRs

Gene Ontology Term	***Count***	***p-value***	***FDR (%)***
GO:0002520: Immune system development	15	1.91×10^{-8}	3.02×10^{-5}
GO:0030097: Hemopoiesis	14	2.16×10^{-8}	3.41×10^{-5}
GO:0048534: Hemopoietic or lymphoid organ development	14	8.76×10^{-8}	1.38×10^{-4}
GO:0045580: Regulation of T cell differentiation	7	8.60×10^{-7}	0.001359
GO:0002521: Leukocyte differentiation	10	1.11×10^{-6}	0.001747
GO:0045582: Positive regulation of T cell differentiation	6	1.23×10^{-6}	0.001951
GO:0045321: Leukocyte activation	11	1.70×10^{-6}	0.002693
GO:0045619: Regulation of lymphocyte differentiation	7	2.39×10^{-6}	0.003781
GO:0002684: Positive regulation of immune system process	10	2.73×10^{-6}	0.004309
GO:0045621: Positive regulation of lymphocyte differentiation	6	3.33×10^{-6}	0.005262

GO:0046649: Lymphocyte activation	10	4.70×10^{-6}	0.007428
GO:0050870: Positive regulation of T cell activation	8	5.53×10^{-6}	0.008734
GO:0001775: Cell activation	11	6.17×10^{-6}	0.009752
GO:0051251: Positive regulation of lymphocyte activation	8	8.08×10^{-6}	0.012774
GO:0002696: Positive regulation of leukocyte activation	8	1.16×10^{-5}	0.018257
GO:0050867: Positive regulation of cell activation	8	1.62×10^{-5}	0.025558
GO:0050863: Regulation of T cell activation	8	1.62×10^{-5}	0.025558
GO:0030098: Lymphocyte differentiation	8	1.90×10^{-5}	0.030027
GO:0051249: Regulation of lymphocyte activation	8	2.59×10^{-5}	0.040908
GO:0030217: T cell differentiation	7	2.76×10^{-5}	0.04356
GO:0002694: Regulation of leukocyte activation	8	4.00×10^{-5}	0.063158
GO:0045058: T cell selection	5	6.65×10^{-5}	0.105094
GO:0050865: Regulation of cell activation	8	6.80×10^{-5}	0.107401
GO:0002252: Immune effector process	6	1.38×10^{-4}	0.218176
GO:0033077: T cell differentiation in the thymus	5	2.30×10^{-4}	0.362727
GO:0042110: T cell activation	7	2.43×10^{-4}	0.38295
GO:0042981: Regulation of apoptosis	14	2.47×10^{-4}	0.389793
GO:0043067: Regulation of programmed cell death	14	2.98×10^{-4}	0.469488
GO:0010941: Regulation of cell death	14	3.12×10^{-4}	0.491456
GO:0033554: Cellular response to stress	12	4.00×10^{-4}	0.629557
GO:0045061: Thymic T cell selection	4	4.06×10^{-4}	0.639966

The top functional categories (FDR < 1%) enriched among genes having H3K4me3 DMRs from DAVID shows a large number of immune-related functions. Count refers to the number of genes in the gene list annotated with the given GO ID. P-values were obtained from a modified Fisher exact test performed by DAVID which tests the enrichment of the corresponding functional category in the given gene list against the population (chicken genome). FDR correction was performed using the Benjamini-Hochberg procedure [172].

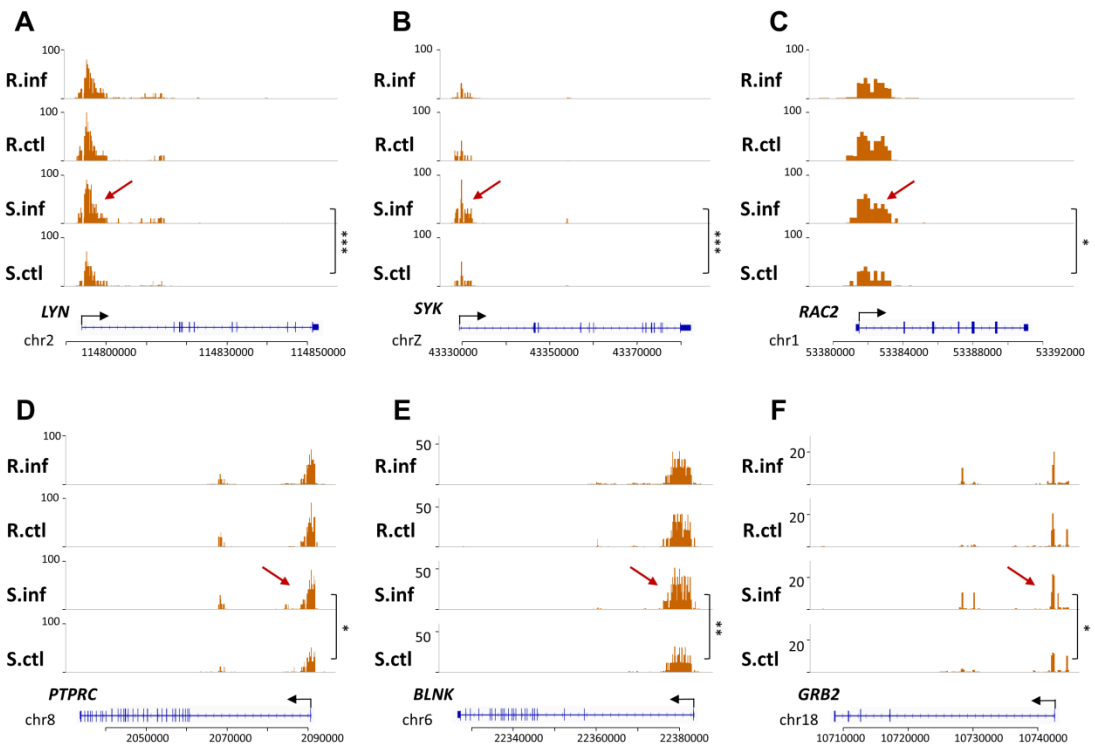


Figure 2.12. Differentially marked regions detected by WaveSeq suggest increased B cell activation in susceptible chickens.
Several genes involved in the B cell activation such as *LYN* (a), *SYK* (b), *RAC2* (c), *PTPRC* (d), *BLNK* (e) and *GRB2* (f) show increased levels of H3K4me3 in infected birds from the S group as shown by the arrowheads. In contrast, there are no significant changes in the R group. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. S.inf = infected S group, S.ctl = control S group, R.inf = infected R group, R.ctl = control R group.

Increased B cell activation in susceptible birds as a result of MD

To investigate the functional implications of observed epigenetic differences, we searched for overlaps between H3K4me3 DMRs and gene promoters and were able to map 241 regions to 310 Ensembl genes (Appendix III). Functional annotation of these genes with DAVID [170] revealed significant enrichment of various immune-related functions, such as, hemopoeisis, positive regulation of lymphocyte activation, response to DNA damage stimulus and regulation of apoptosis (Table 2.1). Thus,

there appeared to be a significant activation of the immune system in infected birds of the S group, consistent with the observed response at the early cytolytic stage of the disease in susceptible birds.

Several genes having H3K4me3 DMRs were involved in the PANTHER [173] B-cell signalling pathway ($p = 1.3 \times 10^{-3}$) such as *LYN, SYK, GRB2, PTPRC, RAC2* and *BLNK*, indicative of increased B cell activation in the infected S group. The signalling molecules CD45, Lyn and Syk, gene products of *PTPRC, LYN* and *SYK*, respectively, are major players in the early stages of B cell antigen receptor signalling. These genes work together with *BLNK* and *GRB2* to activate B cells via the NF-κB mediated pathway while *BLNK* and *RAC2* may also activate B cells via the ERK, p38 or jun signalling cascades. H3K4me3 levels on all these genes were unchanged in the R group but were significantly higher in the infected S group after MDV infection (Figure 2.12). Three of these genes – *LYN*, *SYK* and *RAC2* – had reported expression in bursal cells [174] which suggests that the tissue-specific activation of these genes in the bursa might lead to increased B cell activation in susceptible birds.

MDV primarily targets B cells during early stages of the disease as these cells provide the first line of defence via the host humoral immune response. B cells surround the invading virus particles and have increased rates of infection and atrophy. The infection of B cells, in turn, induces the activation of CD4+ T cells which consequently become more vulnerable to virus infection [130]. The increase in B cell activation indicated by elevated levels of H3K4me3 on key genes involved in the pathway suggests the presence of an increased number of activated B cells in susceptible birds and a possible increase in the number of activated CD4+ T

lymphocytes. The larger population of cells vulnerable to infection by MDV at the early cytolytic stage of the disease in susceptible birds, could, therefore, result in increased levels of infection and higher mortality in the latter stages of the disease.

Discussion

The analysis of ChIP-Seq data poses several challenges including a diverse array of enrichment patterns, the lack of true biological controls and confounding factors such as sequencing depth, mappability and G/C content. In the presence of these sources of bias, it is important to have methods of analysis robust to various data characteristics that also preserve prediction accuracy. In response to these issues, we have developed a novel data-driven ChIP-Seq analysis algorithm named WaveSeq which is capable of detecting both punctate and diffuse enrichment regions and is free of distributional assumptions. WaveSeq utilizes non-parametric modeling of ChIP-Seq data using Monte Carlo sampling and a randomized algorithm to accurately estimate the empirical distribution of reads in the absence of a control.

With the aid of a variety of public data sets we were able to demonstrate that WaveSeq has high accuracy and performs favourably in comparison with several published methods of analysis in detecting punctate and diffuse enrichment regions (Figure 2.13). WaveSeq also performed with comparable accuracy in the absence of control data. Previous studies have observed that the background signal of ChIP-Seq data is non-random [77] and the ability to distinguish regions of true signal from background could be potentially improved if this non-randomness is accounted for. The improved detection capacity exhibited by WaveSeq in the absence of a control

data set suggests that the non-parametric modeling approach is successful in capturing the data characteristics leading to higher prediction accuracy.

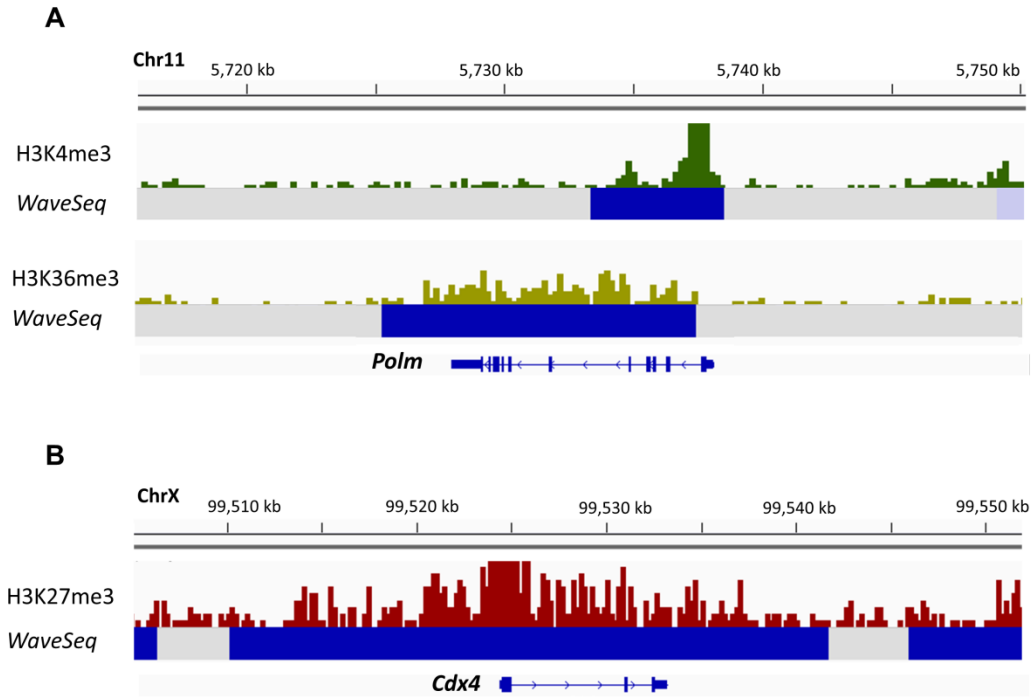


Figure 2.13 WaveSeq detects a broad variety of enrichment regions with high accuracy. Examples of WaveSeq peak calls on MEF histone modification data. (a) WaveSeq detects H3K4me3 and H3K36me3 marks on the housekeeping gene Polm located on chromosome 11 and (b) a broad peak of H3K27me3 on the developmental transcription factor Cdx4 which is silenced in differentiated cell populations.

The rapid advance of epigenetics and the advent of cost-effective next-generation sequencing technologies have led to complex experimental designs being employed to investigate various topics such as the epigenetics of disease response. WaveSeq is capable of being used in such an experimental setting and helps make relevant biological discoveries. We illustrate this by using our algorithm to analyze a complex H3K4me3 data set to investigate the differences in the epigenetic effects of MDV infection in inbred chicken lines having divergent responses to MD. WaveSeq detects the presence of H3K4me3 DMRs on key genes involved in the B cell activation

pathway suggesting the presence of increased numbers of activated B cells in infected individuals of the susceptible line. B cells are the primary targets of MDV at the early cytolytic stage of the disease and infection of these cells by the virus leads to activation of CD4+ T cells which are more vulnerable to infection than naive T cells. Consequently, an increase in the number of MDV-infected cells at this stage of the disease could translate to an increased viral load and a worse prognosis in susceptible birds at the latter stages of infection. Thus, epigenetic differences between the two lines could have a major impact on disease progression indicating that epigenetic marks play an important role in regulating disease response.

The absence of distributional assumptions in WaveSeq makes it potentially applicable to other forms of next-generation sequencing data. The detection of the genomic locations of nucleosomes is one such area of current interest. A nucleosome positioning experiment typically consists of the sequencing of DNA fragments associated with mono-nucleosomes across the whole genome. The data consists of broad diffuse regions with peaks that repeat approximately every 147 bp, the length of DNA associated with single nucleosomes. Regions of active transcription have lower nucleosome enrichment while high nucleosome density is associated with silent heterochromatin. Thus, differences in nucleosome density between samples could be predictive of transcriptional differences. Sequencing data having such underlying patterns could be highly suited to the wavelet transform framework employed by WaveSeq.

One of the primary drawbacks of WaveSeq is the relatively high number of peak calls for low SNR data such as H3K27me3, which is an unfortunate side-effect of the

sensitivity of the algorithm. However, since peak calls are ranked by FDR, a more stringent criterion can be used to circumvent this issue. Moreover, increased sequencing depth significantly improves discriminative power and is highly recommended particularly for data having diffuse enrichments.

Conclusions

ChIP-Seq experiments having a wide variety of enrichment patterns and a lack of true biological controls pose significant challenges for analysis and interpretation. WaveSeq is a highly sensitive, data-driven method capable of detecting significantly enriched regions in data having diverse characteristics. WaveSeq can detect both punctate and diffuse regions with a high degree of accuracy even in low SNR data sets. Moreover, it performs with comparable accuracy in the absence of control data. WaveSeq is suited for application in complex experimental scenarios, helping make biologically relevant functional discoveries and compares favourably with existing methods of analysis over a broad variety of data types.

3. Marek's Disease Virus Infection Induces Widespread Differential Chromatin Marks in Inbred Chicken Lines

Abstract

Marek's disease (MD) is a neoplastic disease in chickens caused by the MD virus (MDV). Successful vaccine development against MD has resulted in increased virulence of MDV and the understanding of genetic resistance to the disease is, therefore, crucial to long-term control strategies. Also, epigenetic factors are believed to be one of the major determinants of disease response.

Here, we carried out comprehensive analyses of the epigenetic landscape induced by MDV, utilizing genome-wide histone H3 lysine 4 and lysine 27 trimethylation maps from chicken lines with varying resistance to MD. Differential chromatin marks were observed on genes previously implicated in the disease such as *MX1* and *CTLA-4* and also on genes reported in other cancers including *IGF2BP1* and *GAL*. We detected bivalent domains on immune-related transcriptional regulators *BCL6, CITED2* and *EGR1*, which underwent dynamic changes in both lines as a result of MDV infection. In addition, putative roles for *GAL* in the mechanism of MD progression were revealed.

Our results confirm the presence of widespread epigenetic differences induced by MD in chicken lines with different levels of genetic resistance. A majority of observed epigenetic changes were indicative of increased levels of viral infection in the susceptible line symptomatic of lowered immunocompetence in these birds

caused by early cytolytic infection. The *GAL* system that has known anti-proliferative effects in other cancers is also revealed to be potentially involved in MD progression. Our study provides further insight into the mechanisms of MD progression while revealing a complex landscape of epigenetic regulatory mechanisms that varies depending on host factors.

Introduction

Rapid advances in epigenetics have led to the discovery of complex mechanisms of gene regulation involving phenomena such as DNA methylation and chromatin modifications. Methylation of particular histone residues has been found to correlate with specific and often opposing cellular functions, e.g. trimethylation of histone H3 lysine 4 (H3K4me3) is associated with transcriptional start sites (TSSs) of active genes while trimethylation of histone H3 lysine 27 (H3K27me3) is found to mark broad genomic regions for repression. Recent studies have also suggested that characteristic combinations of histone modifications or 'chromatin states' define functional elements of the genome and determine their contribution to transcriptional regulation [175-177]. Moreover, the epigenetic state of host genes can be affected by viral infection leading to tumors in humans [178-180]. Thus, epigenetics constitute a dynamic regulatory framework linking genotypes with environmental factors that could play a major role in differential disease responses among individuals having high genetic similarity.

Marek's disease (MD) is a highly contagious disease caused by an oncogenic α-herpesvirus MD virus (MDV) and characterized by T-cell lymphomas in chickens [123]. Major losses to the poultry industry as a result of MD have largely been

averted due to the success of various vaccination strategies which, remarkably, is also the first instance of the successful control of a natural cancer-causing agent using vaccines [121, 181, 182]. However, the virulence of the virus may have progressively increased as a consequence of vaccine development [183-185]. Several reported instances of vaccine breaks or reduced efficacy of vaccination, therefore, underlines the importance of investigating resistance to the disease as a long-term strategy to control MDV.

Natural resistance to MDV can be divided into two categories: major histocompatibility complex (MHC)-associated resistance, wherein different MHC haplotypes at the B blood group locus confer varying levels of resistance and non-MHC associated resistance in which birds having the same MHC haplotype exhibit vastly different responses to MDV infection. Inbred lines 6_3 and 7_2 developed at the Avian Disease and Oncology Laboratory (ADOL, East Lansing, MI) that we used in this study, fall into the latter category. These lines share a high degree of genetic similarity but have divergent responses to MDV infection completely independent of the MHC. Several studies have attempted to pinpoint factors responsible for conferring resistance [186-188], but confounding factors, such as, tissue types, virus strains and ages of birds have made it difficult to find a consensus. Multiple risk elements are possibly at play in this complex disease, and increased resistance or susceptibility is likely to be produced by a combination of such factors. In this study, we take a closer look at epigenetic factors behind different responses to MD with a view to a deeper understanding of the broader genomic impact of MDV infection.

We utilized the above population of inbred chickens – line 6_3 is highly resistant to MD, while line 7_2 is highly susceptible – and compared the epigenetic effects of MD. Genome-wide maps of H3K4me3 and H3K27me3 in thymus tissues of birds from these chicken lines at the latent stage of MDV infection were generated. We carried out systematic analyses to find differential chromatin marks induced by MDV infection. We also investigated co-localization patterns of the two chromatin modifications to detect putative bivalent domains and the effect of MDV on such domains. The results of our study confirm that Marek's disease has far-reaching effects on the epigenetic landscape of chicken lines with diverse responses to the virus and, thus, furthers our understanding of this complex disease.

Methods

Animals and Viruses

Two specific-pathogen-free inbred lines of White Leghorn either resistant (6_3) or susceptible (7_2) to MD were hatched, reared and maintained in Avian Disease and Oncology Laboratory (ADOL, Michigan, USDA). Four chickens from each line were injected intra-abdominally with a partially attenuated very virulent plus strain of MDV (648A passage 40) at 5 days after hatch with a viral dosage of 500 plaque-forming units (PFU). Infected and control chickens from both lines (n = 4) were terminated at 10dpi to collect thymus tissues. All procedures followed the standard animal ethics and use guidelines of ADOL.

Quantification of MDV loads in Thymus

The MDV gene *ICP4* was used for quantification of viral genomic DNA in thymus as previously described [189]. Quantitative PCR was performed by using 100 ng/µl of genomic DNA on the iCycler iQ PCR system (Bio-Rad, USA) and QuantiTect SYBR Green PCR Kit (Qiagen, USA) (Figure 3.1). The relative MDV loads were determined by normalizing to a single-copy gene *Vim* (vimentin) [190]. The primers for *Vim* are as follows: Forward: 5'-CAGCCACAGAGTAGGGTAGTC-3'; Reverse: 5'-GAATAGGGAAGAACAGGAAAT-3'.

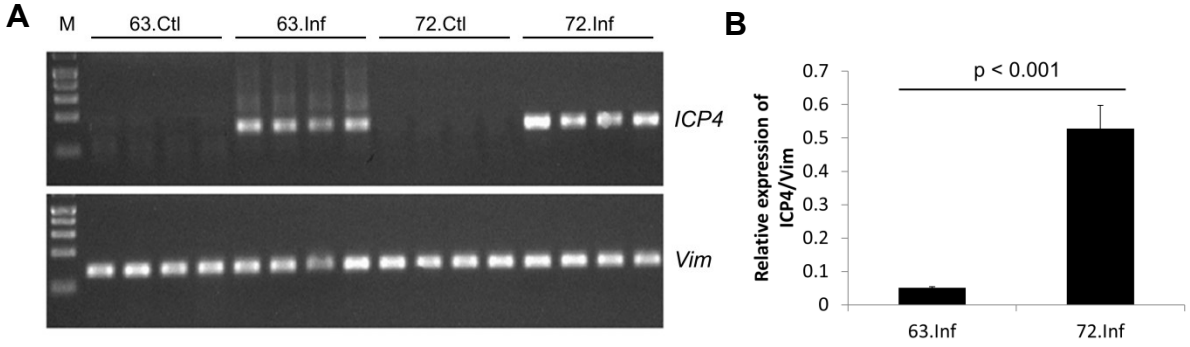


Figure 3.1. Quantification of viral loads in the MDV-challenge experiment using quantitative RT-PCR. The relative virus load is calculated by quantifying viral ICP4 normalized to the single-copy Vim gene (mean ± SEM, n = 4). (a) Only infected birds from the two lines exhibit measurable virus loads, with (b) the susceptible line 72 having a significantly higher number of virus particles ($p < 0.001$).

Chromatin Immunoprecipitation and Illumina Sequencing

Chromatin immunoprecipitation (ChIP) was carried out using thymus samples from MDV infected and controls birds [191]. About 30 mg thymus samples from three individuals were cut into small pieces (1 mm^3) and digested with MNase to obtain mononucleosomes. PNK and Klenow enzymes (NBE, Ipswich, MA, USA) were used to repair the ChIP DNA ends pulled down by the specific antibody. A 3′ adenine was then added using Taq polymerase and a pair of Solexa adaptors (Illumina, USA)

ligated to the repaired ends. Seventeen cycles of PCR was performed on ChIP DNA using the adaptor primers and fragments with a length of about 190 bp (mononucleosome + adaptors) were isolated from agarose gel. Subsequently, cluster generation and ChIP-sequencing (ChIP-Seq) using the purified DNA was performed on the Solexa 1G Genome Analyzer (Illumina, USA) following manufacturer protocols. The antibodies used and the total number of reads obtained for each sample is listed in Appendix IV.

Read Mapping and Summary Counts

Sequence reads obtained from the Illumina 1G Genome Analyzer were aligned to the May 2006 version of the chicken genome (galGal3) using Maq version 0.7.1 [78]. Default alignment policies of Maq were enforced: a valid alignment could have a maximum of two mismatches and if a read aligned equally well to multiple places in the genome, one was chosen at random. If multiple reads mapped to the same genomic location only one was kept to avoid amplification bias. Summary read counts were calculated using non-overlapping windows of 200 bp for visualization and normalized to per million mapped reads in each sample for the purpose of comparisons.

Identification of Significantly Enriched Regions (SERs)

Summarized read counts were subjected to peak calling with SICER [101]. The source code was modified to include support for the chicken genome. Fragment length was specified to be 190. A window size of 200 bp and gap size of 400 bp was used for the analysis. The E-value for estimating significant peaks was set to 100. For

the purposes of comparing different samples, SERs found in similar genomic regions of different samples were merged to obtain a consolidated list as follows: SERs from one sample were used to initialize the list. For each such region *M,* we searched for overlapping SERs in the next sample. In the case of an overlap between *M* and a significant region, *S*, the merged region was updated to include both *M* and *S*. This procedure was iterated over all samples to obtain a consolidated list of merged SERs.

Gene Annotation and Genomic Distribution of SERs

RefSeq and Ensembl gene annotations were downloaded from UCSC genome browser [167]. As there were only 4306 RefSeq genes in the database, we included Ensembl genes in our analysis to improve genome-wide coverage. There were 17858 annotated genes in the Ensembl database, which include validated and predicted genes. Redundancies between the databases were listed and accounted for, yielding a non-redundant list of 18198 genes with 4306 RefSeq genes and 13892 Ensembl genes. We then searched for overlaps between merged SERs and the non-redundant list of annotated genes. For H3K4me3, an SER was annotated with a gene if it overlapped the TSS region of the gene whereas in the case of H3K27me3, a valid overlap constituted an SER overlapping the gene body. To calculate the genomic distribution we counted all SERs having an overlap with one of the following regions: promoter (TSS ± 1 kb), exons, introns, 5' UTR and 3' UTR.

Histone Modification Profiles and Differential Chromatin Marks

Genes were divided into 10 sets based on their absolute expression and representative sets corresponding to high, medium, low and no expression were chosen for

visualization. We defined the gene body as the region between the transcription start site (TSS) and the transcription termination site (TTS). Histone modification profiles surrounding the gene body were calculated in 3 distinct regions: 5000 bp upstream of the 5' end, gene body and 5000 bp downstream of the 3' end of the gene. For reads falling within the gene body, read counts were obtained in bins 5% of the gene length while 1000 bp windows were used for the 5' and 3' flanking regions. The read counts in all cases were normalized to the total number of genes in the categories and total number of reads in the sample. We also compared gene expression to histone modification levels by plotting normalized microarray data (Zhang, H. unpublished data) against reads mapping to (i) TSS ± 500 bp and (ii) the gene body for each gene.

Reads mapping to merged SERs were tested for epigenetic changes induced by MDV infection in lines 6_3 and 7_2 using DESeq [119]. We used the method 'blind' for variance estimation and p-values were corrected for multiple testing using the Benjamini-Hochberg FDR procedure [172]. Statistical significance was defined using a false discovery rate (FDR) threshold of 0.4.

Validation of ChIP, ChIP-Seq and Gene Transcription by Q-PCR

Quantitative real-time RT-PCR was used to validate the quality of the ChIP and gene transcript levels on the iCycler iQ PCR system (Bio-Rad, Hercules, CA, USA). The real-time RT-PCR reactions were performed with a QuantiTect SYBR Green PCR Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. An online primer system (http://frodo.wi.mit.edu/primer3/) was used to design the

primers and four biological and four technical replicates were performed. The primer sequences are shown in Appendix V.

Data Access

Raw and processed ChIP-Seq data discussed in this manuscript were deposited in the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under Series accession number GSE33541.

Results

Genome-wide Distribution of H3K4me3 and H3K27me3

We performed ChIP-Seq experiments on infected and uninfected birds from lines 6_3 and 7_2 to investigate the epigenetic effects of MDV infection. More than 76 million reads from eight samples were mapped to the chicken genome yielding 14418 and 24950 significantly enriched regions (SERs) for H3K4me3 and H3K27me3, respectively (Table 3.1). We further classified these regions as follows: Ubiquitous SERs were found in all samples and were likely due to similarities in the genetic background of the chickens. Line-specific SERs were present in only one line either before or after MDV infection, while condition-specific SERs appeared in both lines but only in individuals with the same infection status.

Ubiquitous SERs formed the largest percentage of all enriched regions, accounting for 74.2% and 23.3% in H3K4me3 and H3K27me3 samples, respectively. In the case of H3K4me3, there were large differences in the number of specific SERs - more than 2000 line-specific SERs were found in line 63, compared to about 300 in line 72.

Similarly, we found 50% more line-specific SERs of H3K27me3 in line 63 (6568) compared to line 72 (4494). However, upon closer examination, most of the line-specific and condition-specific SERs were revealed to have low read counts (Appendix V1) corresponding to regions of low enrichment.

Table 3.1. Significantly enriched regions (SERs) and associated genes in each sample.

		H3K4me3		***H3K27me3***	
	Samples	**SERs (%)**	**Genes**	**SERs (%)**	**Genes**
Line-Specific	**63I**	647 (4.5)	78	3477 (13.9)	615
	63N	594 (4.1)	71	2514 (10.1)	896
	63I,63N	924 (6.4)	190	577 (2.3)	150
	72I	105 (0.7)	16	1658 (6.6)	451
	72N	126 (0.9)	11	2506 (10)	346
	72I,72N	73 (0.5)	17	330 (1.3)	89
Condition-specific	**63I,72I**	97 (0.7)	35	2061 (8.3)	579
	63N,72N	47 (0.3)	9	66 (0.3)	22
Ubiquitous	**63I,63N,72I,72N**	10691 (74.2)	9475	5831 (23.4)	2942
	Total	**14418**	**10206**	**24950**	**7904**

63I: line 6_3 infected, 63N: line 6_3 control, 72I: line 7_2 infected, 72N: line 7_2 control.

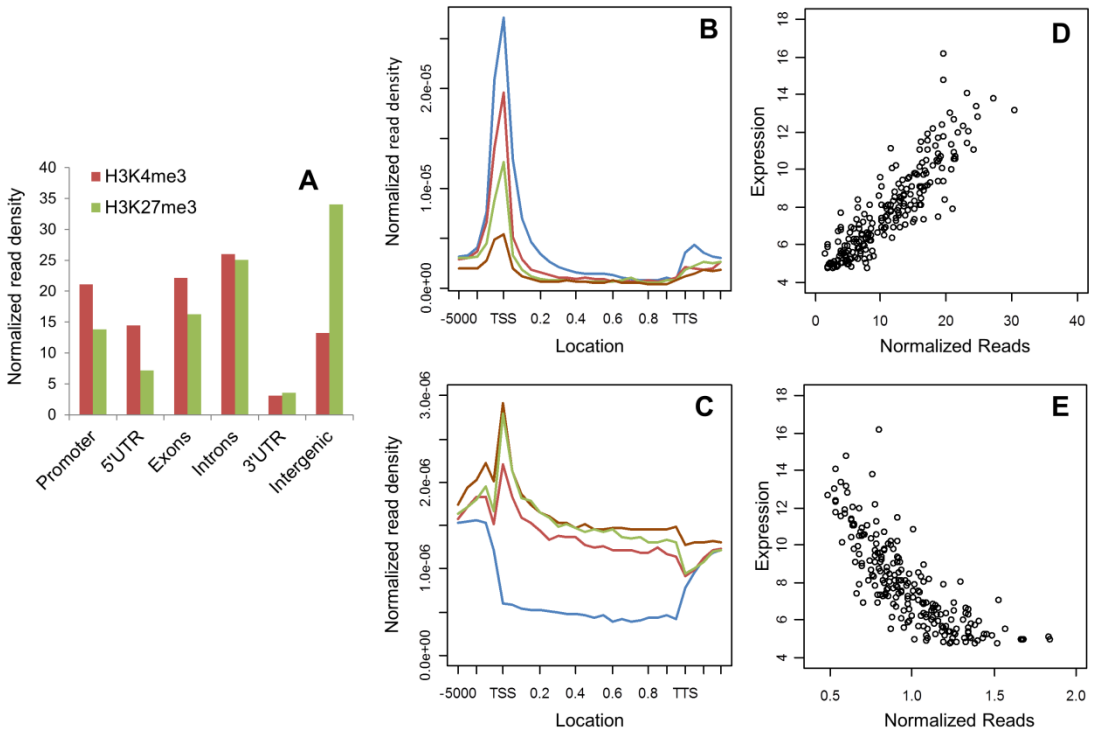


Figure 3.2. Genomic distribution of SERs and relationship between histone marks and gene expression.
(a) Distribution of SERs over different genomic elements. (b-e) Relationship between gene expression and histone marks in infected line 6_3 birds. Plots of histone modifications around the gene body (b, c) in genes having high (blue), medium (red), low (green) and no activity (brown). (d,e) A comparison of epigenetic marks and transcriptional levels. Similar trends were observed in other experimental groups (Appendices VIII-X).

Genes were divided into five regions – promoter, 5' untranslated region (UTR), exons, introns and 3' UTR – and the distribution of SERs across these elements was probed (Figure 3.2 A). We found a large number of intergenic regions marked by H3K27me3, consistent with high levels of this mark associated with areas of silent heterochromatin. In the case of H3K4me3, a larger proportion of SERs were found around the promoter, exons and the 5' UTR, while similar proportions of H3K4me3 and H3K27me3 SERs were present in introns and 3' UTRs. A comparison of the genomic distributions of SERs in the different samples (Figures 3.3 A, B) showed a

similar number of H3K4me3 SERs across the promoter, exons and the 5' and 3' UTRs of genes. Line 6_3 contained a higher number of intronic and intergenic SERs as compared to line 7_2 although this did not appear to change as a result of MDV infection. On the other hand, a greater number of H3K27me3 SERs were found in the infected samples although these levels were similar in the two different lines.

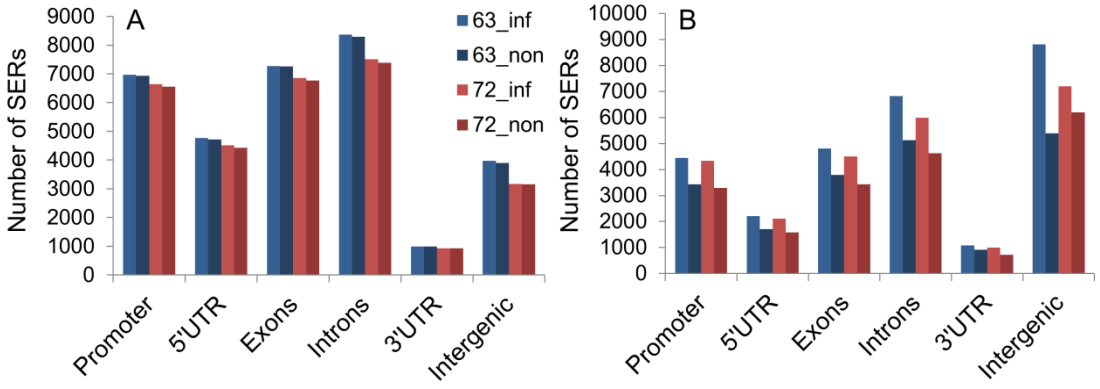


Figure 3.3. Distribution of SERs over different genomic elements.
(a) H3K4me3 and (b) H3K27me3 SERs separated by samples. 63_inf: line 6_3 infected, 63_non: line 6_3 control, 72_inf: line 7_2 infected, 72_non: line 7_2 control.

To analyze the relationship between histone modifications and gene expression, histone modification profiles surrounding the TSS and gene body were plotted for genes ranked by their expression level (Figures 3.2 B-E and appendices VII-X). As expected, a strong positive correlation was observed between gene expression and H3K4me3 marks with a distinct peak around the TSS and little to no enrichment within the gene body. On the other hand, H3K27me3 showed negative correlation with gene expression with a peak near the TSS followed by a broad plateau across the gene body. However, the latter relationship was non-linear – genes with lower expression had similar levels of H3K27me3 marks and levels were markedly distinct only at higher expression levels (Figure 3.2 C, E).

Table 3.2 Differential SERs identified in thymus

	H3K4me3		*H3K27me3*	
Comparison	**Differential SERs***	**Genes**	**Differential SERs***	**Genes**
63I vs 63N	9	4	42	1
72I vs 72N	30	13	5	0
63N vs 72N	148	46	1094	65
Total	**179**	**59†**	**1116**	**66**

*FDR < 0.4. † Some genes are shared between different comparisons. 63I: line 6_3 infected, 63N: line 6_3 control, 72I: line 7_2 infected, 72N: line 7_2 control.

Differential H3K4me3 Marks on Genes Related to MD

We conducted a comprehensive analysis of genome-wide chromatin marks to find significant differences in MDV-induced responses in line 6_3 and 7_2. We used two sets of comparisons: First, to assess the influence of MDV infection within each line, we compared the infected and the non-infected samples from the same line. Secondly, the non-infected samples from the two lines were compared to detect line-specific effects. As a result of this analysis we found 179 differential H3K4me3 SERs and 1116 differential H3K27me3 SERs that mapped to 59 and 66 genes, respectively (Table 3.2). A majority of differential SERs were found in the comparison between non-infected samples of the two lines (Appendix X, XI) with several observed on genes that have been associated with MDV infection.

MX1 is a zinc-finger transcription factor that has antiviral properties against a large number of viruses. *MX1* was upregulated after MDV infection [192] although its contribution to MD progression is unknown. MDV infection induced a highly significant increase in H3K4me3 enrichment in the promoter region of *MX1* in both lines (line 6_3: $p = 1.28 \times 10^{-7}$, line 7_2: $p = 4.26 \times 10^{-9}$; Figure 3.4 A). We observed a concurrent increase in transcript levels after MDV infection in line 7_2 ($p = 0.0264$;

Figure 3.4 B); *MX1* expression in line 6_3 showed a similar trend (fold change = 38.22, p=0.085) although mRNA levels were much lower.

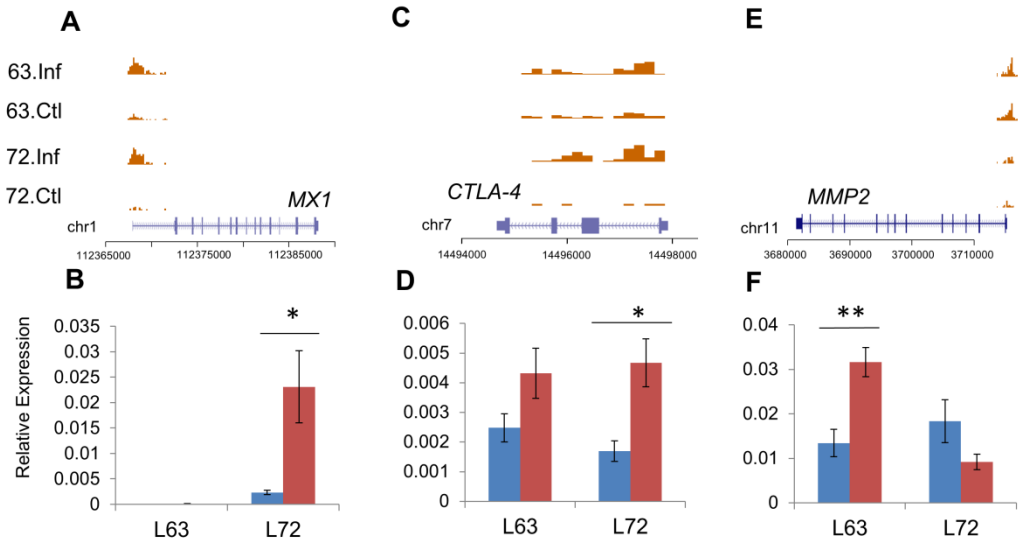


Figure 3.4. Genes related to MD show differential H3K4me3 marks.
MX1 (a, b) and *CTLA-4* (c, d) show increased H3K4me3 marks and higher expression in infected individuals from both lines; *MMP2* (e, f) exhibits higher levels of H3K4me3 in susceptible line 7_2. n = 4; * = significant at $p < 0.05$; ** = significant at $p < 0.01$; *** = significant at $p < 0.001$.

CTLA-4 is a cell surface glycoprotein expressed on CD4+ and CD8+ T lymphocytes that is a powerful negative regulator of T-cell activation [193]. The CTLA4 protein is expressed on T lymphocytes soon after activation and regulates T-cell proliferation, production of IL-2 and also supports the function of T_{reg} cells that suppress immune response [194]. Previous studies have reported an increase in *CTLA-4* expression after MDV infection [195]. We found an increase in H3K4me3 enrichment in line 7_2 as a result of MDV infection ($p = 0.0003$) and there was a similar trend in line 6_3 (Figure 3.4 C). Consistent with the above, there was a significant increase in transcript levels after MDV infection in line 7_2 ($p = 0.04$) and a small increase in line 6_3 (Figure 3.4 D).

MMP2 plays a key role in the degradation of the extra-cellular matrix, and an increase in expression has been associated with increasing tumor cell migration and tumor angiogenesis [196, 197]. *MMP2* was upregulated during the neoplastic stage of MD infection in susceptible birds [198] but downregulated in response to MDV infection during early lytic infection in susceptible and resistant chickens [192]. We observed a slight increase in H3K4me3 enrichment after MDV infection in both lines, while line 7_2 exhibited significantly lower levels than line 6_3 ($p = 0.0016$; Figure 3.4 E). This was coupled with increased *MMP2* expression in line 6_3 after infection ($p = 0.0068$) while there was no such change in line 7_2 (Figure 3.4 F).

Genes Related to Cancers Show Epigenetic Changes in Response to MD

We observed differential histone marks on several genes that have been associated with other cancers but not in the context of MDV infection. Insulin-like growth factor 2 binding protein 1 (*IGF2BP1*) is an RNA-binding factor that regulates the translation of mRNAs produced by certain genes like *IGF2* and *ACTB*. Increased expression of *IGF2BP1* has been implicated in the development and progression of cancers of various organs, e.g. lung, brain, breast and colon [199-202]. There was no change in the H3K4me3 enrichment levels induced by MDV infection although a significantly higher level of enrichment was present in line 7_2 ($p = 4.21 \times 10^{-13}$; Figure 3.5A). Transcript levels in line 7_2 were much higher than in line 6_3, but reduced in response to MDV infection ($p = 0.044$) (Figure 3.5 B).

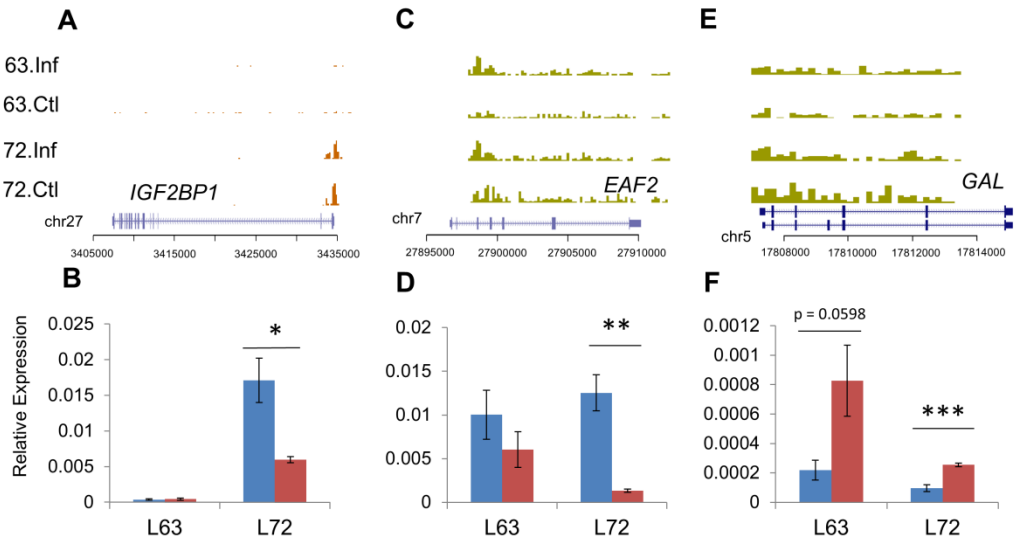


Figure 3.5. MD induces epigenetic changes in genes related to various cancers. *IGF2BP1* (a, b) shows differential H3K4me3 marks and increased expression in susceptible birds while *EAF2* (c, d) and *GAL* (e, f) have differential H3K27me3 levels on the gene body. n = 4; * = significant at $p < 0.05$; ** = significant at $p < 0.01$; *** = significant at $p < 0.001$.

ELL associated factor 2 (*EAF2*) is a testosterone regulated apoptosis inducer and tumor suppressor. Inactivation of *EAF2* has been shown to lead to tumors in multiple organs [167]. There was a significant increase in H3K27me3 levels after MDV infection in line 6_3 ($p = 0.0414$) while among uninfected chickens these levels were markedly higher in line 7_2 ($p = 0.0138$; Figure 3.5 C). However, *EAF2* expression was drastically reduced after MDV infection in line 7_2 ($p=0.0016$) but showed only a small decrease in line 6_3 (Figure 3.5 D).

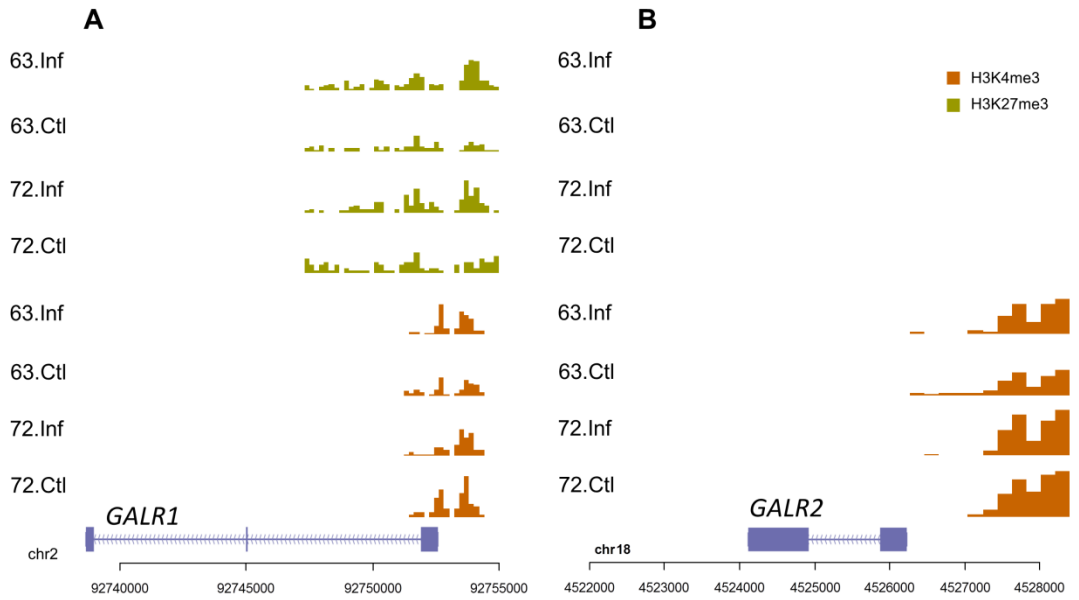


Figure 3.6. Significant H3K4me3 and H3K27me3 enrichment around *GALR1* and *GALR2*.
(a) The anti-proliferative *GAL* receptor *GALR1* exhibited both active and repressive marks. There is no change in H3K4me3 levels but a definite increase in H3K27me3 levels after infection in line 7_2. (b) No significant histone marks observed on *GALR2*.

Galanin (*GAL*) is a neuropeptide that modulates various physiological functions, such as, inhibition of insulin secretion and stimulation of growth hormone secretion. Three galanin receptors are known (*GALR1*, *2* and *3*): the expression of *GALR1* has anti-proliferative effects while *GALR2* can be both anti- or pro-proliferative in function. Therefore, the *GAL* system is considered to be a promising candidate for detection and treatment of various cancers [203, 204]. We observed an increase in H3K27me3 levels on *GAL* after infection in both lines (Figure 3.5 E). Also, expression levels were significantly lowered after MDV infection in line 7_2 ($p = 0.00087$) while there was a similar trend in line 6_3 ($p = 0.051$; Figure 3.5 F). Interestingly, *GALR1* had both H3K4me3 and H3K27me3 enrichments (Figure 3.6 A) although *GALR2* showed no significant histone marks (Figure 3.6 B).

Chromatin Co-localization Patterns Reveal Putative Bivalent Genes

Regions of chromatin having both active and repressive marks are said to be bivalent and have been shown to play important roles in development and genetic imprinting [20, 205]. For example, bivalent domains have been shown to mark promoters of genes that are subsequently silenced in tumors by DNA hypermethylation indicating their importance in cancer [206]. A mono-allelic bivalent chromatin domain that controls tissue-specific genomic imprinting at a specific locus was recently found in mice [205]. To investigate the presence of such bivalent chromatin states and the possible effect of MDV infection, we defined bivalent genes as those having H3K4me3 reads (TSS ± 500 bp) greater than 30 reads per million mapped reads (RPM) and H3K27me3 reads (gene body) greater than 2 RPM, respectively (~85th percentile). This filtering process yielded a list of 99 putative bivalent genes (Appendix XII).

Functional annotation clustering of the above genes using DAVID [170, 207] revealed significant enrichment of several immune-related functions. These included transcription factor *EGR1* which is reported to have tumor suppressor properties, genes involved in lymphocyte activation and differentiation such as *BCL6*, *CD4* and *SMAD3* and genes *TLR3* and *TIRAP* that are part of the toll-like receptor signaling pathway. Bivalent domains were also present on a variety of transcription factors with immune-related functions such as *CITED2*, a transactivator that regulates NF-κB, *MYC* a transcription factor associated with hematopoetic tumors and *RHOB* a Ras family homolog that mediates apoptosis in tumor cells after DNA damage. Moreover,

all genes involved in the top five functional annotation clusters showed higher chromatin levels in line 7_2 primarily after MDV infection (Appendix XIII).

Bivalent Domains are Altered in Response to MD

We further investigated the effect of MD on bivalent chromatin domains observed on *BCL6*, *CITED2*, *EGR1*, *CD4* and *TLR3* (Figures 3.7 and 3.8). Interestingly, three of these genes, *CITED2, BCL6* and *EGR1*, showed identical epigenetic and transcriptional signatures.

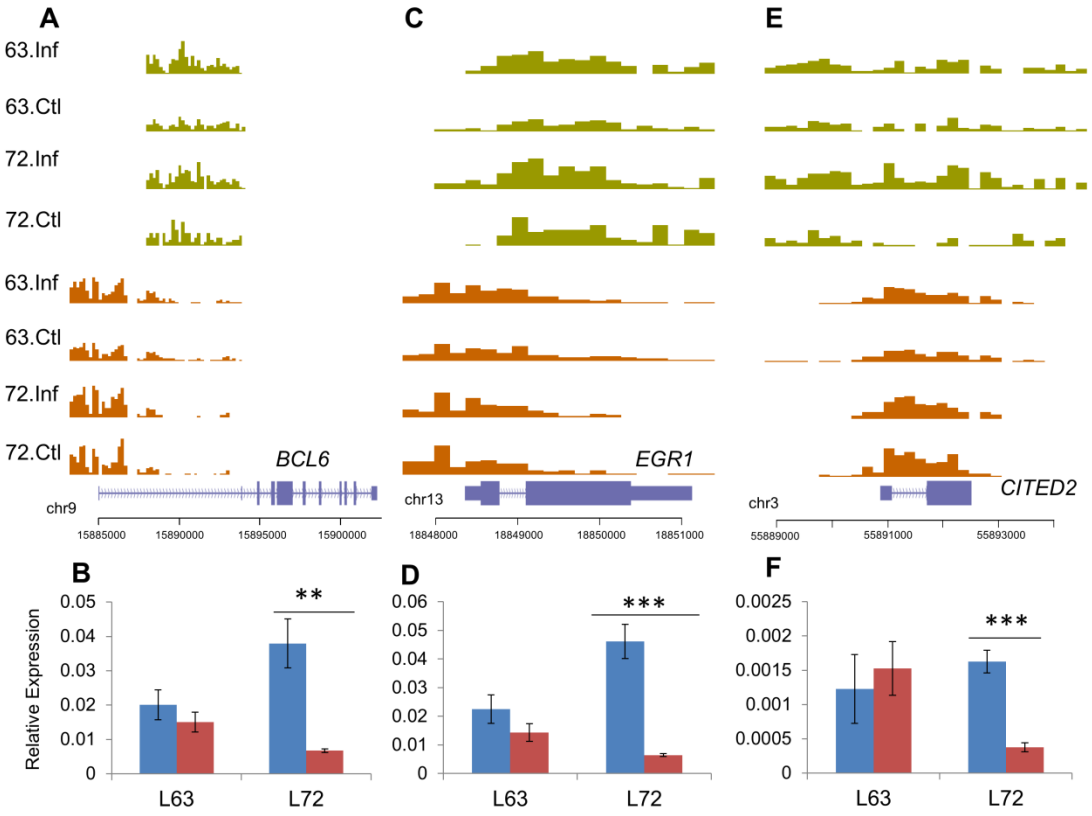


Figure 3.7. Bivalent domains on transcriptional regulators are altered by MD. H3K4me3 (orange) and H3K27me3 (green) profiles and associated transcript levels of *BCL6* (a, b), *EGR1* (c, d) and *CITED2* (e, f). In all three cases we observe a slight increase in H3K27me3 induced by MDV infection in line 7_2 and a concurrent significant decrease in transcript levels while increase in active and repressive marks appear to cancel each other out in line 6_3.

CITED2 is a member of the p300/CBP co-activator family that has intrinsic histone acetyltransferase activity and plays a major role in regulating and coordinating multiple complex cellular signals to determine the expression level of a gene [208]. B-cell CLL/lymphoma 6 (*BCL6*) is a zinc finger protein that functions as a transcriptional repressor which was downregulated at 15 dpi in spleen tissues from F1 progeny ($15I_5$ X 7_1) of MD-susceptible chickens [195]. *EGR1* belongs to a group of early response genes induced by a variety of signaling molecules such as growth factors, hormones and neurotransmitters that is believed to play a major role in cell proliferation and apoptosis [209]. Overexpression of *EGR1* promotes tumor growth in kidney cells [210] but suppresses growth and transformation in other cell types, e.g. fibroblasts and glioblastoma cells [211].

In each of the above genes, both active and repressive chromatin marks were increased in response to infection in line 6_3 chickens. However, in line 7_2, there was a definite increase in H3K27me3 marks but no change in H3K4me3 (Figures 3.7 A, C, E). Transcript levels were in agreement with this observation: infected line 7_2 chickens showed a significant downregulation (*CITED2*: p=0.0004; *BCL6*: p=0.0048; *EGR1*: p=0.0005; Figures 3.7 B, D, F), while there were no such changes in line 6_3.

On the other hand, *TLR3* and *CD4* showed a slight increase in H3K4me3 marks after MDV infection while there were no appreciable changes in H3K27me3 levels. In keeping with the epigenetic changes, there was a small increase in gene expression in infected birds from both lines (Figure 3.8).

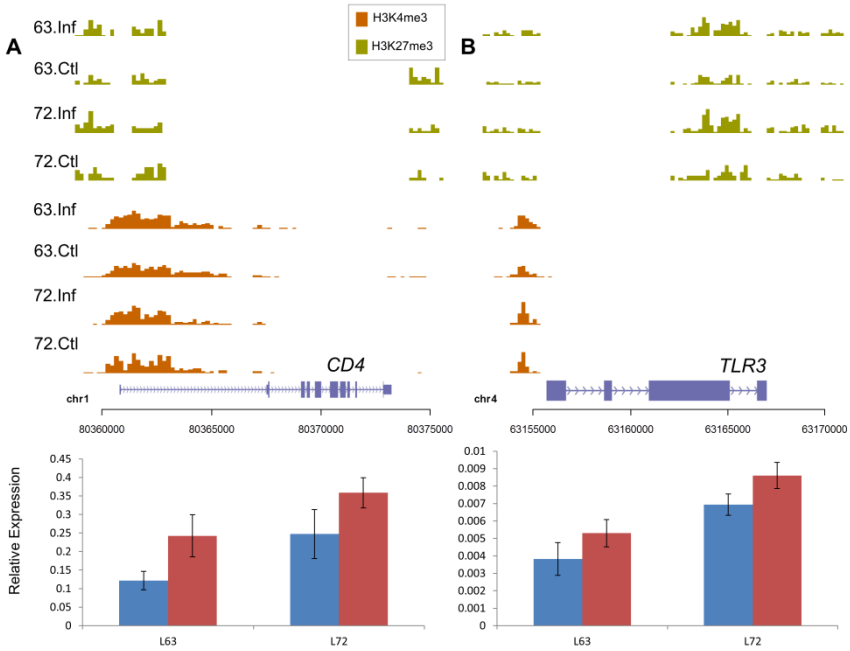


Figure 3.8. Bivalent domains on some genes are unaffected by virus infection. MDV infection has no effect on the bivalent domains or transcription levels of CD4 and TLR3.

Discussion

Immune parameters that are known to play a major role in genetic resistance to MDV are correlated with innate immune responses, such as NK cell activity, production of nitric oxide and cytokines, such as, interferons. Recent studies have identified host cytokines such as IL-18 and IFN-γ that contribute to the initiation and continuation of latency [212]. However, cytokine levels can undergo rapid flux in response to infection, and consistent with this, we did not observe any epigenetic changes associated with these genes in the MHC-congenic lines used in our study (Figure 3.9). This suggests the existence of other extrinsic factors responsible for transcriptional variations between resistant and susceptible chickens at this stage of the disease.

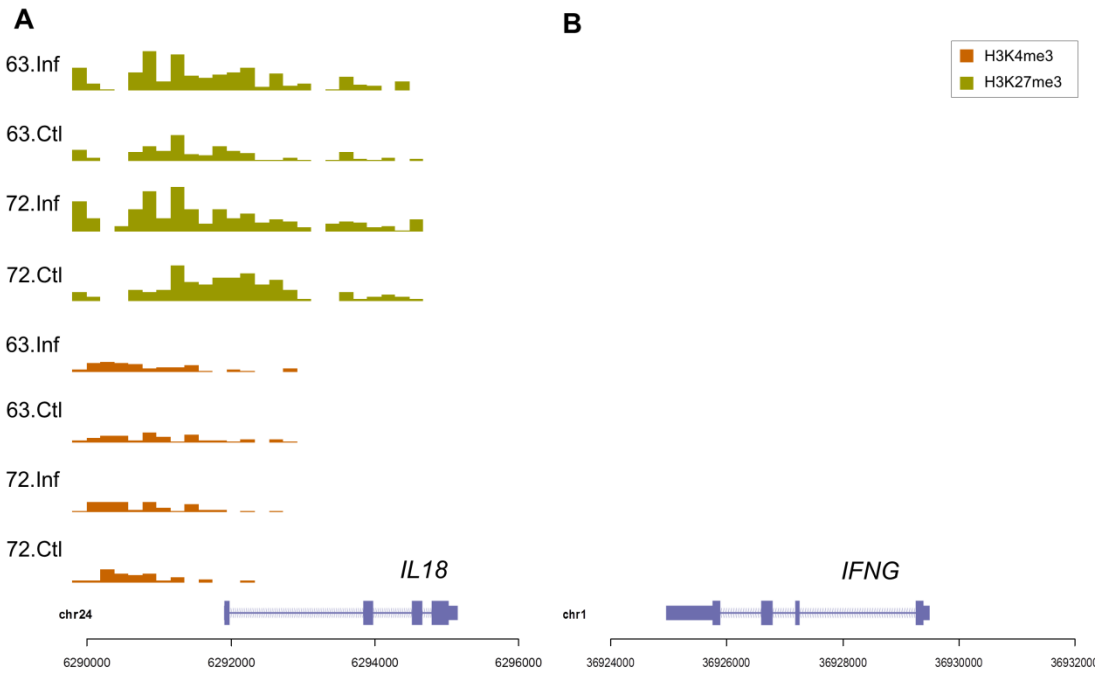


Figure 3.9. Epigenetic profiles of host cytokines (a) IL-18 and (b) IFN-γ.
IL-18 does not show any notable changes in response to MDV infection while IFN-γ does not show any SER.

A global comparison of histone modification levels in the two inbred chicken lines yielded some interesting results. As expected, a majority of SERs were found in all the experimental groups, indicating a high level of epigenetic similarity between the lines in addition to inherent genetic similarity. In the case of H3K27me3, the percentage of ubiquitous SERs was relatively low (23.4%), although this was likely due to lower precision of the peak caller for broad chromatin marks. Besides, most of the SERs detected in a subset of samples corresponded to regions of low enrichment, which may also be the reason behind the relatively low number of differential SERs detected in our study.

Genes that have been previously associated with MD and various other cancers showed differential marks that are either MD-induced (*MX1, CTLA-4, EAF2* and *GAL*) or line-specific (*IGF2BP1* and *MMP2*). The increase in H3K4me3 marks

around the TSS of *MX1,* a gene with known antiviral properties, appeared to be correlated with upregulated expression in both lines in response to MDV infection. However, lowered overall mRNA levels in the resistant line suggest additional factors could be involved in the regulation of this gene. Similarly, increased mRNA levels of the lymphocyte surface marker *CTLA4* is possibly due to the presence of larger numbers of T cells in MDV infected birds as a result of higher levels of infection. *EAF2* functions as an apoptosis inducer in addition to being a tumor suppressor, and therefore, its downregulation could contribute to higher tumor incidence rates in line 7_2. However, it is not clear why a significant increase in H3K27me3 levels did not have any effect on transcript levels in the resistant line.

IGF2BP1 is believed to act by stabilizing the mRNA of the *c-myc* oncogene and therefore, the higher expression of this gene and a more stable c-myc gene product might play a role in increasing MD susceptibility in line 7_2 birds via increased cell proliferation and transformation. The matrix metalloprotease *MMP2* is upregulated after infection in the resistant line 6_3, similar to the previously observed increase at the neoplastic stage of MD. However, mRNA levels were similar in the two lines before infection contrary to the difference in H3K4me3 levels suggesting that additional factors are responsible for regulating this gene.

The correlation between observed differential histone marks and transcript levels was moderate at best. Indeed, differential H3K4me3 marks were strongly predictive of gene expression levels but the correlation between H3K27me3 and mRNA levels was relatively poor. There could be various reasons for this – indeed, H3K27me3 levels had a non-linear relationship with gene expression with higher marks showing little

difference in the effect on expression. Therefore, in this tissue, the levels of H3K27me3 may not be a very good indicator of gene expression levels. Also, the transcription of these genes might be controlled by other factors with the change in H3K27me3 levels only incidental.

Bivalent domains were detected on transcriptional regulators *BCL6*, *CITED2* and *EGR1* and the galanin receptor *GALR1*. The epigenetic and transcriptional signatures observed on these genes indicated that they were poised at the latent stage of the disease, but with crucial differences in the two lines. Increased repressive marks in the susceptible line correlated with significant downregulation of the genes, while in line 6_3, the increase in both marks appeared to compensate for each other with no eventual effect on gene transcription. This suggests that some 'poised' bivalent genes can become highly repressed even with a relatively small increase in H3K27me3 marks. The change in the chromatin levels could also be correlated with an increase in cell populations having the repressive mark. Taken together, the above findings point towards the existence of finely balanced epigenetic control of transcription, which may be necessary to mount a rapid response by the immune system. However, this machinery could potentially be hijacked by a pathogen and result in an aberrant phenotype. The effect of MDV infection on the bivalent domain on *GALR1*, in particular, suggests the repression and potential loss of its anti-proliferative effects. Thus, the galanin system possibly plays an important and hitherto unknown role in MD progression and could be a novel target for long-term control of the disease.

One of the major players in MDV-induced malignant transformation is Meq, a virus-encoded oncoprotein that has diverse functions, e.g. transactivation, chromatin

remodeling and regulation of transcription. Meq interacts with and sequesters the tumor suppressor protein p53, resulting in the dysregulation of cell-cycle control [123] and inhibition of the transcriptional and apoptotic activities of the protein [213]. Several genes that show epigenetic changes in response to MDV infection have been associated with p53. Downregulation of *CITED2* stabilizes the p53 protein leading to its accumulation [214]. The *BCL6* gene product is believed to contribute to lymphomagenesis by inactivation of p53 [215]. Besides, *EAF2* has also been shown to interact with and suppress the function of p53 [167]. The downregulation of all of the above genes in susceptible birds after MDV infection points towards the increased production of p53 and a robust anti-tumor response. That we still observe higher tumor incidence rates in this line, suggests the presence of large amounts of inactivating viral Meq protein which, in turn, indicates that increased numbers of MD-infected cells are present in the susceptible line at this stage of the disease. A majority of tumors are believed to result from the viral transformation of CD4+ T cells some of which are infected at the latent stage of MD [216]. The larger number of virus-infected cells produced in the susceptible line is possibly due to lowered immunocompetence as a result of the early stages of infection. Thus, a more detailed investigation of the early cytolytic stage of MD is necessary to shed further light on the causes behind the divergent response to MD observed in these birds.

Whole tissues represent a mixture of various cell populations, and observed epigenetic changes might be due to a change in chromatin marks in a particular cell type or a variation in the relative number of cells carrying active or repressive histone marks. However, such *in vivo* studies are representative of the host response at a

systems level wherein different cell types might interact in a cooperative manner to fight infection. Thus, while the study of pure cell populations is likely to yield greater discriminative power, the investigation of tissue macroenvironments is, perhaps, closer to reality.

This study focused on the thymus tissue as it is a major immune organ and contains a significant population of T lymphocytes in various stages of differentiation. Our earlier study of the MDV-induced transcriptome in these birds indicated the presence of line-specific differences at the latent stage of MD [217]. In addition, birds susceptible to MD suffer thymic atrophy during the early stages of infection [218], indicating the importance of understanding changes in this organ to elucidate the mechanisms involved in disease progression. Ongoing studies in our lab include other tissues, e.g. spleen [219], and a time-course through the various stages of infection, to further investigate the systemic effects of MD and the epigenetic basis of MD resistance.

Conclusions

We studied the effect of latent MDV infection on two chromatin modifications in inbred chicken lines exhibiting different degrees of resistance to MD. Several genes showed changes in histone enrichment and this response was often significantly different between the two chicken lines. A detailed analysis of co-localization patterns of the chromatin marks revealed the presence of bivalent domains on a number of immune-related transcriptional regulators. More importantly, these domains showed marked changes in response to MDV infection and provide further evidence of the far-reaching epigenetic effects of MD. Our results suggest putative

roles for the *GAL* system in MD progression. A majority of the differential chromatin marks are also suggestive of increased levels of viral infection in the susceptible line symptomatic of lowered immunocompetence in these birds at early stages of the disease. In summary, our study provides further insight into the mechanisms of MD progression while revealing a complex landscape of epigenetic regulatory mechanisms. Further work is necessary to fully elucidate the underlying mechanisms of MD, but our results suggest that this is a promising step towards a deeper understanding of this complex disease.

4. Temporal Chromatin Signatures Induced by Marek's Disease Virus Infection in Bursa of Fabricius

Abstract

Marek's disease (MD) is a highly contagious, lymphomatous disease of chickens induced by a herpesvirus, Marek's disease virus (MDV) that causes major annual losses to the poultry industry. Similar to other herpesviral infections, MD pathogenesis involves multiple stages including early cytolytic and latency, and transitions between these stages are governed by several host and environmental factors. The success of vaccination strategies has led to increased virulence of MDV and selective breeding of naturally resistant chickens is seen as a viable alternative. While multiple gene expression studies have been performed in resistant and susceptible populations little is known about the epigenetic effects of infection. Thus, in this study, we investigated temporal chromatin signatures induced by MDV by analyzing early cytolytic and latent phases of infection in the bursa of Fabricius of MD-resistant and –susceptible birds. Several pathways that have been previously reported in connection with MD, including apoptosis, p53 signaling and cytokine cytokine receptor-interaction, displayed changes in histone modification marks. In addition, several novel pathways were enriched. The neuroactive ligand receptor-interaction pathway showed marked reductions in H3K4me3 marks, particularly in MD-resistant chickens and several genes belonging to the spliceosome pathway showed increased H3K4me3 marks in resistant birds at the latent stage of infection.

Variations in chromatin marks suggest greater inflammation in susceptible chickens at the early cytolytic stage of infection, while the resistant line exhibited recuperative symptoms. During latent MD, the resistant line showed widespread reduction in H3K4me3 marks suggesting epigenetic silencing. Our observations regarding chromatin profiles were also largely in agreement with previous reports. The temporal analysis of chromatin signatures, therefore, revealed further clues about the epigenetic effects of MDV infection. Further studies are necessary to understand the functional implications of the observed variations in histone modifications.

Introduction

Marek's disease (MD) is a highly infectious disease caused by an α-herpesvirus, Marek's disease virus (MDV), that affects chickens worldwide. MD pathogenesis can be divided into three major stages: an early cytolytic phase, which occurs between 3 and 6 days post infection (dpi), is characterized by the infection of B lymphocytes, the first major targets of MDV. The infected B cells enter circulation and induce the activation of CD4+ T cells which in turn become infected. In subsequent stages of the disease, CD4+ T cells form the primary vehicle for MDV multiplication and dissemination, along with a smaller percentage of other cells including B and CD8+ T lymphocytes. Around 7 dpi, the infection enters a latent phase defined by the absence of expressed viral antigens and virus production. This switch to latency is believed to be governed by many viral and host factors, such as, viral interleukin (vIL)-8, which acts as a chemoattractant for T lymphocytes [220], and upregulated chicken major histocompatibility complex (MHC) class II molecules on infected cells promoting the initiation of host immune response [221]. In MD-resistant chickens, latent infection

persists at a low level in lymphoid tissues and CD4+ T lymphocytes. However, in MD-susceptible genotypes, a second cytolytic phase occurs 2-3 weeks after the primary infection, wherein latently infected lymphocytes are transformed and proliferate rapidly to form tumors in various tissues.

The primary lymphoid organs of spleen, thymus and the bursa of Fabricius, are important focal points of MD progression. Cytolytic infection initiates in the spleen before spreading to other lymphoid organs, which lag behind by a day. This is accompanied by significant cytolysis of B and T lymphocytes in addition to varying levels of inflammatory response. Bursal follicles and the thymic cortex undergo regressive changes in this stage of MD leading to organ weight loss, while there is massive apoptosis of thymocytes. In the spleen, however, inflammation results in an increase in organ weight. The above changes are reduced within two weeks of infection, with the organs almost returning to their normal form and structure. However, in MD-susceptible chickens, a second wave of cytolytic infection around 14-21 dpi results in marked inflammation, severe atrophy of bursa and thymus and permanent immunosuppression.

There have been several studies of the effect of MD, particularly in the spleen, but relatively few concerning the bursa of Fabricius [222, 223]. The latter is a primary lymphoid organ evolutionarily unique to birds and critical to the development of the B cell lineage [224]. B lymphocytes in all the major lymphoid organs, as mentioned above, are the primary targets of the virus in the early stages of the disease [129]. Embryonal bursectomy resulted in the abolition of early lytic infection along with reduced viremia and tumor formation, in spite of comparable MD incidence [223].

Bursal atrophy was observed in MD-susceptible line $L7_2$ chickens with the effect reduced in the MD-resistant line $L6_3$ individuals [129], while the bursa-dependent immune system was impaired in infected chickens [225]. It is, therefore, evident that the bursa of Fabricius plays an important role in MD pathogenesis, and it is vitally important to understand the effect of MDV on this organ.

In this study, we used chromatin immunoprecipitation followed by sequencing (ChIP-Seq) to analyze temporal chromatin marks induced by MDV infection. For this work, we utilized a population of inbred chicken lines having contrasting responses to the disease and focused on the bursa of Fabricius. In doing so, we extended our previous studies [219, 226] to include both the cytolytic and latent phases of MD. Our primary goal was to investigate the dynamic changes of chromatin induced by MDV infection to uncover the biological pathways that could be affected by variations in histone modification enrichments. The biological consequences of chromatin profiles are context-specific and similar patterns can lead to a variety of outcomes [227]. Therefore, we focused on *changes* of chromatin enrichments as evidence of possible epigenetic regulation.

Materials and Methods

Animals and viruses

Two specific-pathogen-free inbred lines of White Leghorn, either resistant ($L6_3$) or susceptible ($L7_2$) to MD, were hatched, reared and maintained in Avian Disease and Oncology Laboratory (ADOL, Michigan, USDA). Eight chickens from each line were injected intra-abdominally with a partially attenuated very virulent plus strain of

MDV (648A passage 40) at 14 days after hatch with a viral dosage of 500 plaque-forming units (PFU). Another eight chickens were not infected as age-matched controls. Infected and control chickens (n=4) from both lines were terminated at 5 or 10dpi to collect bursa tissues. All procedures followed the standard animal ethics and use guidelines of ADOL.

Analysis of ChIP-Seq data

Chromatin immunoprecipitation (ChIP) was carried out using bursa samples from MDV infected and controls birds as described elsewhere [8]. Briefly, about 30 mg bursa samples were digested with micrococcal nuclease followed by end-repair with PNK and Klenow enzymes (NBE, Ipswich, MA, USA) and ChIP with the specific antibody. This was followed by addition of 3' adenine, Illumina adaptor ligation, PCR amplification (17 cycles) and size-selection (~ 150 bp). This was followed by cluster generation and sequencing on the Illumina Hi-Seq 2000. The antibodies used and the total number of reads obtained for each sample is listed in Appendix XIV.

Sequence reads were aligned to the May 2006 version of the chicken genome (galGal3) using bowtie version 0.12.7 [80]. Default alignment policies of bowtie were enforced: a valid alignment could have a maximum of two mismatches and if a read aligned equally well to multiple places in the genome, one was chosen at random. If multiple reads mapped to the same genomic location, only one was kept to avoid amplification bias.

Promoter Clustering

Promoters were defined as a 2 kb region surrounding the transcription start site (TSS) of a gene, e.g. TSS ± 1000 bp. Reads mapping to promoter regions of Ensembl genes [228] were tabulated into a matrix and analyzed using edgeR [118]. Separate analyses were performed for H3K4me3 and H3K27me3. Diffscores for each gene g were calculated as:

$$DS_g = sgn\left(logFC_g\right) * -\log_{10} p_g$$

where, sgn(*) is the signum function, logFCg and p_g are the log-fold change and p-values calculated by edgeR. Hierarchical clustering was performed in R [229] with the hclust() function using the Ward's minimum variance method to calculate distances. Clustering heatmaps were generated using the package ggplot [230]. For visualization purposes, DS values greater than 2 were replaced by 2 and those less than -2 by -2.

RNA-Seq Data Analysis

RNA-Seq reads obtained from Illumina Hi-Seq 2000 were analyzed as above. Detailed analysis of this data set can be found elsewhere (Fei, Z. et al. unpublished). Transcript abundances were approximated by the numbers of reads mapping to exons. As we did not intend to perform transcriptome assembly as part of this work, we did not perform splice-junction mapping. Read counts for Ensembl genes were extracted and tabulated for analysis with edgeR as above. Note that several transcripts had no mapped reads, indicating undetectable levels of expression. The DS scores were calculated as for the ChIP-Seq data. Hierarchical clustering was performed with the

Ward's minimum distance criterion and the clustering dendrogram was cut at height 150 to produce 19 clusters.

Co-clustering Analysis

We compared the RNA-Seq and ChIP-Seq clustering results by adopting the technique used in [38]. Briefly, overlaps between each pair of RNA-Seq and ChIP-Seq clusters were tabulated and tested for independence using a χ^2-test with simulated p-values (10000 iterations). Simulated p-values were used as the table of counts was likely to contain several zero counts in which case the test may be rendered inappropriate.

Results

Promoter clustering by dynamic chromatin changes

We sampled two critical time-points of MD progression, 5 and 10 dpi, representing early cytolytic and latent stages of MD, respectively. ChIP-Seq was performed on bursal tissues obtained from MD-resistant line $L6_3$ and MD-susceptible line $L7_2$ chickens. Two histone H3 trimethylation marks having opposing effects on gene regulation were profiled – H3 lysine 4 trimethylation (H3K4me3), which is associated with the 5' end of active genes, and H3 lysine 27 trimethylation (H3K27me3) which marks broad regions for silencing. To uncover gene promoters with similar dynamic patterns of chromatin we examined the 2 kb region centered around the transcription start sites (TSSs) of 16426 annotated genes in the chicken genome from the Ensembl database [228], which included most of the RefSeq genes in addition to predicted

genes and miRNAs. The promoter read counts were compared between infected and control groups within each line using edgeR [118]. We quantified the differences between MD-infected and control individuals by using log-fold changes (log*FC*) and p-values output by edgeR to score each promoter (diffscore). Thus, a p-value of 0.001 with a negative fold-change was scored as -2, while the same p-value with a positive fold-change was scored as +2. Subsequently, hierarchical clustering of diffscores was performed using the Ward's minimum distance criterion. A traditional threshold-based approach attempts to discover the largest variations. In contrast, our measure was aimed at being more inclusive as we were interested in finding enriched pathways. We believe this approach increases the sensitivity of our analysis towards detecting subtle variations in chromatin marks, which might still have an important role in determining transcriptional regulation.

We manually curated the clustering dendrogram and chose a cut height of 400 to obtain a list of 14 clusters (A-N; Figure 4.1). Two of the clusters (F, H) showed finer patterns that were revealed using a cut height of 150 on each clustering sub-tree and resulted in 3 (F1, F2, F3) and 2 (H1, H2) clusters, respectively. In addition, cluster C contained only 2 genes and was subsequently dropped from the analysis. Thus, the hierarchical clustering of diffscores resulted in a set of 16 clusters of promoters showing distinct dynamic patterns of H3K4me3 and H3K27me3.

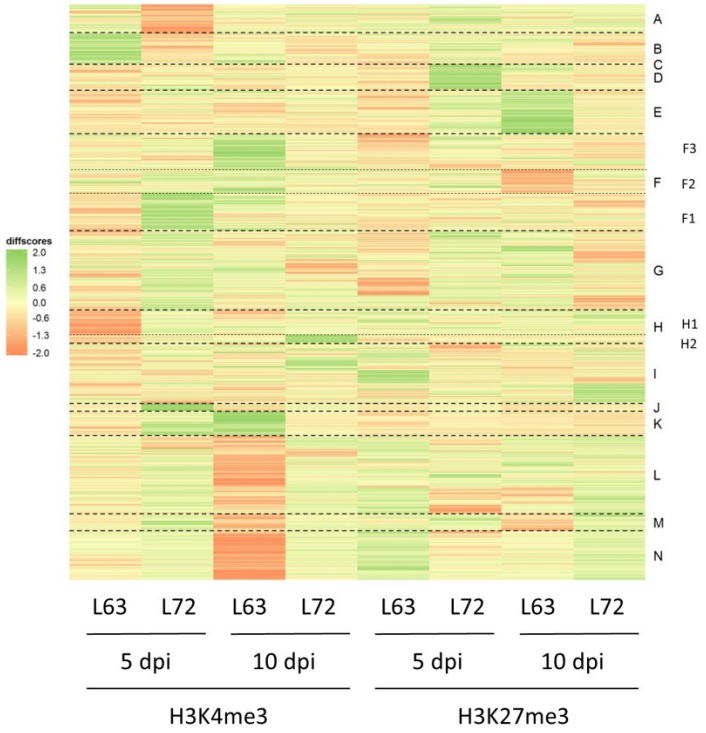


Figure 4.1. Hierarchical clustering of diffscores reveals dynamic chromatin changes. Unsupervised clustering of diffscores reveals striking patterns of chromatin as distinct clusters of promoters exhibit strong trends at each time-point. Line $L6_3$ shows a dramatic decrease in H3K4me3 marks at 10 dpi (clusters L, M, N), while both lines display corresponding changes at 5 dpi.

Several interesting trends were apparent from the above analysis. Distinct clusters of promoters exhibited changes in chromatin enrichment at the cytolytic and latent phases of infection. Moreover, disjoint sets of genes shared similar chromatin signatures in the two inbred chicken lines. For instance, cluster B consisted of genes showing a significant increase in H3K4me3 enrichment in $L6_3$ at 5 dpi, while cluster H1 demonstrated the opposite trend in the same line. In contrast, cluster F1 genes displayed increased H3K4me3 enrichment in $L7_2$ at the same time-point, while cluster A showed a decrease in promoter H3K4me3. Thus, the chromatin landscape revealed the dynamic nature of the epigenetic response in the two chicken lines at different stages of MD.

We conducted functional analysis of the clustered genes to uncover biological pathways and other functional terms associated with differences in chromatin enrichment induced by MDV infection. Clusters displaying similar trends were grouped together (Table 4.1) before gene set enrichment analysis with DAVID [170, 207].

Table 4.1. Cluster grouping based on similar chromatin trends.

DPI	Line	Trend*	Clusters
5	6_3	H3K4me3 ↑	B
		H3K4me3 ↓	H1
		H3K27me3 ↑	E
		H3K27me3 ↓	F2
	7_2	H3K4me3 ↑	F1, J
		H3K4me3 ↓	A
		H3K27me3 ↑	D
10	6_3	H3K4me3 ↑	F3, K
		H3K4me3 ↓	L, M, N
	7_2	H3K4me3 ↑	H2

*The trends summarized above are based on strong observed patterns in the corresponding clusters.

Apoptosis and p53 pathways show early H3K4me3 changes particularly in MD-susceptible chickens

At the early cytolytic stage, genes involved in the p53 signaling pathway (KEGG: gga04115) in both the resistant and susceptible lines displayed changes in H3K4me3 enrichment (Figure 4.2). However, there were several key differences. In line $L7_2$, genes associated with stress signals such as DNA damage, e.g. *ATM*, *CHEK2* and *STT3A*, exhibited increased H3K4me3 marks, while downstream p53 targets which induce the apoptosis pathway, e.g. *ZMAT3* and *CYCS*, showed similar chromatin patterns. In the resistant line, increased H3K4me3 enrichment was present on genes which are also involved in increased apoptosis (*CASP3*, *CASP8*, *BID* and *SHISA5*),

while upstream genes (*CHEK2*, *STT3A*) mirrored the changes observed in the susceptible line only at the later time-point. Moreover, in line $L6_3$, we saw increased H3K4me3 on genes associated with inhibition of angiogenesis and metastasis (*CD82*), and those that can promote DNA repair and damage prevention (*RRM2* and *SESN1*), which were absent in the susceptible line.

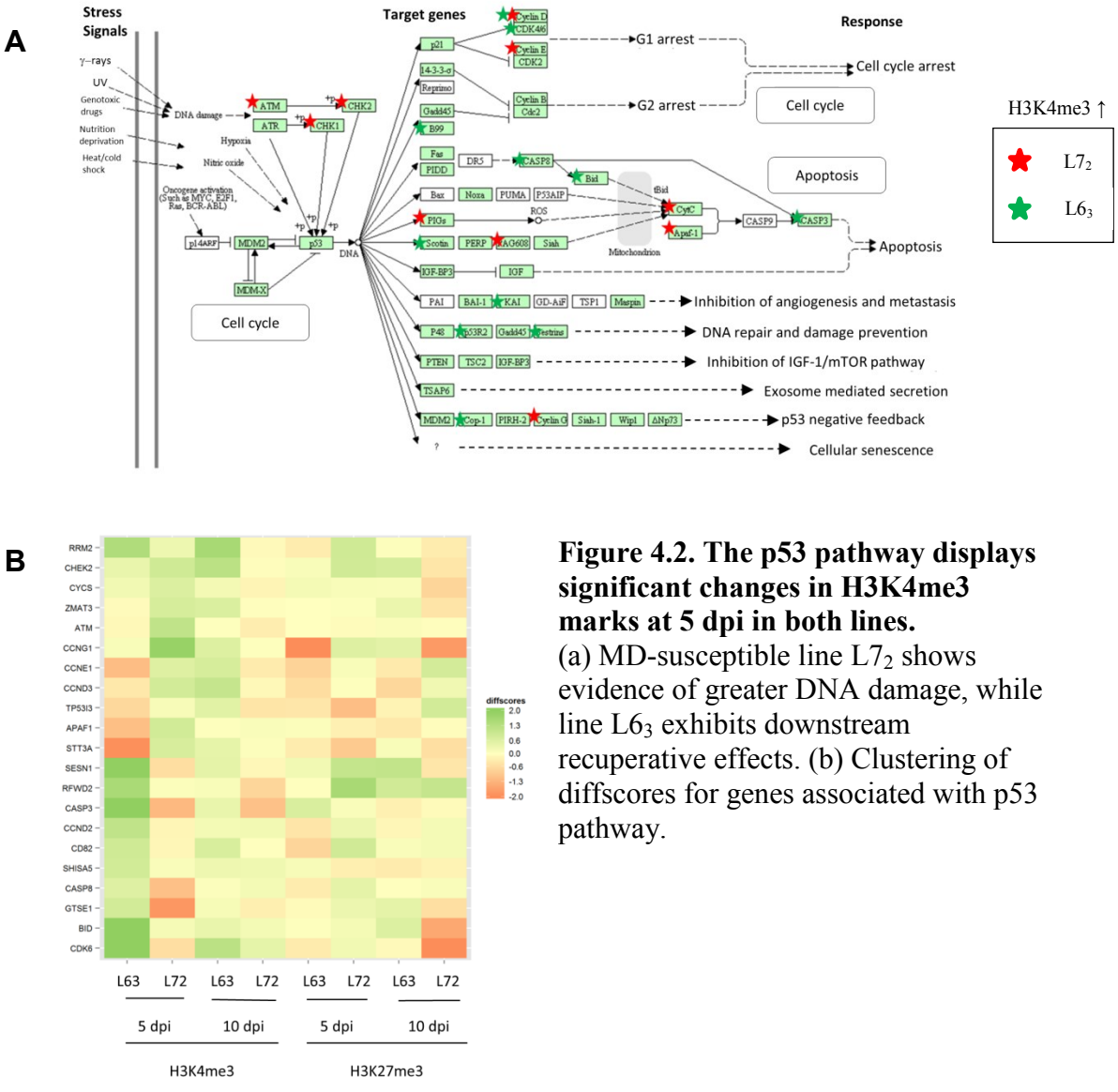


Figure 4.2. The p53 pathway displays significant changes in H3K4me3 marks at 5 dpi in both lines. (a) MD-susceptible line $L7_2$ shows evidence of greater DNA damage, while line $L6_3$ exhibits downstream recuperative effects. (b) Clustering of diffscores for genes associated with p53 pathway.

Several genes associated with the apoptosis pathway (KEGG: gga04210), displayed perturbed chromatin marks in response to cytolytic infection in line $L7_2$ (Figure 4.3).

The pro-inflammatory cytokine *IL1B* and downstream gene *MYD88*, which can induce the NF- κB signaling pathway, along with genes involved in PI3K-Akt signaling, such as, the nerve growth factor *NGFB*, showed increased H3K4me3 on their promoters. However, other genes exhibited contrasting signals. For instance, there was increased promoter H3K4me3 on *PI3KR2*, but a reduction on *PI3KCG*, an increase on cAMP-dependent protein kinase *PRKAR1B*, but a corresponding reduction on *PRKACB*. Other important genes associated with the apoptosis pathway, *FADD* and *CFLAR*, exhibited reduced H3K4me3 marks in the susceptible line while apoptosis-inhibitor *BIRC2* exhibited the same trend in both lines at 5 dpi.

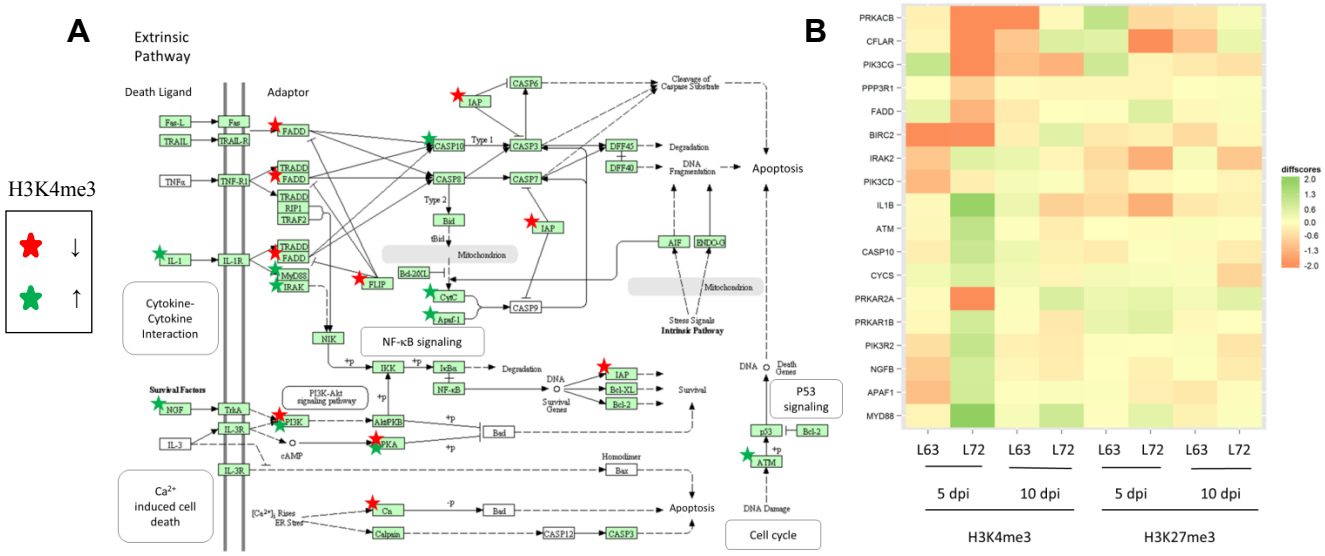


Figure 4.3. Apoptosis pathway shows H3K4me3 changes in line $L7_2$ at 5 dpi. (a) KEGG pathway and (b) clustering heatmap. Members of the NF-κB signaling pathway display increased H3K4me3 enrichment at 5 dpi, while apoptosis-related genes *FADD, CFLAR* and *BIRC2*, have reduced promoter H3K4me3.

Interestingly, the ubiquitin-mediated proteolysis pathway (KEGG: gga04120), which has been linked to the regulation of p53, also displayed significant changes in H3K4me3 marks in the susceptible line (Appendix XV). All three classes of enzymes involved in ubiquitination: ubiquitin-activating enzymes (E1s), ubiquitin-conjugating

enzymes (E2s) and ubiquitin-protein ligases (E3s) showed increased H3K4me3 enrichment in infected individuals including *UBA3*, *7* (E1), *UBE2A*, *2R2* (E2) and multiple classes of E3 enzymes and associated complex subunits, e.g. *ITCH* (HECT-type), *CBL* (U-box type), *PIAS4* (single RING-finger type), *SKP1*, *FBXO2* and *SOCS3* (multiple subunit RING-finger type).

Highly perturbed chromatin on the neuroactive ligand-receptor interaction pathway in MD-resistant chickens

Several genes involved in the neuroactive ligand-receptor interaction pathway (KEGG: gga04080), which is a collection of neural stimulatory molecules and their receptors, displayed striking changes in chromatin marks in the resistant line $L6_3$, at both stages of the disease (Figure 4.4).

Certain components of the pathway showed reduced H3K4me3 enrichment in the resistant line at 5 dpi. This included various G-protein coupled receptors (GPCRs), such as, the dopamine receptors (*DRD4*, *DRD5*), histamine receptor *HRH4*, 5-hydroxytryptamine (5-HT) receptor *HTR2A*, etc. However, a larger proportion of associated molecules demonstrated H3K4me3 reductions at the latent stage of MD including virtually all classes of GPCRs, (e.g. *DRD2*, *HTR1D*, *1E* and *1F*), among a variety of others, e.g. GABA receptors (*GABRA2*, *B2*, *D* and *G1*) and the growth hormone receptor *GHR*. In addition, several genes belonging to this pathway also displayed increased H3K27me3 marks at this time-point.

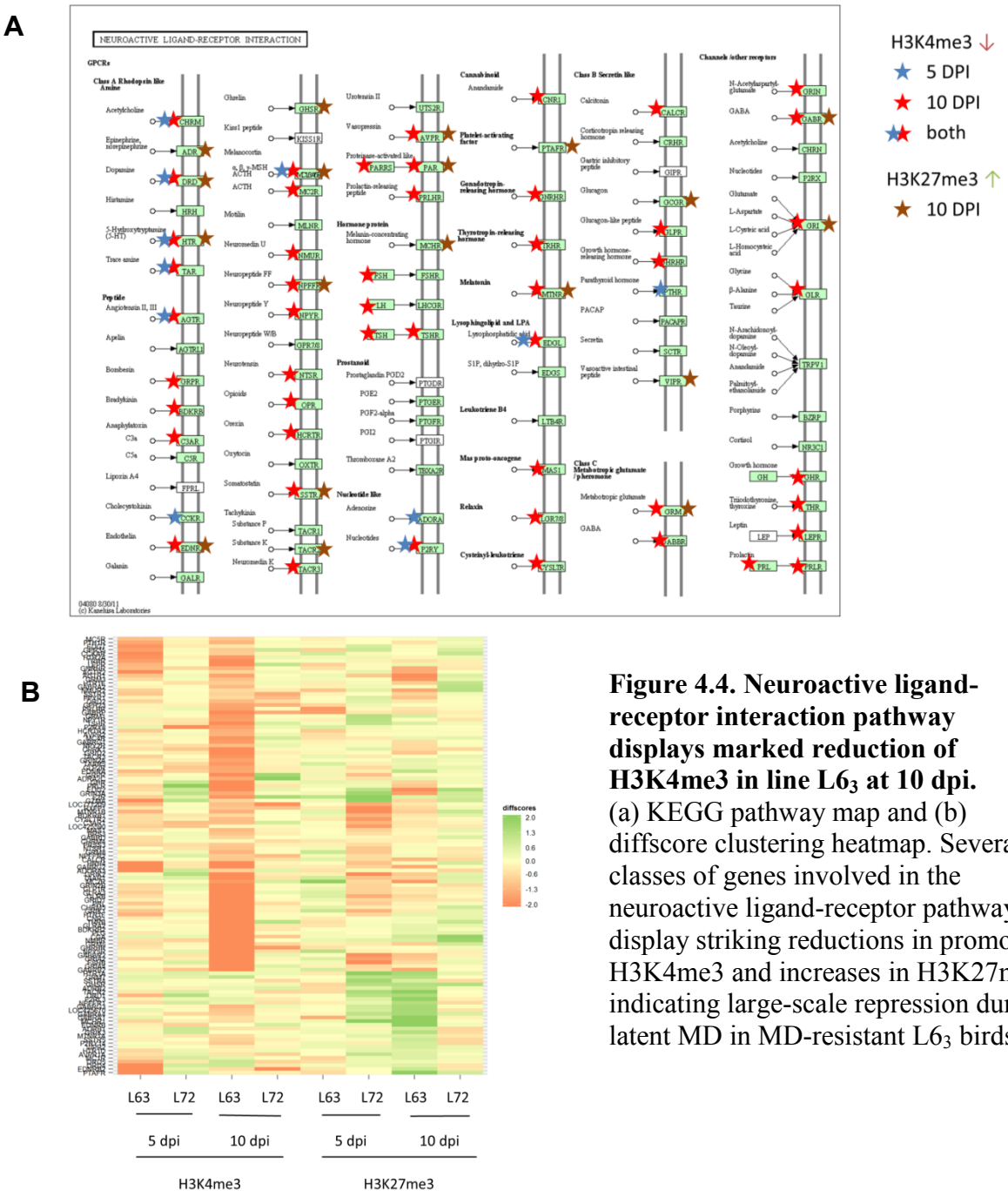


Figure 4.4. Neuroactive ligand-receptor interaction pathway displays marked reduction of H3K4me3 in line L6$_3$ at 10 dpi. (a) KEGG pathway map and (b) diffscore clustering heatmap. Several classes of genes involved in the neuroactive ligand-receptor pathway display striking reductions in promoter H3K4me3 and increases in H3K27me3 indicating large-scale repression during latent MD in MD-resistant L6$_3$ birds.

Thus, various members of the NLR interaction pathway displayed H3K4me3 reductions at either time-point or increased promoter H3K27me3 at 10 dpi, each suggesting possible repression. Some genes belonging to this pathway were also previously reported to be associated with MD. For instance, the proinflammatory

protease granzyme A (*GZMA*) has been shown to be upregulated during early cytolytic MD in susceptible chickens [162]. We found a significant reduction of H3K4me3 on the promoter of *GZMA* in line $L6_3$ at 5 dpi, while an increase was evident in susceptible chickens. Also, the growth hormone gene *GH1* has been associated with MD resistance [187], and shown to be upregulated in susceptible chickens [186]. Interestingly, the growth hormone receptor *GHR* displayed reduced promoter H3K4me3 in infected $L6_3$ chickens at 10 dpi while a slight increase was evident at 5 dpi in susceptible birds.

Signature cytokines and cytokine receptors show H3K4me3 alterations at the latent stage

Several cytokines and cytokine receptors (CCR interaction pathway, KEGG: gga04060) showed changes in H3K4me3 marks at 10 dpi (Figure 4.5). This included notable chemokine *IL8*, fractalkine receptor *CX3CR1* and interferons *IFNA* (*LOC768614*) and *IFNB*, all of which had reduced H3K4me3 in response to infection in the resistant line $L6_3$. Certain subfamilies were especially well-represented in this group, such as, hematopoietic interleukins (*IL7*, *12B*, *15*, *PRL* and *TPO*), and receptors *LEPR*, *OSMR* and *PRLR*; platelet-derived growth factors (PDGFs) *FIGF* and *HGF*, and PDGF receptors (*FLT1*, *KDR*, *KIT* and *MET*); IL-1 family receptors (*IL1R1*, *2* and *IL18RAP*) and TGFβ family receptors *ACVR2B* and *BMPR2*. In contrast, some components of this pathway displayed increased promoter H3K4me3 in line $L7_2$ at this time-point, which included the inflammatory cytokine *IL6*, interleukin receptors *IL7R* and *21R*, TNF superfamily receptors *TNFRSF1B*, *11B* and *FASLG*, IL-10 family receptor *IL22RA1* and TGFβ receptor II-like *LOC424261*.

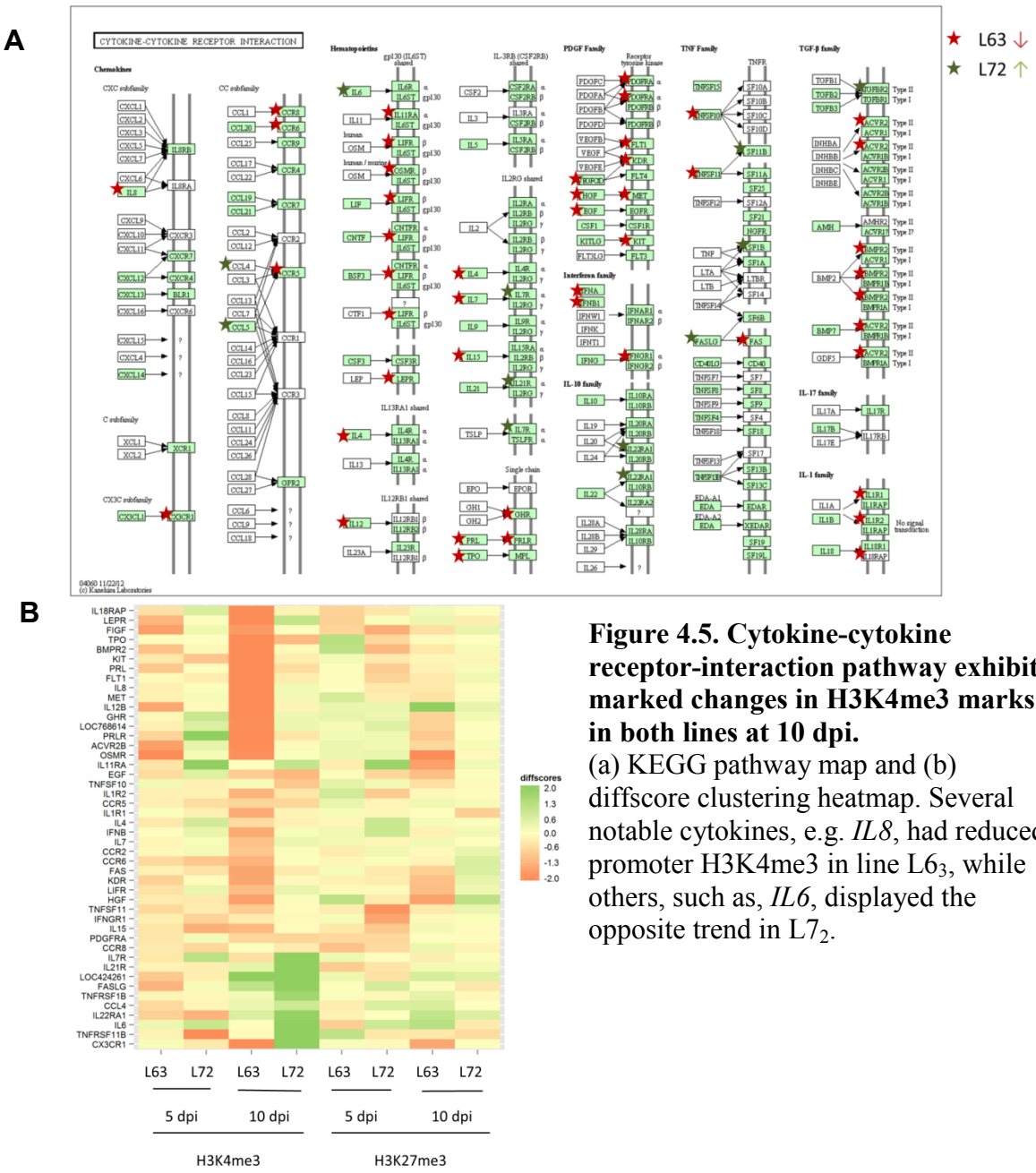


Figure 4.5. Cytokine-cytokine receptor-interaction pathway exhibits marked changes in H3K4me3 marks in both lines at 10 dpi. (a) KEGG pathway map and (b) diffscore clustering heatmap. Several notable cytokines, e.g. *IL8*, had reduced promoter H3K4me3 in line $L6_3$, while others, such as, *IL6*, displayed the opposite trend in $L7_2$.

The expression profiles of host cytokines in response to MDV infection has been studied previously [131, 160, 161]. The interleukin *IL6* is upregulated in MD-susceptible chickens [160] while inducible nitric oxide synthetase (iNOS) is higher in MD-resistant than susceptible chickens after 5 dpi [141]. Our observations of increased H3K4me3 on *IL6* in the MD-susceptible line suggesting activation, and an

evident reduction on *iNOS* at 5 dpi indicating repression, thus, appear to be consistent with the above. Moreover, there were several novel genes showing marked differences between the two lines. The chemokine receptor *CX3CR1*, displayed significantly increased promoter H3K4me3 in the susceptible line at 10 dpi, while the reverse was true of $L6_3$. The interleukin receptor *IL11RA* showed a marked increase in H3K4me3 enrichment in infected $L7_2$ chickens similar to *IL6* and *IL7R*, while the resistant line showed no change. On the other hand, IL6ST receptors *LEPR*, *OSMR*, TGFβ family receptors *ACVR2B*, *BMPR2*, and interleukin *IL12B* demonstrated reduced H3K4me3 in line $L6_3$ at both time-points. Notable similarities were also apparent between the two lines, e.g. *LOC424261* and *FASLG*, showed increased H3K4me3 in both lines at 10 dpi, while *EGF* displayed corresponding reductions.

MAPK signaling pathway displays H3K27me3 changes in both lines

Among the relatively few promoters with striking changes in H3K27me3 marks, several were associated with the MAPK signaling pathway (KEGG: gga04010) in both lines (Figure 4.6). In line $L7_2$ at the early cytolytic stage, this included elements of the classical MAPK pathway, such as, *PDGFA*, a growth factor involved in cell proliferation and migration, various Ras-related genes e.g. *RASA1* and *MRAS*, tyrosine kinase receptor *NTRK2* and transmembrane calcium channel *CACNG4*. In addition, multiple components of the JNK and p38 MAPK pathways also appeared to have perturbed levels of H3K27me3, such as, the proliferation-regulatory cytokine *TGFB2*, the p38 MAP kinase *MAPK12*, Ras-related small GTPase *RAC2*, and the myocyte transcription enhancer factor *MEF2C*.

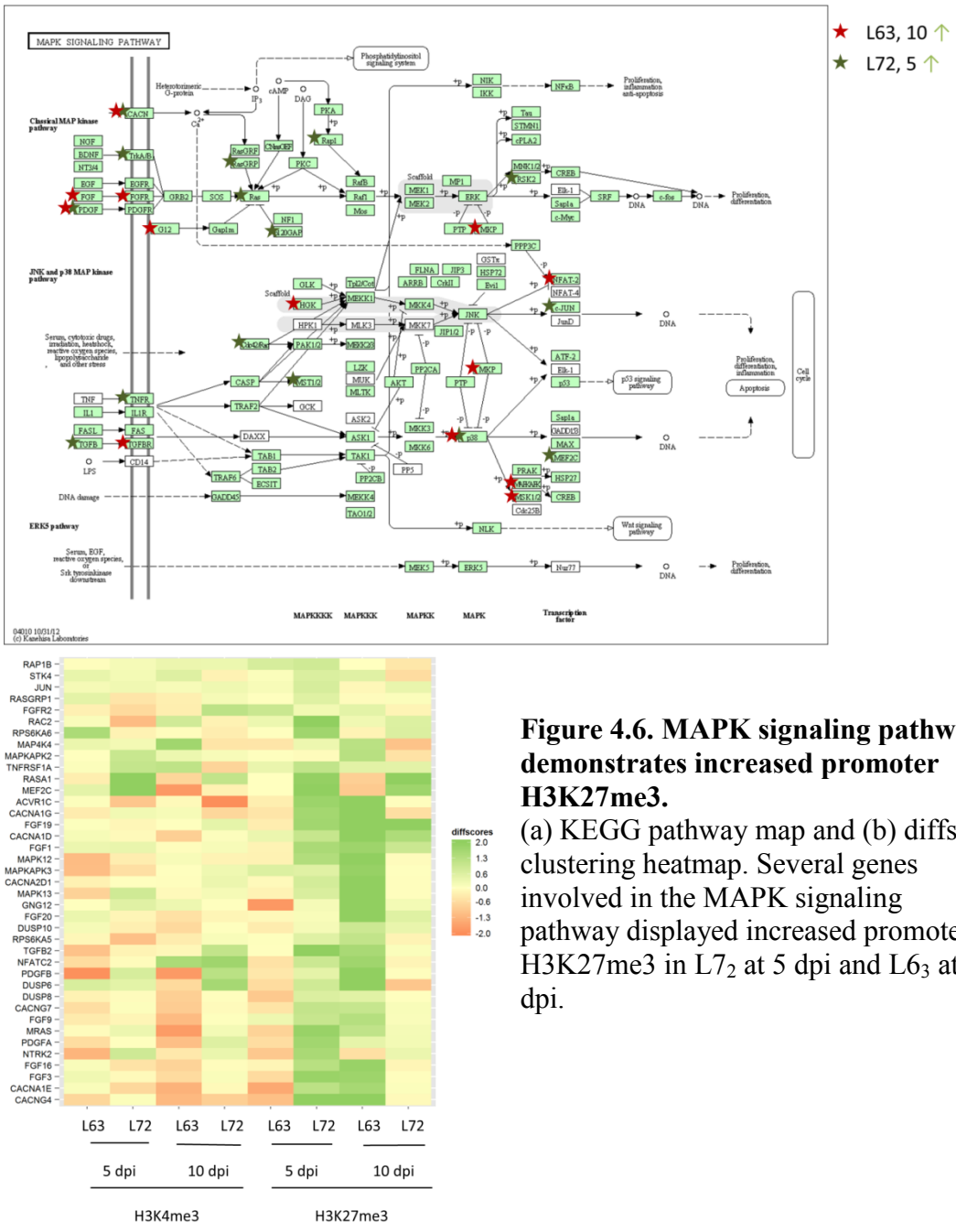


Figure 4.6. MAPK signaling pathway demonstrates increased promoter H3K27me3.
(a) KEGG pathway map and (b) diffscore clustering heatmap. Several genes involved in the MAPK signaling pathway displayed increased promoter H3K27me3 in $L7_2$ at 5 dpi and $L6_3$ at 10 dpi.

In the resistant line, increased H3K27me3 marks were observed at 10 dpi on several components of this pathway. This included several fibroblast growth factors (FGFs; *FGF1*, *3*, *16*, *19* and *20*), *PDGFB* and transmembrane calcium channels (*CACNA1D*, *1E*, *1G* and *2D1*), which were part of the classical pathway. However, a significant proportion of associated genes also belonged to p38 and JNK MAPK signaling, such

as, *TGFβ* family receptor *ACVR1C*, p38 MAP kinase *MAPK13*, dual-specificity phosphatase *DUSP6*, MAPK-activated protein kinases (*MAPKAPK2*, *3*) and *NFATC2*.

Multiple genes showed similar signatures at different time-points in the two lines. For instance, the TGFβ family receptor *ACVR1C* had increased promoter H3K27me3 in line $L7_2$ at 5 dpi and line $L6_3$ at 10 dpi. This was also true of several other genes, such as, growth factors *FGF1*, *3*, *16* and *19*, calcium channels *CACN1G* and *E*, *TGFB2* and MAP kinase *MAPK12*. Interestingly, *RASA1* and *MEF2C*, which showed increased H3K27me3 in line $L7_2$ at both time-points, also exhibited H3K4me3 increases at 5 dpi, but not at 10 dpi. Similarly, *TGFB2*, *PDGFB*, *DUSP6* and *NFATC2*, which displayed higher promoter H3K27me3, particularly at 10 dpi in line $L6_3$, demonstrated increased H3K4me3 in the susceptible line at the same time-point. Taken together, our results suggest overall silencing of this pathway in the susceptible line during cytolytic infection, which is abrogated by the latent stage. In contrast, the silencing occurs in the resistant line at a later stage of infection.

Novel pathways display chromatin variations

At the latent stage of the disease, the focal adhesion pathway (KEGG: gga04510), which consists of several sub-pathways, such as, extra-cellular matrix (ECM) interaction, CCR interaction and MAPK signaling, was highly represented in the resistant line $L6_3$ (Appendix XVI). The genes displaying reduced H3K4me3 marks included several collagens (*COL4A1*, *4A2*, *5A2* and *11A1*), laminins (*LAMA1*, *2*), thrombospondins (*THBS1*, *4*) and integrins (*ITGA1*, *A4*, *B1*, *B6* and *B8*). In addition, various components of the actin cytoskeleton regulatory mechanism such as

ARHGAP5 and *RHOA*, cytoskeletal protein *VCL*, protein kinases *ROCK1* and *ROCK2*, and other elements of the focal adhesion pathway such as AKT kinases (*AKT1*, *2*) and adherens junction component *CTNNB1* (catenin-β), displayed reduced H3K4me3 enrichment. Decreased H3K4me3 marks were also present on growth factor *IGF1*, IGF-1 receptor (*IGF1R*), oncogene *FYN* and SHC signaling adaptors *SHC3* and *SHC4*. Integrin signaling is believed to play an important role in MDV transformation [231], while some collagens are downregulated during lytic MD in MD-susceptible chickens [162]. Our results suggest that this pathway might undergo epigenetic regulation in response to MDV infection. Moreover, reduced H3K4me3 and possible downregulation of pro-neoplastic *IGF1* and its receptor, along with oncogene *FYN*, in the resistant line is also an interesting finding.

Another interesting pathway that contained a large number of genes with increased H3K4me3 marks, particularly in line $L6_3$ at 10 dpi, was the spliceosome pathway (KEGG: gga03040; Figure 4.7), which consists of molecules that regulate pre-mRNA splicing, such as, small nuclear ribonucleoproteins (snRNPs) U1-U6 and spliceosome-associated proteins (SAPs). Genes belonging to each of the above components of this pathway displayed increased H3K4me3 marks in response to MDV infection, e.g. *SNRPD1* and *D3* (U1), *SF3A1* and *PHF5A* (U2), *PRPF4* and *PPIH* (U4/6), *EFTUD2* (U3) and *BCAS2* (Prp19 complex). Some of these genes, such as, *ZMAT2*, *EFTUD2* and *PRPF8*, also demonstrated increased promoter H3K4me3 in the susceptible line at 5 dpi, further evidence of a possible 'phase-difference' in epigenetic response to the disease depending on the level of MD-susceptibility.

Moreover, this is a novel pathway, which, to the best of our knowledge, has not been previously reported in the context of MD progression.

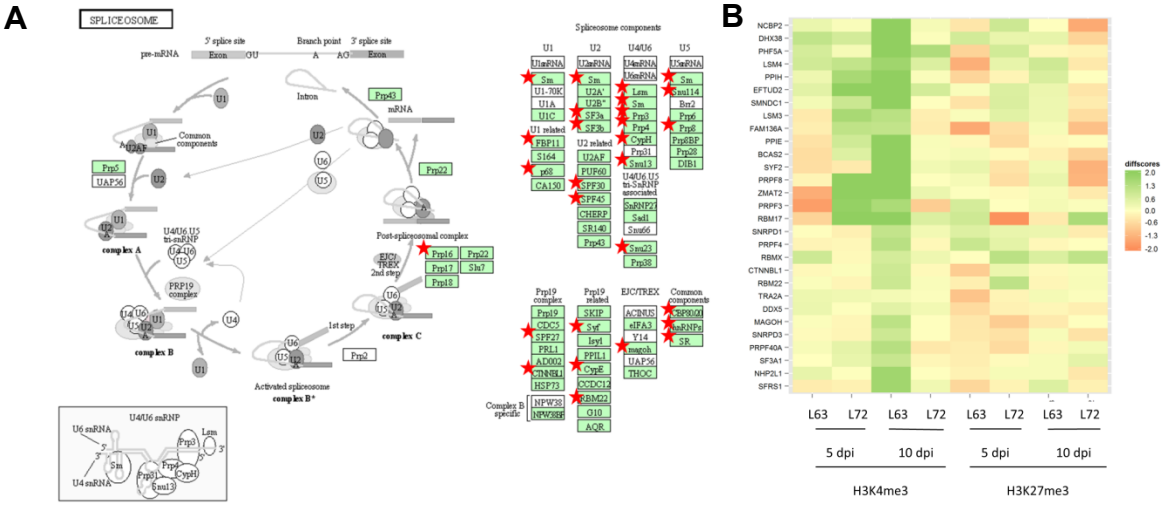


Figure 4.7. The spliceosome pathway shows increased H3K4me3 marks particularly in $L6_3$ at 10 dpi.
(a) KEGG pathway map and (b) diffscore clustering heatmap. Several genes belonging to this pathway had increased promoter H3K4me3 in resistant birds during latent infection, while some showed the same trend at the earlier time-point in susceptible $L7_2$ chickens.

Immune-related microRNAs demonstrate characteristic signatures

MicroRNAs (miRNAs) are short, non-coding RNAs that play a major role in post-transcriptional regulation via translational repression or mRNA destabilization. Several miRNAs have been shown to play major roles in immune response, e.g. miR-146 is a possible tumor suppressor [232] and along with miR-155 is believed to contribute to innate immunity [233]. The miR-17~92 cluster is thought to function as an oncogene, promoting cell proliferation and suppressing apoptosis [234]. We extracted a list of 449 chicken miRNAs and clustered their promoters using diffscores. The miRNAs formed characteristic clusters as observed in the case of coding transcripts (data not shown). To examine the broader epigenetic effect of MD,

we compiled a list of candidate miRNAs from various reports [235-238] and examined their temporal chromatin profiles (Figure 4.8).

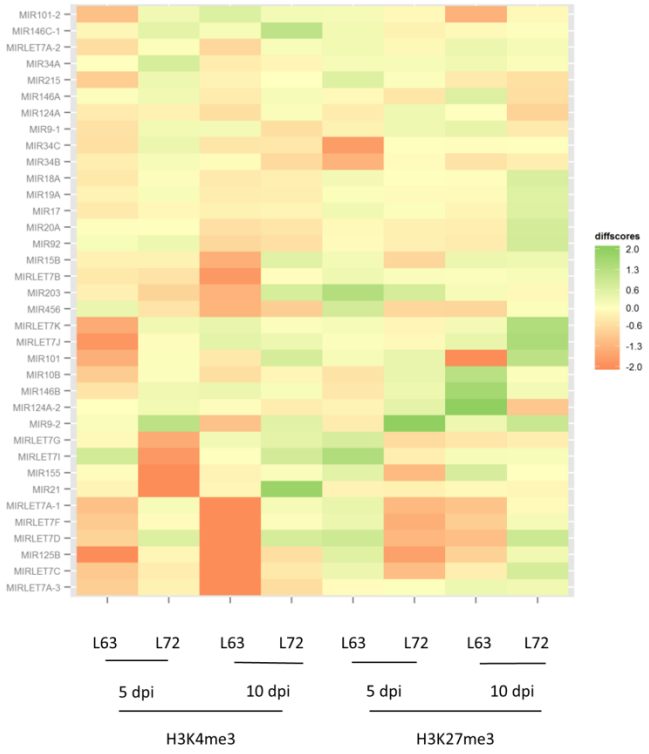


Figure 4.8. Selected immune-related miRNAs display repressive changes in chromatin marks.
Several members of the let-7 family had reduced promoter H3K4me3 marks in line $L6_3$ birds either at 5 or 10 dpi, while other important miRNAs, e.g. *gga-mir-21* and *gga-mir-155* displayed reductions in H3K4me3 in line $L7_2$ at 5 dpi.

Several miRNAs displayed predominantly repressive changes in chromatin in both lines. Multiple immune-related miRNAs *gga-mir-155* [239], *gga-mir-21* [240] and *gga-let-7i* [241] had reduced promoter H3K4me3 in the susceptible line at 5 dpi, but in the case of *gga-mir-21* and *gga-let-7i*, this trend was reversed at 10 dpi. Several other members of the let-7 family displayed reduced H3K4me3 in the resistant line $L6_3$ at the two time-points: *gga-let-7a-2*, *j* and *k* at 5 dpi and *gga-let-7a-1*, *a-3*, *b*, *c*, *d* and *f* at 10 dpi. *Gga-mir-125b*, which has been linked to certain cancers [242, 243],

showed reduced H3K4me3 in line $L6_3$ at both time-points. At 10 dpi, *gga-mir-101*, which can play a role in curbing autoimmune reactions [244], had reduced H3K27me3 marks in line $L6_3$, while other immune-related miRNAs, *gga-mir-10b*, *124a-2* and *146b* displayed the reverse trend. The inhibition of miR-10b has been associated with reduced metastasis [245], while loss of miR-124a functions as a tumor suppressor [246]. Interestingly, both these miRNAs displayed increased H3K27me3 only in the resistant line, suggesting line-specific silencing.

Chromatin signatures distinguish genes with similar expression patterns

We compared the chromatin marks with RNA-Seq data from the same tissue to look for possible correlations. The expression data was analyzed with edgeR, as above, and diffscores obtained from the comparison of infected and control samples were clustered to obtain a set of 19 groups (Appendix XVII). The two clustering results showed definite correlation (χ^2-test, $p < 10\text{-}6$). The χ^2-residuals were subsequently plotted (Figure 4.9).

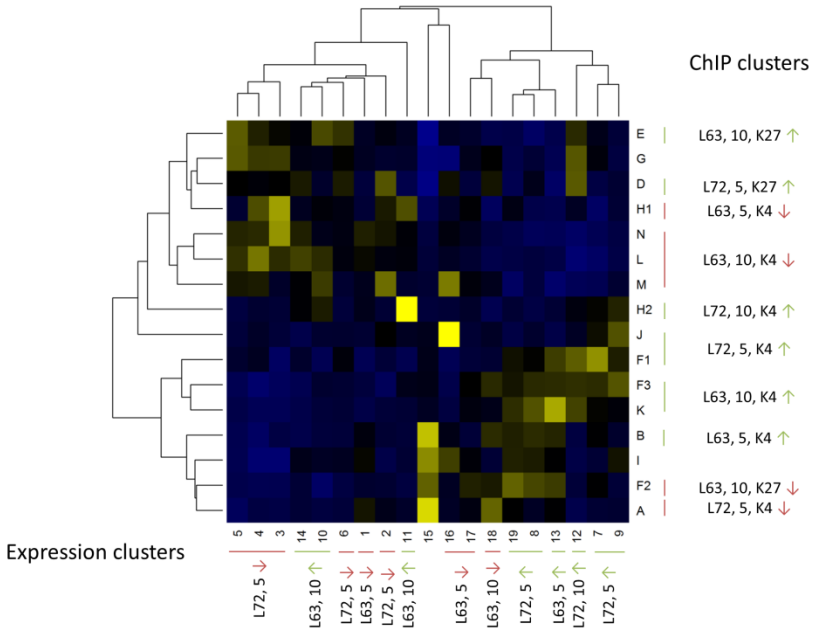


Figure 4.9. Hierarchical clustering of diffscores for RNA-Seq data and co-clustering with ChIP clusters.
Co-clustering analysis reveals that genes with the same expression profiles can have diverse chromatin signatures.

We found that genes with similar temporal expression patterns displayed remarkable diversity in chromatin marks and vice-versa. There were a few clusters that showed a certain level of correlation between the chromatin marks and expression. For example, chromatin clusters F1 and J which displayed increased H3K4me3 marks in the susceptible line at 5 dpi overlapped expression clusters 7 and 9 which consisted of genes upregulated at that time-point in the same line. However, such correlations were largely low. This is consistent with several prior reports [247] that emphasize the diversity of the epigenetic regulatory landscape as evidenced by the expanding array of histone modifications with specific roles.

Discussion

The histone code is a universal, multi-layered guide to the transcriptional regulatory machinery that allows tremendous diversity to be encoded into the genome, while providing an essential link between the genetic material and environmental cues. Interpreting the biological consequences of variations in chromatin marks is exceedingly complex and can be likened to an attempt to discern the outcome of a voluminous treatise from the preface. The task of understanding the broader genomic effects of a complex disease, such as MD, from epigenetic profiling is a similarly daunting undertaking. Our prior studies of the epigenetic effects of latent MD on resistant and susceptible chicken lines [219, 226] have provided us with some perspective. However, the chromatin landscape is dynamic and temporal analyses of histone modification profiles are necessary to obtain a more complete picture. Thus, in contrast to the candidate gene approach of the earlier studies, we conducted a more comprehensive functional analysis of chromatin variations induced by MD.

Major functional differences in response to MDV infection

There were broad similarities, together with striking differences, between the resistant and susceptible lines in response to MD. The most striking difference was the NLR interaction pathway, with variations in chromatin marks on a wide variety of genes. We have previously reported the possible association of this pathway to MD response via miRNAs [219], but the sheer number of differentially-marked genes suggests a significantly greater level of involvement and warrants further investigation. Several cytokines and cytokine receptors showed reductions in promoter H3K4me3 in MD-

resistant chickens, while others displayed the reverse trend in the susceptible line. The MAPK signaling pathway, which had significant representation in the proteome of an MDV-transformed cell line [231], displayed H3K27me3 increases predominantly in the resistant line at 10 dpi. Also, several genes in the spliceosome pathway showed increased promoter H3K4me3 in line $L6_3$.

Interestingly, in the latter two cases, certain genes shared similar chromatin profiles in both lines, but at different time-points. This suggests a possible 'phase-difference' in the epigenetic response to MD depending on the level of susceptibility of the chickens. Also, a large proportion of chromatin changes were repressive in nature (H3K4me3 ↓, H3K27me3 ↑), and appeared at a later stage of the disease in the resistant line. Epigenetic reprogramming of host genes by viruses and other pathogenic microbes has been associated with gene silencing [248] and it is possible that this is another example of such a phenomenon.

Apoptosis in both lines during lytic MD

Virus-induced apoptosis or programmed cell death can occur either as a result of host defense mechanisms eliminating infected cells or as a mode of increased replication and spread of virus particles [249]. We observed an enrichment of the apoptosis pathway among differentially marked promoters in line $L7_2$ during lytic infection, with possible involvement of NF-κB signaling (*IL1B* and *MYD88*). However, certain other genes, which are also critical for inducing apoptosis, e.g. caspases *CASP3*, *CASP8*, Bcl-2 family death regulator *BID*, and *SHISA5*, which can induce apoptosis in a p53-dependent manner, displayed increased H3K4me3 marks in the resistant line at the same time-point. NF-κB signaling is an important component of host immune

response to infection, and is frequently associated with inflammatory diseases and tumors [250]. NF-κB also plays a major role in MDV-induced transformation of CD30+ lymphocytes [163]. Early stages of MD have been associated with inflammatory changes in susceptible chickens [251], and the activation of NF-κB signaling could be part of an inflammatory response in line $L7_2$ chickens. Therefore, while higher levels of apoptosis are possibly clearing greater numbers of infected cells and thus, lowering viral load in line $L6_3$ chickens, the activation of a different subset of genes could be causing inflammatory response in line $L7_2$.

At the early cytolytic stage, the p53 pathway demonstrated significant changes in H3K4me3 in both lines, but the genes differentially marked in each line suggested contrasting outcomes – the susceptible line displayed signs of greater DNA damage, while the resistant line showed evidence of increased DNA repair and recuperative effects. The p53 protein functions as a tumor suppressor and is known to be targeted and inhibited by the viral oncoprotein Meq [123, 213]. We have previously observed variations in chromatin profiles of genes associated with p53 [226], but this is the first direct evidence of multiple components of this pathway undergoing epigenetic variations at early stages of MD. The E3 ubiquitin ligase, Mdm2 [252], responsible for p53 degradation did not exhibit epigenetic changes suggesting activation (cluster F2), but several other components of the ubiquitin-mediated proteolysis pathway displayed increased H3K4me3 in the susceptible line, suggesting possible activation of this pathway during cytolytic infection.

Novel candidates for epigenetic regulation

The variations in chromatin profiles of some MD-associated genes, such as, *IL6* and *GZMA* (increased H3K4me3 in line $L7_2$), which were upregulated in susceptible chickens [160, 162], suggested epigenetic regulation in response to virus infection. We also observed increased H3K4me3 around the pro-inflammatory cytokine *IL1B* (line $L7_2$ at 5 dpi), which was upregulated in brain tissue of chickens infected with MDV [251]. In addition, several novel candidates were also revealed. For instance, *CX3CR1*, which is important for efficient chemotaxis of macrophages to apoptotic lymphocytes [253], displayed contrasting trends in the two lines. Various cytokines sharing the IL6ST subunit have been found to induce proliferation in cases of multiple myeloma [254]. Receptors belonging to the above class showed changes in chromatin marks, e.g. *IL11RA* in line $L7_2$ and *LEPR*, *OSMR* in line $L6_3$. Previous reports have also indicated the involvement of the pro-inflammatory cytokine induced gene *IRG1* in MD susceptibility. This gene is preferentially upregulated in susceptible chickens [162] and involved in inflammatory response via the action of *MYD88* which displayed increased promoter H3K4me3 (and mRNA levels) in the susceptible line at 5 dpi. *MYD88* is an essential regulator of immunity to invading microbes, particularly the activation of T cell responses [255] and, thus, could be an interesting candidate for further study. Moreover, the increase of both H3K4me3 and H3K27me3 in line $L7_2$ on *RASA1* and *MEF2C* suggest their involvement in MD-susceptibility.

The let-7 family of miRNAs have diverse physiological roles and its deregulation has been associated with many human cancers [241]. Several members of this family exhibited striking reductions in promoter H3K4me3, primarily in the resistant line

(except *gga-let-7i*), which suggested possible epigenetic silencing in response to MDV infection. Oncomir *gga-mir-21*, which has also been associated with several cancers [240, 256, 257], displayed H3K4me3 variations in the susceptible line (decrease at 5 dpi and increase at 10 dpi). Also, H3K27me3 appeared to target certain immune-related miRNAs, e.g. *gga-mir-10b* and *gga-mir-124a-2*, only in the resistant line suggesting their involvement in MD-resistance. Previous studies of miRNA expression profiles conducted in our lab [238] suggested large scale down-regulation of host miRNAs during late cytolytic MD (spleen, 21 dpi) in susceptible chickens. Our results indicate a different scenario at the early stages of MD which further underlines the importance of temporal analyses to uncover a truer picture of transcriptional regulation.

Conclusions

In summary, we conducted a comprehensive analysis of the temporal chromatin landscape induced by MDV in two inbred chicken lines with contrasting responses to the disease. We investigated the variations in chromatin marks in response to virus infection to uncover biological pathways possibly under epigenetic control. In doing so, we eschewed a traditional threshold-based analysis, instead utilizing the entire gene set and unsupervised clustering to find groups of promoters that displayed similar patterns of chromatin. Our approach revealed several interesting pathways with large proportions of genes displaying variations in chromatin, such as neuroactive ligand-receptor interaction and apoptosis. Epigenetic variations suggested a heightened inflammatory response during lytic MD in the susceptible line while there appeared to be increased apoptosis and greater virus clearance in the resistant

line. At the latent stage of infection, the resistant line demonstrated widespread reduction in promoter H3K4me3 suggesting epigenetic silencing. Our observations with regard to certain MD-related genes were largely in agreement with previous reports. In addition, we uncovered several novel genes and miRNAs that undergo epigenetic regulation and are possibly associated with MD-resistance or susceptibility.

5. Conclusions and Future Directions

The long-term objective of our laboratory is to understand the epigenetics of MD, the complex disease of poultry, and the mechanisms of MD-resistance and susceptibility. Our chosen model to achieve this goal is a population of inbred chicken lines 6_3 and 7_2 from ADOL, MI that are naturally either highly MD-resistant or highly MD-susceptible. In keeping with this, the focus of my graduate research has been two-fold: the development of novel methods for analyzing genome-wide epigenetic data, e.g. histone modifications, and the application of these and other methods to the data generated from the above chicken population. The works presented here constitute novel contributions in both these areas.

We developed WaveSeq, a novel algorithm for peak-detection in ChIP-Seq data that is accurate, sensitive and robust to diverse enrichment patterns. Our approach is unique as we do not make any restrictive (and erroneous) assumptions about the data distribution, which is a feature of virtually all existing tools primarily for the purposes of computational efficiency. The accuracy of our method relies on the discriminative power of wavelets for pattern recognition. We employed Monte Carlo sampling techniques to estimate the distribution of wavelet coefficients, effectively constraining the wavelet space for pattern detection. Finally, we assign significance scores to predicted peaks by utilizing a novel permutation procedure. WaveSeq performed favorably in comparison with existing methods, particularly for diffuse histone modification data. We believe our method addresses a long-unfulfilled need

of the scientific community and we are working towards a full release on the R platform free for public use.

In conjunction with the development of WaveSeq, we embarked on our study of the histone modification landscape in our chosen animal model. Our work in spleen and thymus tissues of MD-resistant and susceptible chickens resulted in the first publications related to chromatin marks in poultry [219, 226]. Due to the novelty of our approach, these studies have a strong exploratory element. In our investigations of latent MD in thymus, we employed ChIP-Seq to profile H3K4me3 and H3K27me3 in matched infected and control birds from lines 6_3 and 7_2. Several genes previously implicated in MD progression, e.g. *MX1* and *CTLA-4*, and others associated with various cancers, such as, *IGF2BP1* and *GAL*, exhibited line-specific or condition-specific enrichments. Moreover, bivalent chromatin domains, thought to be predominantly associated with developmental genes, were observed on several genes. Three of these genes were p53-associated transcription factors, *EGR1*, *BCL6* and *CITED2*, and associated chromatin signatures showed identical responses to MDV infection. Thus, we demonstrated that MDV induces large-scale variations in chromatin marks, with differential effects in resistant and susceptible chickens.

The next step in our journey was the extension of our efforts to the temporal evolution of chromatin marks in response to MDV infection. We conducted this experiment only 12 months after our initial studies and even within this short interval, Illumina sequencers had improved by several orders of magnitude making it possible to include multiplexing in our protocol. As a result, we were able to generate a much more comprehensive data set, including two time-points of MD progression and two

biological replicates in our experimental design. As histone modifications are context-specific, we reasoned that *changes* in chromatin enrichment are evidence of epigenetic regulation. Following this intuition, we analyzed promoter regions of annotated genes and miRNAs for *differential* H3K4me3 and H3K27me3 enrichments. The results of this analysis were quantified using a measure we termed 'diffscore', which we subjected to hierarchical clustering for evidence of coregulation. Functional analysis of clustered promoters revealed several interesting features: during early cytolytic MD, the susceptible line showed evidence of greater DNA damage and inflammation (possibly via NK-κB signaling), while resistant chickens appeared to have higher apoptosis rates and recuperative symptoms (downstream p53 targets). At the latent stage, line $L6_3$ displayed marked repressive changes on the neuroactive ligand-receptor interaction pathway. Several immune-related miRNAs, e.g. multiple members of the let-7 miRNA family, showed reduced H3K4me3 at 10 dpi in line 6_3, while others, such as, *gga-mir-21* and *gga-mir-155* displayed similar trends in the susceptible line at the earlier time-point. In addition, various MD-associated genes, e.g. *IL6*, *GZMA* and *IL1B*, displayed repressive changes in the susceptible line after infection, consistent with reported trends. Thus, this extensive study gave us further insights into the epigenetic effects of MD, although further work is necessary to confirm some of these findings.

In many of our studies, we adopted the commonly-used approach of assigning functional significance to chromatin marks by annotating putative enrichments with adjacent genes. Whilst analysis of such gene lists for functional enrichment can well serve as indications of biological involvement, there is always the possibility that

there are additional factors that determine the biological outcome. For instance, assaying a representative set of histone modifications is a valid approach to investigate the chromatin landscape. However, the transience of a majority of histone marks suggests that such a view is nothing more than a snapshot of a dynamic system under a specific set of conditions and definitive predictions based on such a fleeting picture is fraught with the possibility of error. Therefore, extensive temporal analyses are necessary as the maturation of NGS technology makes such experimental designs more accessible. However, the validation of ChIP-Seq findings remains difficult and time-consuming, with causal relationships nearly impossible to prove.

The discovery of numerous histone modifications led to the 'histone code' hypothesis, which proposed the existence of a system of epigenetic marks that can define the functional elements of the genome in a combinatorial and deterministic manner. However, over the years this simplistic view has been replaced by the understanding that chromatin signatures comprise a nuanced and subtle network, which only forms part of the transcriptional regulation machinery. Instead of studying each such component in isolation, integrative approaches are necessary, wherein multiple sources of information, such as, transcription factor binding, DNA methylation, copy number variations and gene expression, are studied together with histone modifications. The recent tool, ZINBA [102], attempts to provide a general solution to the problem by using a mixture regression approach. However, the extensive computational requirements are major limiting factors for such analyses and can only be justified by a sizeable improvement in accuracy. Thus, amidst the

profusion of peak callers there is an urgent need for large-scale benchmarking efforts in order to make the choice of a suitable method a little easier.

Appendices

Appendix I. Sequencing results showing the antibody used and read numbers for each sample from bursa of Fabricius at 5 days post infection.

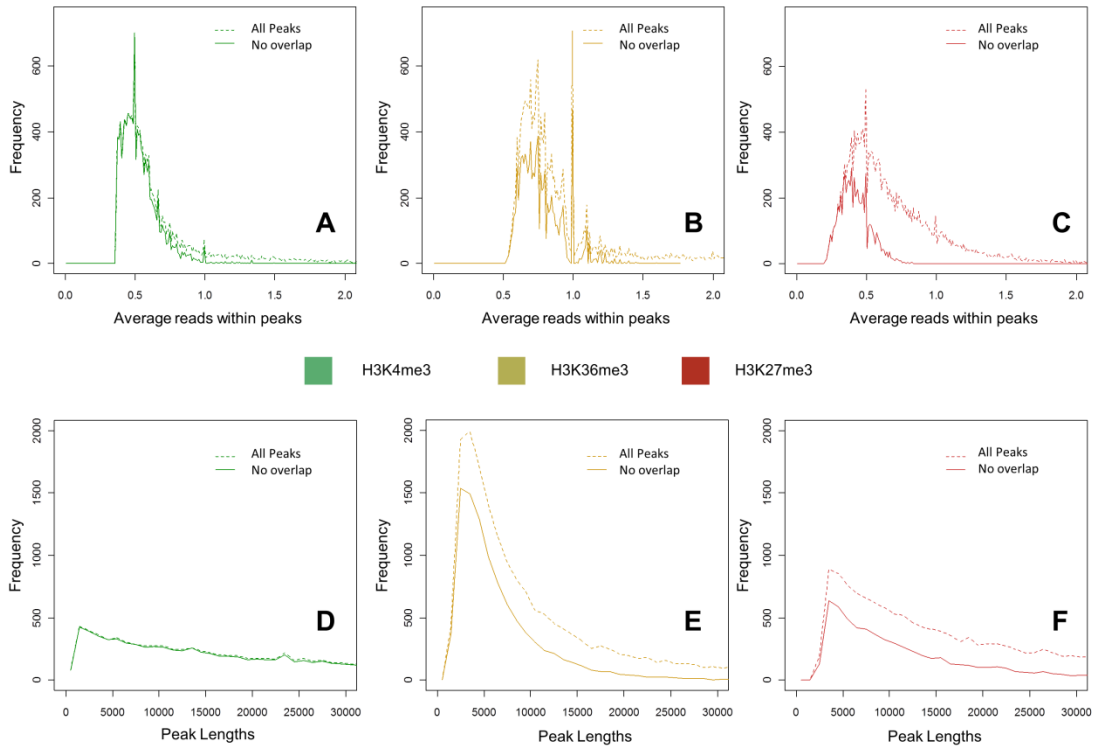
Antibody	Sample	Raw	Mapped	Mapped%	Non-redundant	Non-redundant%
H3K4me3	R.ctl	9623392	6284246	65.30178	5483970	56.98583
(Millipore,	R.inf	5589297	4641859	83.04907	3763076	67.32646
Cat. #17- 614)	S.ctl	8333094	4790077	57.48257	4432040	53.18601
	S.inf	7298525	5189925	71.10923	4479908	61.38101

$Mapped\% = \frac{\#\ mapped\ reads * 100}{\#\ raw\ reads}$; $Non-redundant\% = \frac{\#\ non-redundant\ reads * 100}{\#\ mapped\ reads}$.

R.ctl = line 6_3 control, R.inf = line 6_3 infected, S.ctl = line 7_2 control, S.inf = line 7_2 infected.

Appendix II. RSEG peaks not detected by WaveSeq have low average read counts and are possibly false positives.

Average read counts within RSEG peaks (a, b & c) and peak length distributions (d, e & f) in the H3K4me3 (a & d), H3K36me3 (b & e) and H3K27me3 (c & f) data. The solid lines correspond to all peaks called by RSEG (All Peaks) and the dashed lines represent those peaks that are not detected by WaveSeq (No overlaps). These plots show that WaveSeq detects a majority of large RSEG peaks in the H3K27me3 and H3K36me3 data. However, most of the H3K4me3 peaks detected by RSEG are very large and appear to be false positives. The average read counts output by the program were plotted.



Appendix III. List of H3K4me3 DMRs and overlapping genes

The chromosome, start and end columns refer to the significant DMRs detected by WaveSeq. The columns S.inf and S.ctl contain the normalized reads (per million) mapped to the DMRs in the infected and control samples of the S group, respectively. P-values are calculated by WaveSeq using an exact binomial test and fold change = (S.inf+1)/(S.ctl+1). The columns RefSeq_ID and Ensembl_ID contain RefSeq and Ensembl genes that overlap the corresponding DMRs.

Chr	Start	End	RefSeq_ID	Ensembl_ID	S.inf	S.ctl	FC	p-value	FDR
chr1	9000	20199	CD69	ENSGALG00000009761	119.86	64.03	1.86	5.82E-05	0.009021
chr1	5832600	5843199	-	ENSGALG00000006713	135.05	82.70	1.63	0.000391	0.029701
chr1	14735200	14746199	-	ENSGALG00000008167	186.45	125.47	1.48	0.000645	0.039621
chr1	15029200	15038799	-	-	66.59	32.37	2.03	0.00077	0.043481
chr1	15388000	15397999	PIK3CG	ENSGALG00000008081	220.33	135.15	1.63	7.52E-06	0.002114
chr1	28607600	28615199	-	ENSGALG00000009443	114.25	51.04	2.21	1.04E-06	0.000451
chr1	32535200	32546799	-	ENSGALG00000009665	167.99	111.05	1.51	0.000938	0.048411
chr1	35649000	35656599	-	-	78.21	131.31	0.60	0.000301	0.043824
chr1	46275800	46287999	-	ENSGALG00000019338	214.26	144.84	1.48	0.000255	0.023634
chr1	48574800	48582599	-	ENSGALG00000011516	67.06	32.84	2.01	0.000562	0.037928
chr1	52668000	52674199	-	ENSGALG00000011531 ENSGALG00000023146	101.17	53.16	1.89	0.000136	0.0155
chr1	53366800	53374799	CYTH4	ENSGALG00000012454	136.22	84.35	1.61	0.000555	0.037928
chr1	53379200	53385599	RAC2	ENSGALG00000012490 ENSGALG00000012456	122.90	73.72	1.66	0.000553	0.037928
chr1	61677000	61684599	-	-	99.53	56.00	1.76	0.00049	0.035743
chr1	63058000	63067799	ADIPOR2	ENSGALG00000013000	134.12	75.85	1.76	5.42E-05	0.008658
chr1	64218000	64224999	-	ENSGALG00000013057	97.90	26.46	3.60	7.97E-11	1.73E-07
chr1	69309800	69325199	-	ENSGALG00000014011	240.66	136.57	1.76	9.05E-08	6.87E-05
chr1	70414000	70423799	PPFIBP1	ENSGALG00000014106	117.29	61.67	1.89	3.26E-05	0.006499
chr1	71352400	71360599	-	ENSGALG00000014203	87.15	46.78	1.84	0.000477	0.035335
chr1	75541600	75547199	CCND2	ENSGALG00000017283	91.59	50.56	1.80	0.000699	0.040987
chr1	78757000	78769599	FOXM1	ENSGALG00000013424	132.71	82.46	1.60	0.000771	0.043481
chr1	79043200	79047599	AICDA	ENSGALG00000013420 ENSGALG00000014280	135.05	82.70	1.63	0.000391	0.029701
chr1	80183600	80188799	-	-	144.86	81.52	1.77	3.20E-05	0.006499
chr1	80320800	80330199	MLF2	ENSGALG00000014468	145.10	92.62	1.56	0.000698	0.040987
chr1	80674800	80685199	ZYX	ENSGALG00000014688	154.91	90.26	1.71	5.02E-05	0.008282
chr1	81189400	81201199	-	-	227.81	142.48	1.59	1.13E-05	0.002901
chr1	92041400	92056799	-	ENSGALG00000015398	373.84	253.53	1.47	1.84E-06	0.000682
chr1	94886400	94900399	-	-	242.76	153.35	1.58	8.77E-06	0.002377
chr1	95046800	95056999	-	-	182.48	121.69	1.50	0.000545	0.037807
chr1	95259600	95270999	LOC396098	ENSGALG00000015461	369.40	245.26	1.50	6.36E-07	0.000292
chr1	101540800	101552599	SAMSN1	ENSGALG00000015679	190.42	121.69	1.56	0.000109	0.013225
chr1	109612200	109622399	RUNX1	ENSGALG00000016022	142.53	83.64	1.70	0.000101	0.01283

chr1	116747800	116754799	-	ENSGALG00000016261	103.51	57.18	1.80	0.000344	0.028366
chr1	119277200	119286199	-	-	82.48	43.95	1.86	0.000618	0.039418
chr1	125765800	125777399	MOSPD2 FANCB	ENSGALG00000016569	146.73	83.17	1.76	3.77E-05	0.007225
chr1	126826000	126835199	TLR7	ENSGALG00000016590	146.50	86.24	1.69	9.88E-05	0.012633
chr1	129227600	129244599	-	ENSGALG00000021892	135.05	85.53	1.57	0.000912	0.047596
chr1	140524800	140538599	-	-	96.96	52.69	1.82	0.000373	0.029344
chr1	144368400	144377599	TNFSF13B	ENSGALG00000016852	166.83	104.20	1.60	0.000194	0.020125
chr1	144389600	144403199	ABHD13 LIG4	ENSGALG00000016853 ENSGALG00000016854	194.63	121.69	1.59	4.63E-05	0.007893
chr1	170074000	170090599	-	ENSGALG00000016947	371.04	233.21	1.59	2.17E-08	2.73E-05
chr1	170225000	170237199	-	ENSGALG00000016954	227.57	148.62	1.53	5.31E-05	0.008567
chr1	170986400	170988799	-	ENSGALG00000016964	12.29	38.08	0.34	0.000306	0.044243
chr1	172334400	172345599	LCP1	ENSGALG00000016986	146.73	85.77	1.70	7.22E-05	0.010434
chr1	172404000	172418999	-	ENSGALG00000016988	179.21	101.60	1.76	3.66E-06	0.001156
chr1	173527200	173532399	-	ENSGALG00000019094	50.00	21.74	2.24	0.000767	0.043481
chr1	173562400	173571599	-	ENSGALG00000017008	85.28	45.37	1.86	0.000572	0.038095
chr1	173671800	173675199	-	-	11.81	36.22	0.34	0.000346	0.046853
chr1	174685200	174694799	CKAP2	ENSGALG00000017025 ENSGALG00000017026	116.59	67.10	1.73	0.00036	0.029344
chr1	178865000	178878399	BRCA2 ZAR1L	ENSGALG00000017073 ENSGALG00000017074	169.16	97.82	1.72	1.20E-05	0.003024
chr1	183458400	183467199	PSPC1	ENSGALG00000017142	107.01	63.80	1.67	0.000919	0.047759
chr1	183588800	183597199	PARP4	ENSGALG00000017146 ENSGALG00000017148	125.24	75.85	1.64	0.000499	0.036272
chr1	183916600	183922399	-	-	82.48	43.48	1.88	0.000618	0.039418
chr1	185429600	185435999	-	ENSGALG00000017174	69.16	32.37	2.10	0.000296	0.026037
chr1	193389200	193400599	-	ENSGALG00000017247	178.74	105.15	1.69	1.69E-05	0.003902
chr10	1013400	1022399	-	-	122.67	70.41	1.73	0.000215	0.021151
chr10	6554200	6560799	-	ENSGALG00000003809	115.42	69.23	1.66	0.00086	0.045898
chr10	13802400	13821199	-	-	319.63	195.17	1.63	5.01E-08	4.01E-05
chr10	14303800	14313599	-	ENSGALG00000006505	109.58	63.80	1.71	0.000561	0.037928
chr10	16315600	16324799	-	-	156.08	91.21	1.70	4.24E-05	0.007537
chr10	16330400	16340199	-	ENSGALG00000006949	63.80	123.37	0.52	1.29E-05	0.005
chr11	1044000	1050999	-	ENSGALG00000021442	64.02	30.48	2.07	0.000588	0.038465
chr11	3337200	3348599	-	-	106.54	59.54	1.78	0.000314	0.026814
chr11	18750000	18760399	COTL1	ENSGALG00000017644 ENSGALG00000020995 ENSGALG00000005651	178.04	99.71	1.78	2.39E-06	0.000865
chr12	4733000	4745199	VGLL4	ENSGALG00000004937	199.77	137.28	1.45	0.000849	0.04571
chr12	6958200	6972799	-	ENSGALG00000005237	235.99	134.45	1.75	1.63E-07	0.000103
chr12	7464200	7474199	-	ENSGALG00000005385	88.09	44.18	1.97	0.00016	0.017498
chr12	9208600	9224399	IP6K2	ENSGALG00000005701	360.76	227.30	1.58	4.48E-08	3.78E-05
chr12	11063800	11071799	CHCHD4	ENSGALG00000006328 ENSGALG00000006345	117.53	69.23	1.69	0.000533	0.03747
chr12	13862800	13871199	-	-	95.56	51.75	1.83	0.000339	0.02831

chr13	2220200	2234199	MATR3	ENSGALG00000002478	177.57	105.15	1.68	2.14E-05	0.004852
chr13	3876800	3889799	-	ENSGALG00000002080	167.29	101.60	1.64	6.63E-05	0.009773
chr13	8409200	8425199	CSNK1A1	ENSGALG00000001364	154.68	99.71	1.55	0.000657	0.040051
chr13	8509000	8515399	-	ENSGALG00000001210	54.21	14.89	3.48	1.11E-06	0.000468
chr13	10681400	10695999	UBLCP1	ENSGALG00000003672 ENSGALG00000003691	108.18	60.49	1.78	0.000264	0.023821
chr13	11470600	11481399	-	ENSGALG00000003818	100.94	57.18	1.75	0.00075	0.043081
chr13	13217800	13228399	RPS14 CD74	ENSGALG00000004588 ENSGALG00000004594	57.89	109.35	0.53	6.65E-05	0.015909
chr13	13655800	13670199	HNRNPH1	ENSGALG00000005955	183.88	103.49	1.77	2.59E-06	0.000895
chr13	13690400	13701799	-	ENSGALG00000005989	192.99	126.65	1.52	0.000255	0.023634
chr13	17015400	17030599	-	-	201.17	126.88	1.58	3.97E-05	0.007269
chr13	17456800	17463199	IRF1	ENSGALG00000006785	40.66	14.18	2.74	0.000535	0.03747
chr14	34400	41999	-	ENSGALG00000002796	88.79	41.82	2.10	4.27E-05	0.007537
chr14	820400	829999	PARN	ENSGALG00000003091 ENSGALG00000003111	104.21	58.36	1.77	0.000375	0.029344
chr14	2322400	2336999	-	-	174.30	113.89	1.53	0.00038	0.029563
chr14	3478000	3486799	CARD11	ENSGALG00000004398	143.46	82.46	1.73	5.75E-05	0.009018
chr14	3538000	3546599	SDK1	ENSGALG00000004420	124.54	71.36	1.73	0.000181	0.019029
chr14	6167200	6179399	TBL3	ENSGALG00000005458 ENSGALG00000005465 ENSGALG00000025687 ENSGALG00000005558	122.20	70.18	1.73	0.000215	0.021151
chr14	6639200	6652599	LCMT1	ENSGALG00000005962 ENSGALG00000005973	173.37	116.25	1.49	0.000955	0.048948
chr14	9155000	9169599	SOCS1	ENSGALG00000007158	148.60	84.83	1.74	3.17E-05	0.006499
chr15	4994400	5009799	DDX55	ENSGALG00000003298 ENSGALG00000003314	104.68	57.65	1.80	0.000263	0.023821
chr15	5463000	5469199	-	ENSGALG00000003863	58.41	26.70	2.14	0.000628	0.039475
chr15	5752200	5765599	-	ENSGALG00000004379	125.00	76.08	1.63	0.000672	0.040492
chr15	5954600	5963399	-	ENSGALG00000004493 ENSGALG00000004515	103.04	58.13	1.76	0.000487	0.035682
chr15	6355200	6364599	NAA25 TRAFD1	ENSGALG00000004780 ENSGALG00000004802	101.40	55.76	1.80	0.000287	0.025728
chr15	6479000	6489399	PTPN11	ENSGALG00000023491	85.52	46.78	1.81	0.000832	0.045331
chr15	8923800	8933199	-	ENSGALG00000006695	169.86	109.64	1.54	0.000384	0.029701
chr16	60000	65999	BMA1 TAPBP	ENSGALG00000000158 ENSGALG00000008022	70.56	35.21	1.98	0.000821	0.045177
chr17	5462600	5470999	FAM102A	ENSGALG00000005074	107.95	61.67	1.74	0.000481	0.035409
chr18	6990600	7001199	KPNA2	ENSGALG00000003584	108.18	60.72	1.77	0.000264	0.023821
chr18	10739400	10756399	GRB2	ENSGALG00000008016	175.47	110.82	1.58	0.000141	0.015755
chr19	1228400	1234199	-	-	88.55	45.84	1.91	0.000242	0.022787
chr19	3094400	3108999	-	ENSGALG00000001410	179.44	114.83	1.56	0.000175	0.018648
chr19	7315200	7323999	MIR21	ENSGALG00000021733	180.61	112.23	1.60	8.24E-05	0.011372
chr19	8979800	9002399	EVI2A	ENSGALG00000005588	320.33	200.13	1.60	1.60E-07	0.000103
chr19	9776000	9787599	-	ENSGALG00000006005 ENSGALG00000006011	135.28	84.35	1.60	0.000693	0.040987
chr19	9867600	9874199	-	ENSGALG00000006048	132.01	73.25	1.79	4.56E-05	0.007868

chr2	10329200	10337599	-	ENSGALG00000006723	105.84	60.25	1.74	0.000572	0.038095
chr2	19928000	19939599	RSU1	ENSGALG00000008720	144.86	87.19	1.65	0.000214	0.021151
chr2	22863400	22871199	-	ENSGALG00000009479	60.52	10.87	5.18	8.00E-10	1.22E-06
chr2	25992400	25999999	-	-	75.24	38.75	1.92	0.000641	0.039537
chr2	26788000	26797199	-	ENSGALG00000010777	105.84	58.83	1.79	0.000288	0.025728
chr2	39805400	39818399	TGFBR2	ENSGALG00000011442	222.67	143.19	1.55	4.18E-05	0.007537
chr2	41617800	41627999	-	ENSGALG00000011574	127.57	77.74	1.63	0.000569	0.038095
chr2	46427200	46435599	ELMO1	ENSGALG00000012093	149.30	87.90	1.69	6.55E-05	0.009773
chr2	51290400	51298199	PSMA2	ENSGALG00000019598	116.36	61.67	1.87	4.32E-05	0.007542
chr2	51745000	51759599	-	ENSGALG00000012337 -	200.94	138.93	1.44	0.00088	0.046554
chr2	60674600	60679999	-	-	55.14	22.92	2.35	0.000217	0.021151
chr2	63051600	63064399	RBM24	ENSGALG00000012712	155.14	101.84	1.52	0.000889	0.046703
chr2	82940600	82957199	IKZF1	ENSGALG00000013086	426.64	264.16	1.61	7.42E-10	1.22E-06
chr2	91573200	91582599	INVS	ENSGALG00000013441	111.68	62.38	1.78	0.000241	0.022787
chr2	92042400	92047799	ISG12-2	ENSGALG00000013452 ENSGALG00000013575	45.56	13.47	3.22	3.01E-05	0.006349
chr2	92549400	92568199	-	ENSGALG00000013628	130.66	213.79	0.61	8.67E-06	0.003529
chr2	92807000	92817599	-	-	83.88	42.06	1.97	0.00031	0.026638
chr2	98936800	98945399	C2H18orf1	ENSGALG00000013886	118.46	65.21	1.80	0.000109	0.013225
chr2	109884800	109893599	-	-	138.55	86.72	1.59	0.000623	0.039475
chr2	114789000	114816799	LYN	ENSGALG00000018967	401.41	254.71	1.57	1.02E-08	1.40E-05
chr2	129130000	129140599	NBN	ENSGALG00000015912	84.11	40.40	2.06	9.65E-05	0.012633
chr2	154012200	154019999	LY6E	ENSGALG00000015913 ENSGALG00000016152	77.34	33.79	2.25	3.30E-05	0.006499
chr20	6692000	6696799	-	ENSGALG00000004859	59.58	15.36	3.70	2.55E-07	0.000155
chr20	8313200	8319399	-	ENSGALG00000005609	52.57	21.50	2.38	0.000371	0.029344
chr20	8425800	8435999	SLC17A9	ENSGALG00000005711	131.78	81.05	1.62	0.000727	0.042269
chr20	9705200	9712599	-	-	40.89	14.18	2.76	0.000535	0.03747
chr20	9975200	9991399	BCL2L1	ENSGALG00000006211	253.28	150.04	1.68	3.26E-07	0.000171
chr20	10800000	10810199	TPX2 -	ENSGALG00000006267 -	97.20	55.76	1.73	0.000825	0.045191
chr20	11580000	11588199	-	ENSGALG00000007640	133.18	76.79	1.72	9.79E-05	0.012633
chr20	11977600	11988599	-	ENSGALG00000007757	109.11	63.80	1.70	0.000561	0.037928
chr20	12030000	12042999	-	ENSGALG00000007768 ENSGALG00000020895	229.91	139.41	1.64	3.14E-06	0.001059
chr20	12509600	12520999	-	-	178.98	111.05	1.61	9.68E-05	0.012633
chr21	1741400	1750199	SKI	ENSGALG00000001229	90.19	45.60	1.96	0.000134	0.015386
chr21	1906400	1913999	GNB1	ENSGALG00000001334	82.01	43.95	1.85	0.000618	0.039418
chr21	2677400	2691199	LOC419429	ENSGALG00000002005	252.34	150.04	1.68	4.12E-07	0.000208
chr21	4799200	4810199	NBL1	ENSGALG00000004043	98.60	47.97	2.03	2.76E-05	0.00589
chr22	307600	318799	ARHGAP25	ENSGALG00000000132	74.53	36.86	2.00	0.000371	0.029344
chr22	336000	350599	PCNA	ENSGALG00000000165	142.76	82.94	1.71	7.37E-05	0.010555
chr22	2445400	2456199	C22H20orf30 PLEKHA2	ENSGALG00000000169 ENSGALG00000000171 ENSGALG00000003349	118.69	64.74	1.82	7.64E-05	0.010684

chr23	99600	108599	-	-	130.61	73.01	1.78	7.67E-05	0.010684
chr23	176000	188999	-	ENSGALG00000000519	141.83	84.12	1.68	0.000176	0.018648
chr23	248400	257799	-	ENSGALG00000000562	151.87	80.57	1.87	3.48E-06	0.001147
chr23	4078800	4085399	-	ENSGALG00000002021	236.22	145.55	1.62	3.63E-06	0.001156
chr23	5896400	5903799	SRSF10	ENSGALG00000004133	74.53	38.75	1.90	0.000861	0.045898
chr24	4347800	4358599	POU2AF1	ENSGALG00000006809	207.01	124.05	1.66	5.91E-06	0.001759
chr24	5704600	5720999	DDX6 CXCR5	ENSGALG00000021251 ENSGALG00000007675	235.29	133.03	1.76	1.17E-07	8.09E-05
chr26	2975200	2983799	-	ENSGALG00000001373	93.46	52.22	1.77	0.000833	0.045331
chr26	3369000	3371199	MOV10	ENSGALG00000023899	6.38	27.10	0.26	0.000324	0.044602
chr3	2365000	2372799	XPO1	ENSGALG00000004377	128.74	77.97	1.64	0.000451	0.033709
chr3	7631800	7645199	EHD3	ENSGALG00000009086	141.12	89.79	1.57	0.000736	0.042668
chr3	17525200	17536799	SRSF7	ENSGALG00000013825 ENSGALG00000023495 ENSGALG00000013821 ENSGALG00000013819	120.10	72.54	1.65	0.000655	0.040051
chr3	23135000	23144799	-	ENSGALG00000009828	71.03	36.86	1.90	0.000923	0.047825
chr3	23424800	23439399	TRAF5	ENSGALG00000009864	197.67	129.96	1.52	0.000197	0.020321
chr3	24231000	24236399	-	-	98.83	55.05	1.78	0.000637	0.039537
chr3	28942600	28952199	-	-	150.47	85.77	1.75	2.67E-05	0.00579
chr3	31412600	31420399	-	ENSGALG00000010149	86.92	27.88	3.04	2.34E-08	2.73E-05
chr3	32338800	32348799	RASGRP3	ENSGALG00000010435	179.44	117.91	1.52	0.000375	0.029344
chr3	33238000	33247199	EIF2AK2	ENSGALG00000023188 ENSGALG00000010560	150.47	54.58	2.73	1.21E-11	3.06E-08
chr3	34825000	34834799	-	ENSGALG00000010612	110.28	63.56	1.72	0.000437	0.03283
chr3	39374400	39383199	GPR137B	ENSGALG00000010843	161.45	102.31	1.57	0.000331	0.027882
chr3	40011800	40016799	-	-	56.08	25.75	2.13	0.000752	0.043081
chr3	44712200	44724199	-	ENSGALG00000020005	133.18	82.70	1.60	0.000616	0.039418
chr3	47834000	47842799	-	-	110.28	49.38	2.21	1.47E-06	0.000558
chr3	49995400	50006399	-	ENSGALG00000012359	91.12	48.67	1.85	0.000333	0.027921
chr3	58815600	58824999	STX7	ENSGALG00000002930	122.43	71.12	1.71	0.000297	0.026037
chr3	62261000	62271399	NCOA7	ENSGALG00000014834	136.45	82.94	1.64	0.000311	0.026638
chr3	62284400	62292599	-	-	128.04	73.25	1.74	0.000128	0.014869
chr3	66395400	66404799	-	ENSGALG00000014937 ENSGALG00000014940	99.30	57.42	1.72	0.000966	0.049211
chr3	66447400	66454199	-	-	44.16	13.94	3.02	4.71E-05	0.007944
chr3	66455000	66465999	FAM26E	ENSGALG00000014961	190.42	106.56	1.78	1.20E-06	0.000493
chr3	77930400	77938399	-	-	77.57	40.64	1.89	0.0008	0.044674
chr3	85865600	85873199	LMBRD1	ENSGALG00000016174	107.01	62.38	1.70	0.000668	0.040379
chr3	97873800	97877799	-	ENSGALG00000016398	33.41	10.87	2.90	0.000606	0.039317
chr3	97880800	97884199	-	ENSGALG00000016400	21.26	4.25	4.24	0.000911	0.047596
chr3	100739400	100749599	TRIB2	ENSGALG00000016457	146.26	84.59	1.72	5.23E-05	0.008543
chr3	110034800	110045999	-	-	153.98	91.68	1.67	8.69E-05	0.011881
chr4	1435200	1443199	GPR174	ENSGALG00000004111	87.85	47.02	1.85	0.000695	0.040987
chr4	1457800	1463599	ITM2A	ENSGALG00000004107	116.82	69.94	1.66	0.000678	0.040688

chr4	11303800	11313399	-	ENSGALG00000007482	61.92	29.30	2.08	0.000973	0.049383
chr4	12993800	13009999	MAGT1	ENSGALG00000007861 ENSGALG00000023641 ENSGALG00000007863 ENSGALG00000007902	159.82	101.37	1.57	0.000388	0.029701
chr4	18848000	18856399	-	ENSGALG00000009177	87.39	47.73	1.81	0.000695	0.040987
chr4	21107600	21114799	TLR2-2	ENSGALG00000009239	148.83	82.46	1.80	1.61E-05	0.003902
chr4	33107000	33118799	-	ENSGALG00000010022 ENSGALG00000010031	102.10	56.00	1.81	0.000219	0.021151
chr4	36092800	36102799	-	ENSGALG00000010324	99.53	54.82	1.80	0.000343	0.028366
chr4	36171000	36181399	-	ENSGALG00000020220	72.67	21.50	3.27	1.04E-07	7.51E-05
chr4	47665200	47674599	-	-	205.14	133.03	1.54	0.000106	0.013225
chr4	50710200	50723799	TMEM66 SRP72	ENSGALG00000011395 ENSGALG00000024482 ENSGALG00000011403	100.70	57.89	1.73	0.00075	0.043081
chr4	56545800	56554399	METTL14	ENSGALG00000012000	153.74	98.53	1.55	0.000625	0.039475
chr4	58662600	58671799	-	ENSGALG00000012048 ENSGALG00000012063	110.52	60.25	1.82	0.000155	0.017164
chr4	61723000	61732799	DAPP1	ENSGALG00000000056	88.32	48.44	1.81	0.000763	0.043481
chr4	64244000	64254399	-	-	160.98	100.66	1.59	0.000239	0.022787
chr4	70911800	70930799	-	-	299.31	187.61	1.59	4.27E-07	0.000209
chr4	71445000	71450999	-	-	167.53	96.40	1.73	1.42E-05	0.003527
chr4	85898200	85907199	-	ENSGALG00000015690	208.18	133.50	1.56	5.76E-05	0.009018
chr4	86948600	86959599	SLBP	ENSGALG00000015712	105.14	39.22	2.64	3.53E-08	3.15E-05
chr4	86991000	87006399	FAM53A	ENSGALG00000015713	283.88	181.70	1.56	2.52E-06	0.000888
chr4	88807800	88817199	KDM3A	ENSGALG00000015803	123.83	71.59	1.72	0.000232	0.022266
chr4	88860000	88873199	RNF103 RMND5A	ENSGALG00000015809 ENSGALG00000015815	98.13	55.05	1.77	0.000637	0.039537
chr5	11410200	11415799	-	-	114.25	68.76	1.65	0.000804	0.044699
chr5	14740200	14749599	CD81	ENSGALG00000006546	162.15	107.27	1.51	0.000958	0.048954
chr5	15723400	15732999	BRSK2	ENSGALG00000006681	125.24	69.23	1.80	7.06E-05	0.010304
chr5	16703800	16711799	-	ENSGALG00000006841	91.12	45.37	1.99	9.90E-05	0.012633
chr5	20523600	20535799	CD44	ENSGALG00000007849	256.55	164.93	1.55	8.31E-06	0.002293
chr5	25090200	25099999	SPI1	ENSGALG00000008127	205.61	141.53	1.45	0.000685	0.040931
chr5	27791600	27798999	GANC	ENSGALG00000009018 ENSGALG00000009036	116.36	67.10	1.72	0.00036	0.029344
chr5	30238600	30247399	-	-	169.40	98.29	1.72	1.65E-05	0.003902
chr5	36498000	36504599	G2E3		99.77	51.75	1.91	0.00011	0.013225
chr5	45570200	45582199	ZC3H14	ENSGALG00000010616 ENSGALG00000010622	117.99	66.16	1.77	0.000201	0.020456
chr5	47503200	47513199	LOC423422	ENSGALG00000017387	126.87	76.08	1.66	0.000533	0.03747
chr5	48531200	48542199	GLRX5	ENSGALG00000011079	93.93	51.98	1.79	0.000585	0.038465
chr5	50600600	50606999	-	ENSGALG00000011139	103.74	60.72	1.70	0.000944	0.048588
chr5	50899400	50908799	EVL	ENSGALG00000011209	133.88	83.41	1.60	0.000816	0.045177
chr5	52993200	53001999	XRCC3	ENSGALG00000011533 ENSGALG00000011534	103.51	55.29	1.86	0.000166	0.018016
chr5	60273400	60280199	ARF6	ENSGALG00000012267	135.98	81.52	1.66	0.000291	0.025822

				ENSGALG00000012268					
chr6	5447400	5459999	-	ENSGALG00000002414	123.37	73.25	1.68	0.000436	0.03283
chr6	11716200	11725799	DNA2	ENSGALG00000024347 ENSGALG00000004037	149.30	95.46	1.56	0.000661	0.040118
chr6	16445600	16453799	-	-	109.11	64.74	1.67	0.000774	0.043491
chr6	17827200	17845399	PIK3AP1	ENSGALG00000005547	361.69	235.57	1.53	2.77E-07	0.000162
chr6	19510800	19527199	-	-	241.36	162.56	1.48	9.74E-05	0.012633
chr6	19856200	19867999	-	ENSGALG00000006254	185.28	106.56	1.73	4.24E-06	0.001314
chr6	20438400	20448799	-	ENSGALG00000006384	96.73	19.38	4.80	1.45E-13	4.39E-10
chr6	22375200	22386999	BLNK	ENSGALG00000006973	250.71	162.80	1.54	1.70E-05	0.003902
chr6	24225200	24236599	-	ENSGALG00000007753	180.84	121.45	1.49	0.000801	0.044674
chr6	24987600	24994799	-	-	49.62	93.46	0.54	0.000278	0.041705
chr6	25713000	25725799	-	-	176.41	108.22	1.62	6.51E-05	0.009773
chr6	28162800	28172799	ACSL5	ENSGALG00000008840	96.26	52.22	1.83	0.000373	0.029344
chr6	29022200	29032599	DCLRE1A	ENSGALG00000008938	143.69	86.24	1.66	0.000201	0.020456
chr6	29163800	29176399	NHLRC2 -	ENSGALG00000008946 ENSGALG00000008971	117.76	68.52	1.71	0.000389	0.029701
chr7	4553400	4566599	UBE2F	ENSGALG00000003812	96.26	53.64	1.78	0.000536	0.03747
chr7	7149800	7156399	ITGB2	ENSGALG00000007511	125.00	71.59	1.74	0.00014	0.015755
chr7	8893600	8903599	STAT1	ENSGALG00000007651	107.95	40.88	2.60	3.04E-08	2.88E-05
chr7	12449200	12466199	CFLAR	ENSGALG00000008239	197.90	126.18	1.56	9.25E-05	0.012429
chr7	12500800	12511999	CASP8	ENSGALG00000008240 ENSGALG00000008355	184.12	116.72	1.57	0.000103	0.012915
chr7	15658000	15667999	UBE2E3	ENSGALG00000020793	211.92	145.31	1.46	0.000553	0.037928
chr7	18007200	18018999	-	-	78.74	41.59	1.87	0.000888	0.046703
chr7	22601800	22610999	IFIH1	ENSGALG00000011089	96.26	19.38	4.77	1.45E-13	4.39E-10
chr7	29843600	29847599	-	ENSGALG00000012072	38.55	9.69	3.70	2.49E-05	0.005479
chr7	32372600	32382799	-	-	125.24	74.67	1.67	0.000368	0.029344
chr7	34326800	34344399	ARHGAP15	ENSGALG00000012421	245.80	154.53	1.59	6.06E-06	0.00177
chr8	1933800	1944399	-	-	182.01	114.36	1.59	9.22E-05	0.012429
chr8	2083400	2096399	PTPRC	ENSGALG00000002192	214.96	141.06	1.52	0.000126	0.01485
chr8	3630200	3648599	-	-	198.60	123.58	1.60	3.36E-05	0.006531
chr8	4002200	4014599	-	ENSGALG00000021112	117.06	70.88	1.64	0.000726	0.042269
chr8	6853200	6863599	-	-	134.82	81.75	1.64	0.000367	0.029344
chr8	7374200	7384599	-	-	118.69	67.34	1.75	0.000218	0.021151
chr8	8331000	8338999	FAM129A	ENSGALG00000004812	102.81	58.83	1.73	0.000629	0.039475
chr8	10178800	10187799	C8H1orf27	ENSGALG00000005080	88.32	47.73	1.83	0.000527	0.03747
chr8	15813800	15822599	-	ENSGALG00000005105 -	145.56	89.31	1.62	0.000305	0.026604
chr8	25134000	25143599	ORC1	ENSGALG00000010623	169.86	103.49	1.64	7.52E-05	0.010671
chr8	25179000	25195199	GPX7	ENSGALG00000010627 ENSGALG00000010629	119.63	52.22	2.27	3.21E-07	0.000171
chr9	3527800	3547399	KLHL6	ENSGALG00000010633 ENSGALG00000002263	316.36	216.67	1.46	1.68E-05	0.003902
chr9	6137800	6151199	BOK	ENSGALG00000005772	270.57	155.71	1.73	2.66E-08	2.88E-05

chr9	11264000	11273399	-	-	209.35	137.75	1.52	0.000128	0.014869
chr9	25418800	25425399	-	-	64.25	30.48	2.07	0.000588	0.038465
chrZ	861200	868399	ACAA2	ENSGALG00000002793 ENSGALG00000002777	60.98	27.88	2.15	0.000524	0.03747
chrZ	961600	973599	TLX3	ENSGALG00000024979 ENSGALG00000002696 ENSGALG00000021762 ENSGALG00000024938 ENSGALG00000021750 ENSGALG00000018565	96.03	46.55	2.04	3.29E-05	0.006499
chrZ	1553800	1565799	PIAS2	ENSGALG00000001851 ENSGALG00000001843	84.58	39.93	2.09	6.10E-05	0.009352
chrZ	8422800	8427599	CD72	ENSGALG00000021353	94.63	41.59	2.25	5.87E-06	0.001759
chrZ	8786200	8794999	-	-	172.43	74.67	2.29	3.67E-10	6.97E-07
chrZ	9086200	9092599	-	-	111.68	61.91	1.79	0.00017	0.018263
chrZ	9915000	9927399	BRIX1	ENSGALG00000003365 ENSGALG00000003373 ENSGALG00000003387	110.98	49.38	2.22	1.47E-06	0.000558
chrZ	10322400	10330999	LMBRD2 SKP2	ENSGALG00000013377 ENSGALG00000003547	68.93	30.01	2.26	0.000156	0.017176
chrZ	10932400	10940199	WDR70	ENSGALG00000003688 ENSGALG00000003708	54.67	23.63	2.26	0.000539	0.037514
chrZ	13075800	13081399	HMGCS1	ENSGALG00000014862	78.27	34.02	2.26	3.91E-05	0.007239
chrZ	17652400	17658399	-	ENSGALG00000014727	52.57	20.56	2.49	0.000208	0.02088
chrZ	19978000	19985599	CENPK	ENSGALG00000014753 ENSGALG00000014756	73.13	34.73	2.07	0.000206	0.02088
chrZ	20015000	20023799	CZH5orf44	ENSGALG00000014765 ENSGALG00000014767 ENSGALG00000020567	75.47	38.99	1.91	0.000641	0.039537
chrZ	21011200	21014199	-	-	99.53	49.86	1.98	4.83E-05	0.008055
chrZ	22870600	22882999	F2RL1	ENSGALG00000014984	137.85	66.16	2.07	6.98E-07	0.000312
chrZ	23296000	23302799	-	-	71.50	35.92	1.96	0.000606	0.039317
chrZ	25974000	25977999	-	-	60.75	28.35	2.10	0.000846	0.04571
chrZ	26494000	26500999	CBWD1	ENSGALG00000010147	60.98	28.35	2.11	0.000846	0.04571
chrZ	26522600	26529799	-	ENSGALG00000010156	104.68	52.93	1.96	3.81E-05	0.007225
chrZ	27926200	27956199	-	ENSGALG00000023324	18.22	106.56	0.18	2.31E-16	1.17E-12
chrZ	27952200	27956199	-	-	10.05	89.08	0.12	5.52E-17	4.19E-13
chrZ	27962800	27966999	-	ENSGALG00000018479	21.03	3.74	4.65	0.000277	0.041705
chrZ	27968600	27973199	-	ENSGALG00000018479	38.75	5.61	6.02	2.50E-07	0.000252
chrZ	33393000	33402199	PLIN2	ENSGALG00000015090	80.38	38.51	2.06	0.000138	0.015659
chrZ	37218800	37226199	-	-	101.87	47.73	2.11	1.07E-05	0.002796
chrZ	37696800	37702599	-	-	42.99	16.30	2.54	0.000862	0.045898
chrZ	40796200	40807199	DAPK1	ENSGALG00000012608	173.84	84.35	2.05	2.97E-08	2.88E-05
chrZ	41465800	41473399	-	ENSGALG00000012621	75.00	34.73	2.13	0.000107	0.013225
chrZ	41665200	41674599	-	-	65.89	29.06	2.22	0.000262	0.023821
chrZ	42824600	42836199	-	ENSGALG00000010693 ENSGALG00000010694	171.03	92.15	1.85	1.27E-06	0.000507
chrZ	43327200	43335199	SYK	ENSGALG00000015216	153.74	77.03	1.98	6.07E-07	0.000288

chrZ	44067800	44077599	-	ENSGALG00000017686	125.47	56.94	2.18	3.14E-07	0.000171
chrZ	44165800	44173999	HINT1	ENSGALG00000000428	93.46	47.73	1.94	0.000126	0.01485
chrZ	45132000	45138399	-	-	69.39	31.90	2.14	0.000183	0.019172
chrZ	45245400	45250399	-	-	60.75	26.94	2.21	0.000317	0.026881
chrZ	55077800	55084199	LMNB1	ENSGALG00000014692	48.60	19.38	2.43	0.000522	0.03747
chrZ	56311400	56320199	ARSK	ENSGALG00000014672	78.27	34.50	2.23	3.91E-05	0.007239
chrZ	58562000	58569599	-	ENSGALG00000014670 ENSGALG00000014648	55.38	21.74	2.48	0.000121	0.014418
chrZ	59123800	59137999	-	ENSGALG00000014645	419.40	204.86	2.04	4.69E-18	7.12E-14
chrZ	61615400	61623199	TMEM167A	ENSGALG00000015619	57.95	26.70	2.13	0.000878	0.046554
chrZ	62751200	62760399	-	ENSGALG00000015576	64.49	30.48	2.08	0.000588	0.038465
chrZ	62771400	62778799	LOC425215	ENSGALG00000020534	69.86	29.06	2.36	6.57E-05	0.009773
chrZ	65032800	65037799	RAD17 SMC2 PTGR1	ENSGALG00000015571 ENSGALG00000015691	118.93	61.20	1.93	2.45E-05	0.005477
chrZ	66784000	66792199	-	ENSGALG00000001765	69.86	33.55	2.05	0.000469	0.034867
chrZ	67021800	67029999	-	ENSGALG00000001864	68.23	31.43	2.13	0.000255	0.023634
chrZ	69690600	69699599	-	-	83.18	42.53	1.93	0.00031	0.026638
chrZ	69701800	69704199	-	-	6.38	32.24	0.22	2.43E-05	0.00824
chrZ	72281400	72288199	-	ENSGALG00000008204	114.02	55.05	2.05	6.64E-06	0.001901
chrZ	73532600	73536399	-	ENSGALG00000005316	70.56	35.21	1.98	0.000821	0.045177
chrZ	74298400	74309999	-	-	85.75	36.62	2.31	9.81E-06	0.002614

Appendix IV. Sequencing results showing raw and mapped reads for from thymus samples.

Chicken Line	Status	Raw	Mapped	Mapped %	Non-redundant	Non-redundant %
6_3	Infected	13188253	10008826	75.892	4559032	0.455501175
	Control	11901057	10019594	84.1908	5615896	0.560491373
7_2	Infected	10716016	7417387	69.2178	2850746	0.384332919
	Control	6046819	4520471	74.7578	2365956	0.523387054
6_3	Infected	12496897	10152490	81.2401	7523315	0.74103151
	Control	9010579	7471261	82.9165	6836444	0.915032148
7_2	Infected	8754256	7082681	80.9056	5933170	0.837701147
	Control	7838478	5921978	75.5501	3603345	0.608469839

$Mapped\% = \frac{\#\ mapped\ reads * 100}{\#\ raw\ reads}$; $Non-redundant\% = \frac{\#\ non-redundant\ reads * 100}{\#\ mapped\ reads}$.

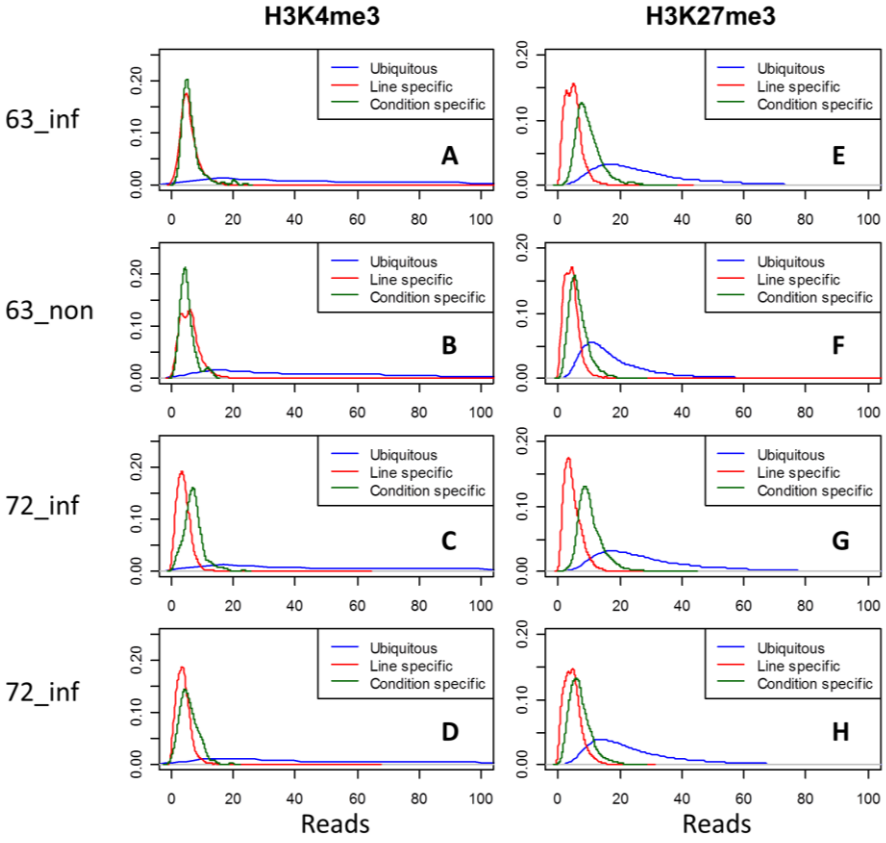
Appendix V. Primers used for quantitative PCR validation.

Genes	Purpose	Primers	Sequence
CTLA-4	Gene expression	F	5'- TCAAACAGACAGGCGACAAG-3'
		R	5'- GGGCTAACATGGCACTGAAT-3'
BCL6	Gene expression	F	5'- ATCCAGTTCACCCGCCACGC-3'
		R	5'- AGAGGCCACTGCAGGCCATCA-3'
CITED2	Gene expression	F	5'- CACGTCAGCCCGGGAGAGGA-3'
		R	5'- TTCCGCCATCTCCCACCTCCC-3'
EGR1	Gene expression	F	5'- AGCACCTTGCGGCAGACACTT-3'
		R	5'- GGAGAAGCGCCCCGTGTAGG-3'
TLR3	Gene expression	F	5'- CCATGGTGCAGGAAGTTTAAGGTGC-3'
		R	5'- CTGGCCAGTTCAAGATGCAGCA-3'
MX1	Gene expression	F	5'- TGGAGGAGCCAGCTGTTGCG-3'
		R	5'- ATTCTGGCCTGAGCAGCGTTGT-3'
MMP2	Gene expression	F	5'- GCTTTCTGCTTAGGCATTGG-3'
		R	5'- GCATTGGCATTTCATGTTTG-3'
IGF2BP1	Gene expression	F	5'- GCGTGACTCCGGCCGACTTG-3'
		R	5'- TGCAGCTCCACTTTCCCCGAA-3'
TNFSF1A	Gene expression	F	5'- CTGCGTCGCTGGCTTCTCTCC-3'
		R	5'- GTTAGGATAACCGTCCCCAGCGA-3'
GAL	Gene expression	F	5'- GCTCCCTGCGAGACACCGTT-3'
		R	5'- GGTTATCTACTGCATGTGGCCCAAG-3'
EAF2	Gene expression	F	5'- GCGGGCCATGGTGTGAGGTG-3'
		R	5'- AGTCATAGCGCACGGTGTGGAA-3'
HAPLN1	Gene expression	F	5'- GCGCATCTCGACTTGGGAGCT-3'
		R	5'- GGCGGGGTCCATTTTCTTCTTGGA-3'
CD4	Gene expression	F	5'- TGTCAACGCCGGATGTATAA -3'
		R	5'- CTTGTCCATTGGCTCCTCTC -3'
GAPDH	Gene expression	F	5'-GAGGGTAGTGAAGGCTGCTG-3'
		R	5'-ACCAGGAAACAAGCTTGACG-3'
GAPDH-ChIP	ChIP validation	F	5'-GTCACGTCCCAGGAGCAG-3'
		R	5'-AGGACCGTGCTAATGAGGAA-3'
MyoD-ChIP	ChIP validation	F	5'-TTGGTGGAGATCATGCCATA-3'
		R	5'-GTTGTGGGCCAGAAACAAGT-3'
K4-Peak-2	ChIP-Seq validation	F	5'-TCCTCCTTATGTGGGGAGTG-3'
		R	5'-GGACCTGTACTCGCAAGCTC-3'

F: forward, R: reverse

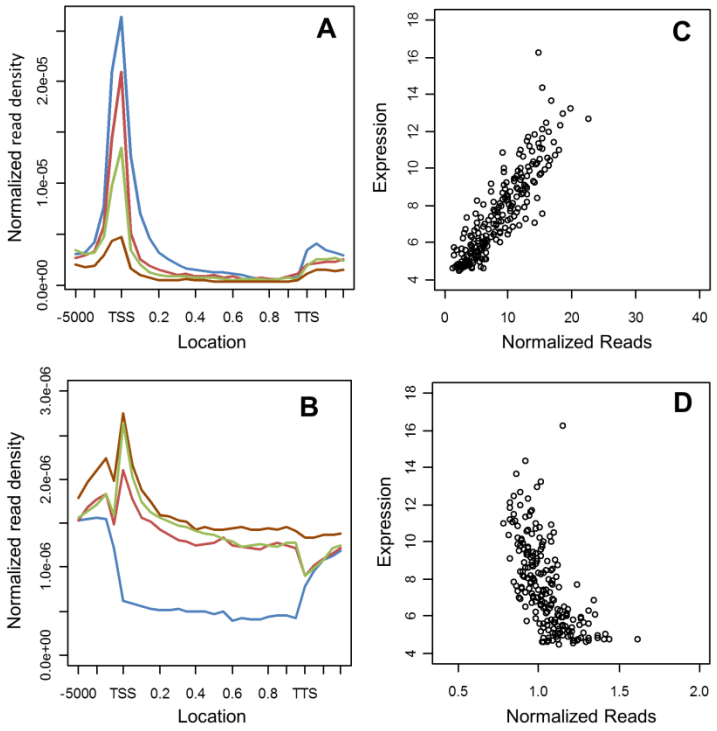
Appendix VI. Probability densities of peak length distributions in different classes of SERs.

Line-specific and condition-specific SERs predominantly correspond to low enrichment regions for both H3K4me3 (a-d) and H3K27me3 (e-h).
63_inf: line 6_3 infected, 63_non: line 6_3 control, 72_inf: line 7_2 infected, 72_non: line 7_2 control.



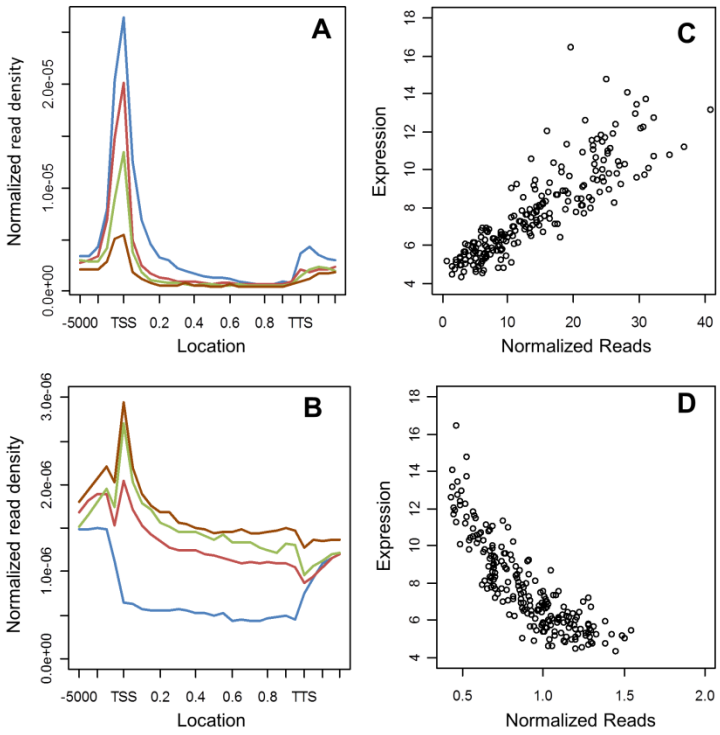
Appendix VII. Relationship between gene expression and histone marks in line 6_3 control samples.

Plots of histone modifications around the gene body (a & b) in genes having high (blue), medium (red), low (green) and no activity (brown). We also compared epigenetic marks with transcriptional levels: H3K4me3 shows positive correlation with gene expression levels (c) while H3K27me3 exhibits a negative relationship (d).



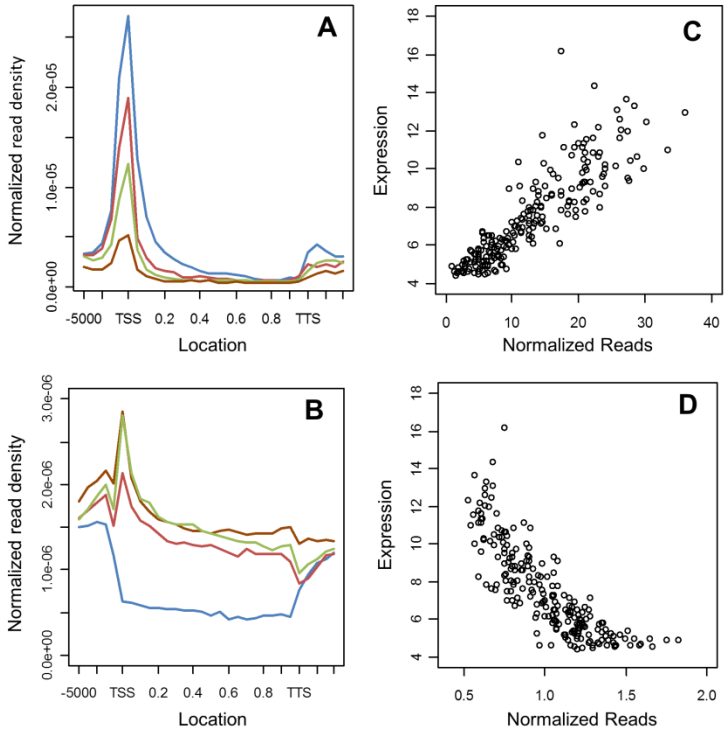
Appendix VIII. Relationship between gene expression and histone marks in line 7_2 control samples.

Plots of histone modifications around the gene body (a & b) in genes having high (blue), medium (red), low (green) and no activity (brown). We also compared epigenetic marks with transcriptional levels: H3K4me3 shows positive correlation with gene expression levels (c) while H3K27me3 exhibits a negative relationship (d).



Appendix IX. Relationship between gene expression and histone marks in line 7_2 infected samples.

Plots of histone modifications around the gene body (a & b) in genes having high (blue), medium (red), low (green) and no activity (brown). We also compared epigenetic marks with transcriptional levels: H3K4me3 shows positive correlation with gene expression levels (c) while H3K27me3 exhibits a negative relationship (d).



Appendix X. Differential H3K4me3 marks.

Genome-wide differential H3K4me3 marks produced by DESeq (FDR < 0.4) and associated genes. P-values from three contrasts are displayed as follows: 63: 63I vs 63N, 72: 72I vs 72N, 63v72N: 63N vs 72N.

SER	Samples	63	63v72N	72	Genes
chr1:51052800-51055599	72	0.716728	0.667995	3.82E-06	ENSGALG00000019325
chr1:170985000-170989599	72	0.002904	0.306603	4.55E-06	ENSGALG00000016964
chr1:53428200-53431799	72	0.052888	0.30538	0.000259	ENSGALG00000012472
chr5:25094000-25097799	72	0.449986	0.005768	0.000298	SPI1
chr7:14495200-14497999	72	0.170978	0.209627	0.000303	CTLA4
chr2:37739800-37741599	72	0.570832	0.021549	0.000589	ENSGALG00000011298
chr1:133186800-133192199	72	0.180816	0.034588	0.000633	P2RY8
chr20:6692200-6696799	72	0.012873	0.337854	0.000765	ENSGALG00000004859
chr6:20438000-20445599	72	0.000887	0.407654	0.000803	ENSGALG00000006384
chr1:112367400-112371799	63,72	1.28E-07	0.415314	4.26E-09	MX1
chrZ:27968600-27972599	63,72	3.41E-06	0.220991	1.41E-05	ENSGALG00000018479
chrZ:27964800-27967199	63,72	7.99E-05	0.982308	1.99E-06	ENSGALG00000018479
chr2:92042800-92046599	63,72	0.000149	0.469067	5.88E-05	ISG12-2
chr2:22865400-22868199	63,72	0.000232	0.015605	6.48E-09	ENSGALG00000009479
chr8:17511400-17515199	63v72N	0.479617	1.59E-15	0.664428	ENSGALG00000008854
chr27:3433400-3435399	63v72N	0.770916	4.21E-13	0.854824	IGF2BP1
chr19:8963800-8966399	63v72N	1	3.22E-12	0.757603	ENSGALG00000005578
chr20:8475000-8478399	63v72N	0.609497	5.00E-08	0.551674	BHLHB4
chr2:42132800-42135999	63v72N	0.885306	5.45E-08	0.11767	ENSGALG00000011613
chr2:125007000-125012399	63v72N	0.606371	1.12E-06	0.653623	ENSGALG00000015732
chr17:2454000-2459199	63v72N	0.257894	2.32E-06	0.969863	ENSGALG00000008472
chr15:11138400-11145599	63v72N	0.348842	3.37E-06	0.669509	ENSGALG00000007891
chr1:34419600-34421399	63v72N	0.326263	7.36E-06	0.0056	NM_205429
chr5:9146600-9149999	63v72N	0.574535	8.79E-06	1	ENSGALG00000005569
chr9:24010000-24011399	63v72N	0.691296	1.19E-05	1	ENSGALG00000009651
chr18:2129800-2131199	63v72N	0.065392	1.84E-05	0.782984	ENSGALG00000001261
chr1:58694400-58697399	63v72N	0.86674	2.08E-05	1	ENSGALG00000012842
chr6:20700600-20706799	63v72N	0.98773	3.56E-05	0.542148	ENSGALG00000006478
chr1:16635600-16638799	63v72N	0.34467	7.53E-05	0.967558	ENSGALG00000007025
chr4:88993400-88995799	63v72N	0.450676	0.000352	0.616238	RHACD8-4
chr2:31255200-31257599	63v72N	0.754165	0.00045	0.501096	ENSGALG00000010977
chr3:78289800-78291599	63v72N	0.400008	0.000464	0.868325	ENSGALG00000015768
chr4:89068600-89071199	63v72N	0.536311	0.000498	0.198913	ENSGALG00000015900
chr16:254000-260599	63v72N	0.167746	0.000532	0.727697	ENSGALG00000024350
chr28:2385800-2387599	63v72N	0.056255	0.000752	0.475928	HMHA1
chr9:24282200-24286799	63v72N	1	0.000792	0.194649	PTX3

chr5:21118200-21127599	63v72N	0.442807	0.000858	0.974084	ENSGALG00000007920
chr4:35196200-35199399	63v72N	0.668563	0.00089	0.08687	ENSGALG00000010119
chr1:86640200-86646399	63v72N	0.702909	0.001168	0.47306	ENSGALG00000015152
chr1:73596000-73597999	63v72N	0.292515	0.001353	0.354935	ENSGALG00000009702
chr1:493800-497599	63v72N	0.828115	0.001482	0.011776	ENSGALG00000013772
chr3:91008600-91012999	63v72N	0.914485	0.001541	0.222185	ENSGALG00000016313
chr28:950600-953199	63v72N	0.065569	0.001577	0.216139	LOC429451
chr11:3713800-3717199	63v72N	0.937602	0.001624	0.144532	MMP2
chr9:6356600-6361199	63v72N	0.286756	0.001771	0.380294	ENSGALG00000024277
chr11:18834600-18838599	63v72N	0.236689	0.001905	0.69083	ENSGALG00000021839
chr5:9901200-9904599	63v72N	0.285583	0.001913	0.59137	ENSGALG00000005662
chr24:5576000-5580199	63v72N	0.410471	0.00203	0.210206	ENSGALG00000024072
chr3:8004400-8011999	63v72N	0.245685	0.002035	0.364679	LBH
chr7:11293600-11296999	63v72N	0.899936	0.002137	0.02535	ENSGALG00000008118
chr5:60369200-60372999	63v72N	0.56105	0.002249	0.180948	ENSGALG00000012295
chr9:18230200-18232799	63v72N	0.203495	0.002359	0.183201	ENSGALG00000008852
chr16:287400-289799	63v72N	0.101651	0.002374	0.836484	YFV
chr6:20271600-20273799	63v72N	0.601004	0.002532	0.389374	ENSGALG00000006315
chr1:80088400-80090199	63v72N	0.036097	0.002575	0.50626	GAPDH
chr17:2197800-2203399	63v72N	0.203224	0.003068	0.72636	ENSGALG00000008623
chr11:540400-541799	63v72N	0.446134	0.003207	0.041007	ENSGALG00000001149
chr6:21718400-21721199	63v72N	0.5093	0.003402	0.723345	CYP26A1
chr1:80531600-80534199	63v72N	0.112136	0.003602	0.417485	ENSGALG00000014570
chr27:4011000-4012599	63v72N	0.027087	0.003641	0.880761	RPL19

63I: line 63 infected, 63N: line 63 control, 72I: line 72 infected, 72N: line 72 control.

Appendix XI. Differential H3K27me3 marks.

Genome-wide differential H3K27me3 marks produced by DESeq (FDR < 0.4) and the associated genes. P-values from three contrasts are displayed as follows: 63: 63I vs 63N, 72: 72I vs 72N, 63v72N: 63N vs 72N.

SER	Samples	63	63v72N	72	Genes
chrZ:61242600-61246599	63	0.000433	0.020859	0.950978	HAPLN1
chr1:200341000-200346199	63v72N	0.005362	8.03E-06	0.004025	PLEKHB1
chr20:13944200-13946199	63v72N	0.011424	0.000339	0.07724	ENSGALG00000021818
chr6:3937800-3940599	63v72N	0.086992	0.000453	0.024937	CHAT
chr1:56375000-56377399	63v72N	0.303852	0.000477	0.000864	SLC41A2
chr4:68468600-68471999	63v72N	0.316764	0.000736	0.015472	CNGA1
chr2:105030600-105041199	63v72N	0.005641	0.001004	0.599687	YES1
chr5:53554000-53556399	63v72N	0.141409	0.001085	0.705233	LOC396507
chrZ:14565400-14567599	63v72N	0.455076	0.001942	0.127439	ISL1
chr9:22131400-22134199	63v72N	0.062622	0.001942	0.013793	SERPINI1
chrZ:53399000-53401399	63v72N	0.453352	0.002075	0.039701	LPL
chr3:83227600-83228999	63v72N	0.298938	0.002459	0.054133	MYO6
chr4:61633000-61634799	63v72N	0.136925	0.002736	0.016587	MTTP
chr15:1026600-1029599	63v72N	0.062622	0.002745	0.173465	TXNRD2
chr17:10782800-10784999	63v72N	0.015864	0.00291	0.604932	LMX1B
chr1:34734400-34739999	63v72N	0.170909	0.003188	0.049861	USP15
chr2:142382800-142388799	63v72N	0.036544	0.003623	0.241009	COL14A1
chr7:38258000-38260599	63v72N	0.635254	0.003987	0.005603	BAZ2B
chr12:13025200-13035799	63v72N	0.248088	0.004195	0.359076	PTPRG
chr1:86479600-86480599	63v72N	0.385484	0.004399	1	CLDND1
chr2:9614800-9616999	63v72N	0.24546	0.004417	0.132706	VIPR2
chr20:13529400-13532599	63v72N	0.346345	0.004507	0.048222	PARD6B
chr8:10056400-10060999	63v72N	0.120295	0.004617	0.611896	PTGS2
chrZ:73849400-73853199	63v72N	0.725354	0.005114	0.001516	SNX2
chr8:25401600-25405399	63v72N	0.058472	0.005189	0.34897	LRP8
chr3:83220400-83222599	63v72N	0.298938	0.005312	0.138533	MYO6
chr12:20438000-20440399	63v72N	0.302634	0.00538	0.063607	ENSGALG00000008546
chr8:10061600-10063599	63v72N	0.919869	0.005598	0.091557	PTGS2
chr20:13611000-13614599	63v72N	0.156188	0.005903	0.04699	PTPN1
chr3:107835800-107838599	63v72N	0.038119	0.006081	1	SELI
chr14:3565400-3566399	63v72N	0.661808	0.006367	0.076777	SDK1
chr19:9934600-9937199	63v72N	0.012778	0.006736	0.684263	ENSGALG00000024472
chr2:19019200-19020999	63v72N	0.657886	0.006888	0.156392	ARL5B
chr3:98841000-98849999	63v72N	0.037394	0.006998	0.200231	MBOAT2
chrZ:31223800-31228399	63v72N	0.539413	0.007079	0.040181	NFIB
chr3:83237600-83239199	63v72N	0.01338	0.007223	0.012756	MYO6

chr5:32322800-32325799	63v72N	0.138907	0.007372	0.157171	LOC423287
chr1:136063000-136065199	63v72N	1	0.007507	0.122353	CNGA3
chr1:119394000-119397399	63v72N	0.338779	0.007685	0.84737	NR0B1
chr1:82674400-82681599	63v72N	0.021767	0.007708	0.839856	TNFRSF1A
chr13:12341200-12342999	63v72N	0.236329	0.008054	0.184991	MFAP3
chr1:82682800-82684999	63v72N	0.09838	0.009211	0.557042	TNFRSF1A
chr4:12203800-12206399	63v72N	0.285549	0.00941	0.072009	CDX4
chr11:3259800-3261399	63v72N	0.221089	0.00946	0.119613	RBM35B
chr15:3199600-3202399	63v72N	0.413584	0.009922	0.151783	STX2
chr5:17802200-17813599	63v72N	0.07643	0.010175	0.161937	GAL,GAL
chr2:142391600-142393199	63v72N	0.924871	0.01037	0.124309	COL14A1
chr7:13565400-13567999	63v72N	0.326274	0.01084	0.087589	ADAM23
chr3:98829000-98834999	63v72N	0.065783	0.01114	0.067188	MBOAT2
chr21:2663200-2673399	63v72N	0.819859	0.011482	0.032016	LOC419429
chr2:32833000-32834599	63v72N	0.426118	0.011758	0.176766	HIBADH
chr20:13521000-13522799	63v72N	0.657886	0.011758	0.065858	PARD6B
chr10:464600-471199	63v72N	0.275583	0.011908	0.056682	RBPMS2
chr9:6139600-6145199	63v72N	0.660708	0.012047	0.012202	BOK
chr9:25551600-25554399	63v72N	0.086893	0.012453	0.023952	ENSGALG00000023481
chr9:22134600-22157599	63v72N	0.105332	0.012688	0.351075	PDCD10
chr10:12022000-12024199	63v72N	0.30003	0.01291	0.54086	GALK2
chr11:10590400-10593399	63v72N	0.108908	0.013312	0.370375	NUDT19
chr23:2790000-2791599	63v72N	0.270433	0.013363	0.000515	PTPRU
chr3:57077400-57080199	63v72N	0.083218	0.013476	0.173465	MAP7
chr15:2961000-2962399	63v72N	0.183799	0.013643	0.08126	STX2
chr2:119649400-119658799	63v72N	0.14125	0.013823	0.172635	COPS5
chr7:27897400-27900599	63v72N	0.041375	0.013832	0.528546	EAF2
chr1:199701000-199703999	63v72N	0.417345	0.014276	0.024195	MAP6
chr1:200229400-200232599	63v72N	0.084398	0.015005	0.817811	PLEKHB1
chr10:9020600-9022599	63v72N	0.773433	0.01506	0.077735	PRTG
chr2:34385400-34391599	63v72N	0.379072	0.015335	0.053537	DAZL
chr14:8924600-8928999	63v72N	0.224468	0.015342	0.566168	CDR2
chr2:32821200-32824599	63v72N	0.136263	0.015538	0.160668	HIBADH
chr19:9764800-9766799	63v72N	0.837869	0.015928	0.066012	ENSGALG00000005995
chr24:932800-935599	63v72N	0.189546	0.016521	0.765579	TNIP1
chr7:12495800-12498999	63v72N	0.369589	0.017077	0.599676	CASP18
chr1:174433000-174438599	63v72N	0.335686	0.017193	0.755401	WDFY2
chr3:81190800-81193199	63v72N	0.473563	0.017221	0.28765	LOC421845
chr15:2944400-2947999	63v72N	0.497322	0.017263	0.033	STX2
chr3:11203000-11210399	63v72N	0.098602	0.017496	0.181967	PPP3R1

63I: line 63 infected, 63N: line 63 control, 72I: line 72 infected, 72N: line 72 control.

Appendix XII. Putative bivalent genes from colocalization analysis of H3K4me3 and H3K27me3.

Gene	Alternative Name	Samples
CITED2	ENSGALG00000013818	L63_inf,L72_inf
BCL6	ENSGALG00000007357	L63_inf,L72_inf,L72_non
EGR1	ENSGALG00000007669	L63_inf,L72_inf,L72_non
TLR3	ENSGALG00000013468	L63_inf,L72_inf,L72_non
ST6GAL1	ENSGALG00000005550	L63_non,L72_inf
TIRAP	ENSGALG00000001077	L72_inf
NECAP2	ENSGALG00000003745	L72_inf
UBB	ENSGALG00000004509	L72_inf
SMAD3	ENSGALG00000007870	L72_inf
ANXA5	ENSGALG00000011885	L72_inf
GCH1	ENSGALG00000012200	L72_inf
YFV	ENSGALG00000024344	L72_inf
LOC417083	ENSGALG00000024350	L72_inf
LOC378902	ENSGALG00000006407	L72_inf,L72_non
ST3GAL6	ENSGALG00000015252	L72_inf,L72_non
RHOB	ENSGALG00000016485	L72_inf,L72_non
CD4	ENSGALG00000014477	L72_non
PLS1	ENSGALG00000002647	L63_inf
ENSGALG00000008952	ENSGALG00000008952	L63_inf
RAB33B	ENSGALG00000009790	L63_inf
C9orf18	ENSGALG00000001352	L63_inf,L63_non,L72_inf,L72_non
ENSGALG00000003545	ENSGALG00000003545	L63_inf,L63_non,L72_inf,L72_non
PFN2	ENSGALG00000010410	L63_inf,L63_non,L72_inf,L72_non
PLEKHA8	ENSGALG00000011185	L63_inf,L63_non,L72_inf,L72_non
ENSGALG00000011364	ENSGALG00000011364	L63_inf,L63_non,L72_inf,L72_non
SIX1	ENSGALG00000022994	L63_inf,L63_non,L72_inf,L72_non
BTBD14A	ENSGALG00000001728	L63_inf,L72_inf
STK10	ENSGALG00000002816	L63_inf,L72_inf
LOC768803	ENSGALG00000003048	L63_inf,L72_inf
CDC25A	ENSGALG00000004934	L63_inf,L72_inf
RAP1GAP2	ENSGALG00000005868	L63_inf,L72_inf
REEP6	ENSGALG00000015189	L63_inf,L72_inf
ENSGALG00000020995	ENSGALG00000020995	L63_inf,L72_inf
ENSGALG00000023324	ENSGALG00000023324	L63_inf,L72_inf
RAB33A	ENSGALG00000024049	L63_inf,L72_inf
LOC419892	ENSGALG00000002568	L63_inf,L72_inf,L72_non
AGPHD1	ENSGALG00000003063	L63_inf,L72_inf,L72_non
ENSGALG00000003598	ENSGALG00000003598	L63_inf,L72_inf,L72_non
SOX30	ENSGALG00000003723	L63_inf,L72_inf,L72_non

ENSGALG00000004884	ENSGALG00000004884	L63_inf,L72_inf,L72_non
IL13RA1	ENSGALG00000006032	L63_inf,L72_inf,L72_non
GLT8D4	ENSGALG00000007804	L63_inf,L72_inf,L72_non
ENSGALG00000007909	ENSGALG00000007909	L63_inf,L72_inf,L72_non
C10orf26	ENSGALG00000008119	L63_inf,L72_inf,L72_non
ENSGALG00000009816	ENSGALG00000009816	L63_inf,L72_inf,L72_non
SPP1	ENSGALG00000010926	L63_inf,L72_inf,L72_non
PRRG4	ENSGALG00000012032	L63_inf,L72_inf,L72_non
C11orf54	ENSGALG00000017219	L63_inf,L72_inf,L72_non
C1orf190	ENSGALG00000017379	L63_inf,L72_inf,L72_non
LOC768635	ENSGALG00000019568	L63_inf,L72_inf,L72_non
ENSGALG00000023347	ENSGALG00000023347	L63_inf,L72_inf,L72_non
MACROD2	ENSGALG00000023773	L63_inf,L72_inf,L72_non
ENSGALG00000023864	ENSGALG00000023864	L63_inf,L72_inf,L72_non
PDE8A	ENSGALG00000005992	L63_inf,L72_non
RAB3B	ENSGALG00000010567	L63_inf,L72_non
MIB1	ENSGALG00000014974	L63_inf,L72_non
PLEKHF2	ENSGALG00000015988	L63_inf,L72_non
TAF12	ENSGALG00000000991	L63_non
RPLP1	ENSGALG00000016172	L63_non
BATF	ENSGALG00000010323	L63_non,L72_inf
ZDHHC18	ENSGALG00000000869	L72_inf
ORAI2	ENSGALG00000001837	L72_inf
TRIM65	ENSGALG00000002209	L72_inf
GSTT1	ENSGALG00000005204	L72_inf
GFI1	ENSGALG00000005940	L72_inf
SLC24A6	ENSGALG00000008337	L72_inf
KCNMB4	ENSGALG00000010044	L72_inf
CYP46A1	ENSGALG00000011162	L72_inf
MYC	ENSGALG00000016308	L72_inf
ENSGALG00000020271	ENSGALG00000020271	L72_inf
ENSGALG00000022653	ENSGALG00000022653	L72_inf
TPCN3		L72_inf
RAB40B	ENSGALG00000001545	L72_inf,L72_non
SPTAN1	ENSGALG00000004719	L72_inf,L72_non
GSTT1	ENSGALG00000006344	L72_inf,L72_non
CCDC40	ENSGALG00000007042	L72_inf,L72_non
NELF	ENSGALG00000008681	L72_inf,L72_non
C14orf174	ENSGALG00000010457	L72_inf,L72_non
CSTB	ENSGALG00000014410	L72_inf,L72_non
LOC421845	ENSGALG00000015865	L72_inf,L72_non
CRYL1	ENSGALG00000017135	L72_inf,L72_non
TMEM22	ENSGALG00000001285	L72_non

DYDC1	ENSGALG00000002432	L72_non
ACSBG1	ENSGALG00000003286	L72_non
PTGS2	ENSGALG00000005069	L72_non
DHRS11	ENSGALG00000005403	L72_non
C16orf45	ENSGALG00000006456	L72_non
C22orf36	ENSGALG00000006588	L72_non
CCDC104	ENSGALG00000008064	L72_non
ENSGALG00000010412	ENSGALG00000010412	L72_non
TPMT	ENSGALG00000012687	L72_non
ENSGALG00000014777	ENSGALG00000014777	L72_non
SNX3	ENSGALG00000015304	L72_non
CCDC125	ENSGALG00000015572	L72_non
TP53I3	ENSGALG00000016502	L72_non
N6AMT2	ENSGALG00000017133	L72_non
ENSGALG00000021811	ENSGALG00000021811	L72_non
ENSGALG00000024306	ENSGALG00000024306	L72_non
SLMAP		L72_non

63_inf: line 6_3 infected, 63_non: line 6_3 control, 72_inf: line 7_2 infected, 72_non: line 7_2 control.

Appendix XIII. Functional annotation clustering of bivalent genes.

Top 5 functional annotation clusters from enrichment analysis of putative bivalent genes using DAVID. P-values were generated by the program and FDR calculated using the Benjamini-Hochberg procedure.

Annotation Cluster 1 Enrichment Score: 2.1845463098819007

Term	Count	P-Value	FDR
GO:0006955~immune response	7	1.98E-04	0.110468
GO:0045087~innate immune response	3	0.012582	0.656019
GO:0006952~defense response	3	0.111974	0.834128

Annotation Cluster 2 Enrichment Score: 1.6487663600383535

Term	Count	P-Value	FDR
GO:0046649~lymphocyte activation	4	0.007963	0.692509
GO:0045321~leukocyte activation	4	0.011364	0.674961
GO:0001775~cell activation	4	0.016505	0.706945
GO:0030097~hemopoiesis	4	0.019778	0.657496
GO:0048534~hemopoietic or lymphoid organ development	4	0.026651	0.735031
GO:0010604~positive regulation of macromolecule metabolic process	6	0.028689	0.733148
GO:0002520~immune system development	4	0.030149	0.700043
GO:0030098~lymphocyte differentiation	3	0.030369	0.679294
GO:0042110~T cell activation	3	0.033032	0.688318
GO:0002521~leukocyte differentiation	3	0.047625	0.729812

Annotation Cluster 3 Enrichment Score: 1.6398579175050156

Term	Count	P-Value	FDR
GO:0010033~response to organic substance	5	0.00998	0.693819
GO:0010604~positive regulation of macromolecule metabolic process	6	0.028689	0.733148
GO:0044093~positive regulation of molecular function	4	0.042032	0.718255

Annotation Cluster 4 Enrichment Score: 1.1692334713765171

Term	Count	P-Value	FDR
GO:0070085~glycosylation	3	0.046075	0.734266
GO:0006486~protein amino acid glycosylation	3	0.046075	0.734266
GO:0043413~biopolymer glycosylation	3	0.046075	0.734266
GO:0009101~glycoprotein biosynthetic process	3	0.055656	0.727326
GO:0009100~glycoprotein metabolic process	3	0.067634	0.736268
GO:0031090~organelle membrane	5	0.123709	0.999998
GO:0012505~endomembrane system	4	0.143513	0.993977

Annotation Cluster 5 Enrichment Score: 1.1590029316201944

Term	Count	P-Value	FDR
GO:0010604~positive regulation of macromolecule metabolic process	6	0.028689	0.733148
GO:0009967~positive regulation of signal transduction	3	0.103793	0.825775
GO:0010647~positive regulation of cell communication	3	0.111974	0.834128

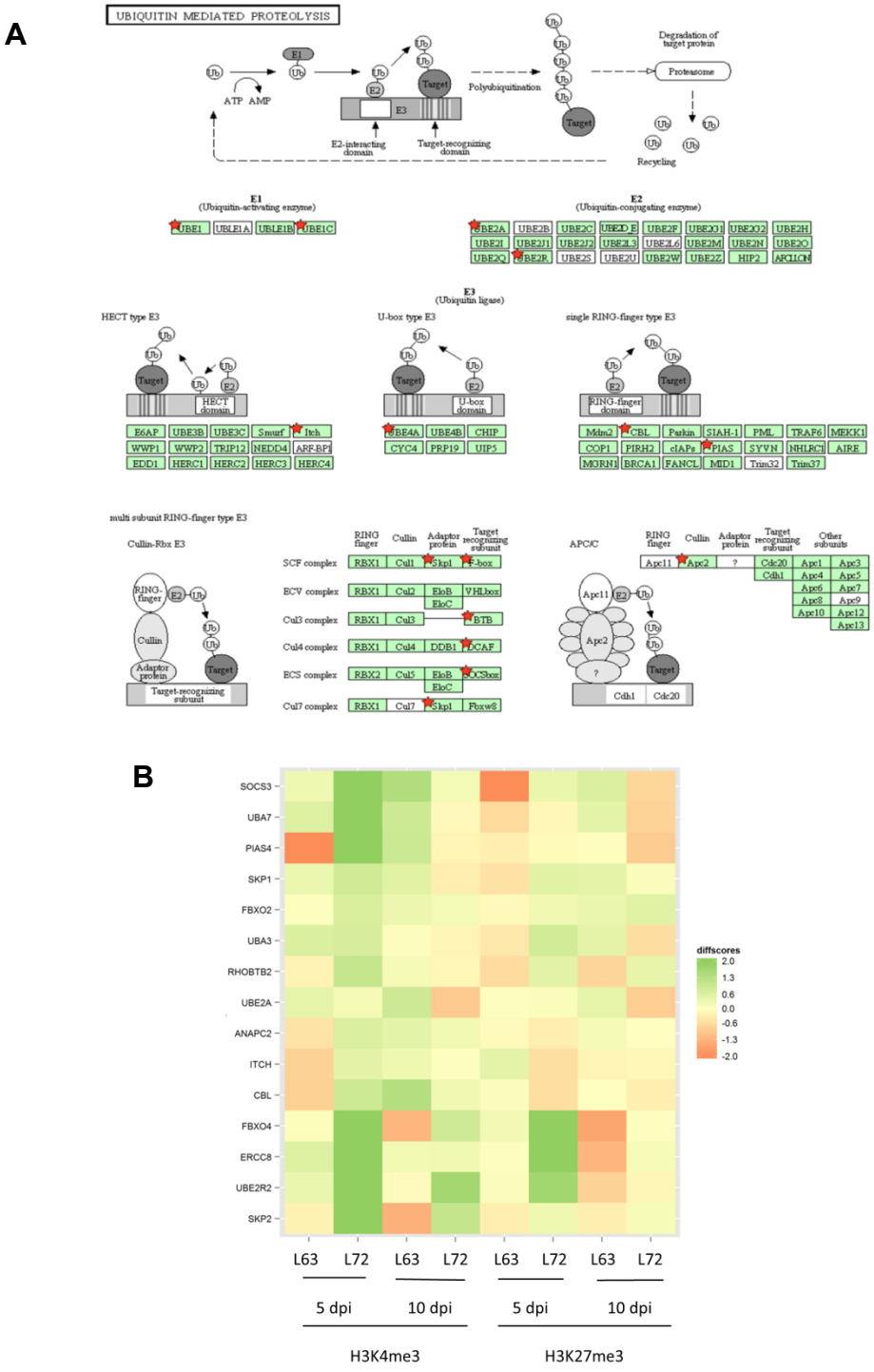
Appendix XIV. Sequencing results showing read numbers for each sample from bursa of Fabricius at 5 and 10 days post infection.

Histone	DPI	Line	Status	Replicate	Total	Mapped	Mapped%	Non-redundant	Non-redundant%
H3K4me3	5	$L6_3$	inf	1	10146707	9469948	93.33026	5990070	63.25346
				2	13131449	12144366	92.48306	7392050	60.86814
			non	1	16445821	15134536	92.02664	9744869	64.38829
				2	17765727	16502268	92.88822	12311597	74.60548
		$L7_2$	inf	1	13685625	12397340	90.58658	8567829	69.11022
				2	15693208	14242909	90.75843	9231076	64.81173
			non	1	22800222	11537587	50.60296	7102076	61.55599
				2	23472442	12579614	53.59312	8094915	64.34947
	10	$L6_3$	inf	1	11207008	10454772	93.28781	7216101	69.02208
				2	8743345	8174684	93.49607	5686370	69.56073
			non	1	16180287	15023018	92.84766	11294506	75.18134
				2	13562130	12506700	92.21782	9863369	78.86468
		$L7_2$	inf	1	17597960	15956904	90.67474	10980332	68.81242
				2	15109789	13756612	91.04437	10963779	79.69825
			non	1	20238612	1872581	9.252517	985452	52.62533
				2	21964876	11188450	50.93792	8784442	78.51348
H3K27me3	5	$L6_3$	inf	1	13033297	12535093	96.17745	9701319	77.39328
				2	9233419	8823724	95.56291	7086270	80.30929
			non	1	14458859	13576043	93.89429	11482089	84.57611
				2	12628694	11812297	93.53538	10014981	84.78436
		$L7_2$	inf	1	17480503	16641372	95.19962	14810698	88.99926
				2	15159625	14573824	96.13578	12974760	89.02784
			non	1	22208836	11979948	53.94226	10340280	86.31323
				2	24457902	17323581	70.8302	14925120	86.15494
	10	$L6_3$	inf	1	8915797	8521946	95.58255	6299382	73.91952
				2	10798509	10335407	95.71143	7510955	72.67208
			non	1	13943995	13132816	94.18259	11627720	88.53943
				2	11968746	11235976	93.87764	9362469	83.32582
		$L7_2$	inf	1	15202340	14301081	94.07158	12719372	88.93993
				2	13530768	12757146	94.2825	11426112	89.56637
			non	1	24191568	9603288	39.69684	5565871	57.95797
				2	21476908	11605789	54.03845	10150591	87.46145

$Mapped\% = \frac{\#\ mapped\ reads * 100}{\#\ raw\ reads}$; $Non-redundant\% = \frac{\#\ non-redundant\ reads * 100}{\#\ mapped\ reads}$.

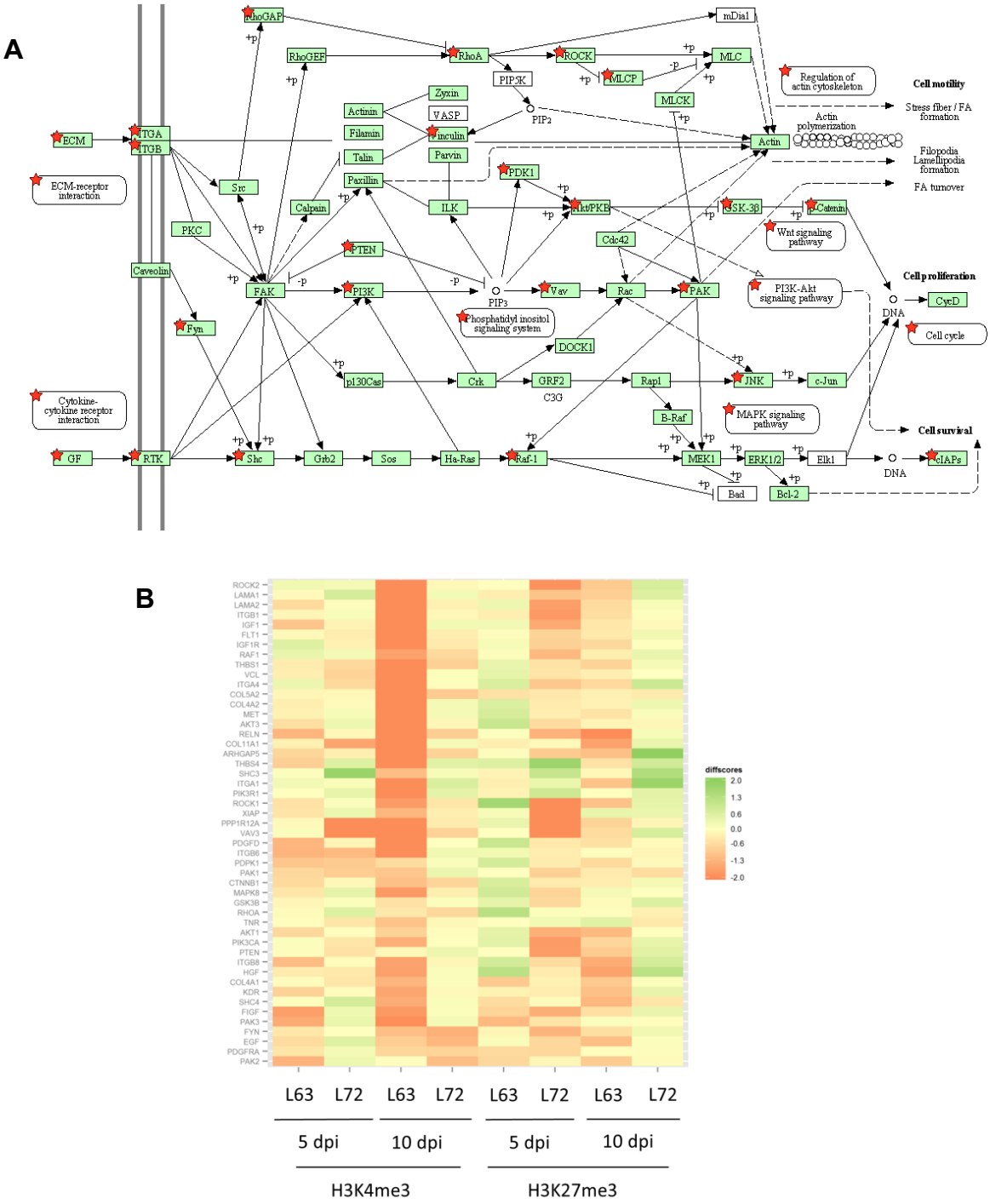
Appendix XV. Ubiquitin-mediated proteolysis pathway displays increased H3K4me3 marks in line $L7_2$ at 5 dpi.

(a) KEGG pathway map and (b) diffscore clustering heatmap. All classes of enzymes, E1, E2 and E3, exhibit increased promoter H3K4me3 indicating up-regulation.



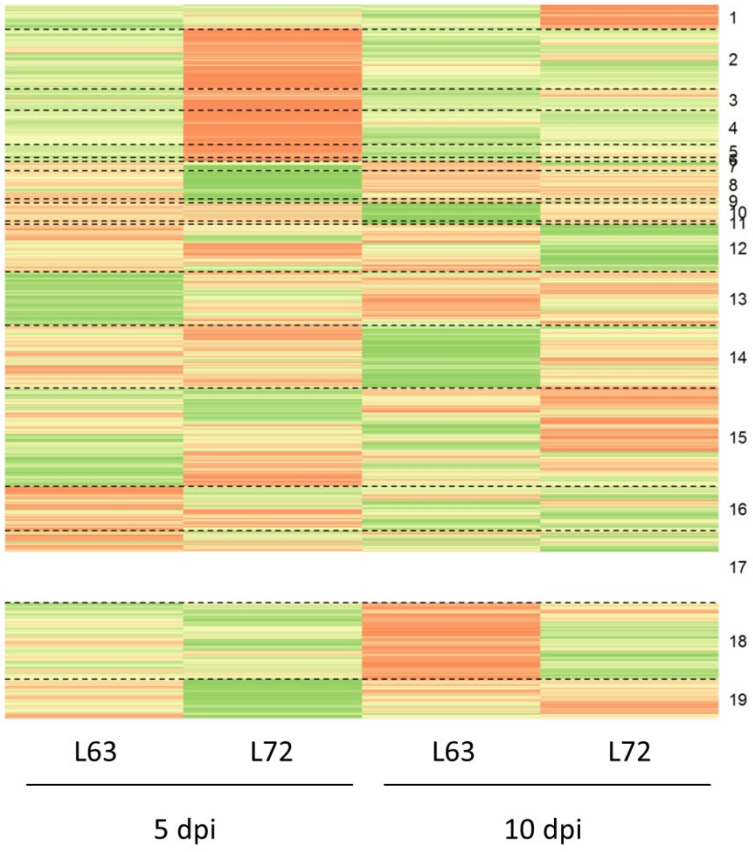
Appendix XVI. Focal adhesion pathway displays reduced H3K4me3 marks in line $L6_3$ at 10 dpi.

(a) KEGG pathway map and (b) diffscore clustering heatmap. Several members of the focal adhesion pathway demonstrate reductions in promoter H3K4me3 in resistant birds during latent infection.



Appendix XVII. Hierarchical clustering of diffscores from differential analysis of RNA-Seq data from Bursa.

Heatmap of clustered diffscores with green denoting upregulation and red representing downregulation after MDV infection. Blank rows in cluster 17 correspond to genes with no mapped reads.



Bibliography

1. Allfrey VG, Faulkner R, Mirsky AE: **Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis**. *Proc Natl Acad Sci U S A* 1964, **51**:786-794.
2. Esteller M: **Aberrant DNA methylation as a cancer-inducing mechanism**. *Annu Rev Pharmacol Toxicol* 2005, **45**:629-656.
3. Herman JG, Baylin SB: **Gene silencing in cancer in association with promoter hypermethylation**. *N Engl J Med* 2003, **349**(21):2042-2054.
4. Nguyen CT, Gonzales FA, Jones PA: **Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation**. *Nucleic Acids Res* 2001, **29**(22):4598-4606.
5. Fahrner JA, Eguchi S, Herman JG, Baylin SB: **Dependence of histone modifications and gene expression on DNA hypermethylation in cancer**. *Cancer Res* 2002, **62**(24):7213-7218.
6. Seligson DB, Horvath S, Shi T, Yu H, Tze S, Grunstein M, Kurdistani SK: **Global histone modification patterns predict risk of prostate cancer recurrence**. *Nature* 2005, **435**(7046):1262-1266.
7. Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, Bonaldi T, Haydon C, Ropero S, Petrie K *et al*: **Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer**. *Nat Genet* 2005, **37**(4):391-400.
8. Kurdistani SK: **Histone modifications in cancer biology and prognosis**. *Prog Drug Res* 2011, **67**:91-106.
9. Knipe DM, Lieberman PM, Jung JU, McBride AA, Morris KV, Ott M, Margolis D, Nieto A, Nevels M, Parks RJ *et al*: **Snapshots: chromatin control of viral infection**. *Virology* 2013, **435**(1):141-156.
10. Wysocka J, Myers MP, Laherty CD, Eisenman RN, Herr W: **Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1**. *Genes Dev* 2003, **17**(7):896-911.
11. Peng H, Nogueira ML, Vogel JL, Kristie TM: **Transcriptional coactivator HCF-1 couples the histone chaperone Asf1b to HSV-1 DNA replication components**. *Proc Natl Acad Sci U S A* 2010, **107**(6):2461-2466.
12. Gunther T, Grundhoff A: **The epigenetic landscape of latent Kaposi sarcoma-associated herpesvirus genomes**. *PLoS Pathog* 2010, **6**(6):e1000935.
13. Clapier CR, Cairns BR: **The biology of chromatin remodeling complexes**. *Annu Rev Biochem* 2009, **78**:273-304.
14. Tamaru H: **Confining euchromatin/heterochromatin territory: jumonji crosses the line**. *Genes Dev* 2010, **24**(14):1465-1478.

15. Herceg Z, Murr R: **Chapter 3 - Mechanisms of Histone Modifications**. In: *Handbook of Epigenetics*. Edited by Trygve T. San Diego: Academic Press; 2011: 25-45.
16. Kouzarides T: **Chromatin modifications and their function**. *Cell* 2007, **128**(4):693-705.
17. Orford K, Kharchenko P, Lai W, Dao MC, Worhunsky DJ, Ferro A, Janzen V, Park PJ, Scadden DT: **Differential H3K4 methylation identifies developmentally poised hematopoietic genes**. *Dev Cell* 2008, **14**(5):798-809.
18. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J: **Differential chromatin marking of introns and expressed exons by H3K36me3**. *Nat Genet* 2009, **41**(3):376-381.
19. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y: **Role of histone H3 lysine 27 methylation in Polycomb-group silencing**. *Science* 2002, **298**(5595):1039-1043.
20. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K *et al*: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells**. *Cell* 2006, **125**(2):315-326.
21. Allan RS, Zueva E, Cammas F, Schreiber HA, Masson V, Belz GT, Roche D, Maison C, Quivy JP, Almouzni G *et al*: **An epigenetic silencing pathway controlling T helper 2 cell lineage commitment**. *Nature* 2012, **487**(7406):249-253.
22. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA *et al*: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome**. *Nat Genet* 2007, **39**(3):311-318.
23. Dillon SC, Zhang X, Trievel RC, Cheng X: **The SET-domain protein superfamily: protein lysine methyltransferases**. *Genome Biol* 2005, **6**(8):227.
24. Jenuwein T, Laible G, Dorn R, Reuter G: **SET domain proteins modulate chromatin domains in eu- and heterochromatin**. *Cell Mol Life Sci* 1998, **54**(1):80-93.
25. Nguyen AT, Zhang Y: **The diverse functions of Dot1 and H3K79 methylation**. *Genes Dev* 2011, **25**(13):1345-1358.
26. Bedford MT, Richard S: **Arginine methylation an emerging regulator of protein function**. *Mol Cell* 2005, **18**(3):263-272.
27. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, Casero RA: **Histone demethylation mediated by the nuclear amine oxidase homolog LSD1**. *Cell* 2004, **119**(7):941-953.
28. Metzger E, Wissmann M, Yin N, Muller JM, Schneider R, Peters AH, Gunther T, Buettner R, Schule R: **LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription**. *Nature* 2005, **437**(7057):436-439.

29. Cloos PA, Christensen J, Agger K, Helin K: **Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease**. *Genes Dev* 2008, **22**(9):1115-1140.
30. Couture JF, Collazo E, Ortiz-Tello PA, Brunzelle JS, Trievel RC: **Specificity and mechanism of JMJD2A, a trimethyllysine-specific histone demethylase**. *Nat Struct Mol Biol* 2007, **14**(8):689-695.
31. Chang B, Chen Y, Zhao Y, Bruick RK: **JMJD6 is a histone arginine demethylase**. *Science* 2007, **318**(5849):444-447.
32. Zhang Y: **Transcriptional regulation by histone ubiquitination and deubiquitination**. *Genes Dev* 2003, **17**(22):2733-2740.
33. Pickart CM: **Mechanisms underlying ubiquitination**. *Annu Rev Biochem* 2001, **70**:503-533.
34. Hickey CM, Wilson NR, Hochstrasser M: **Function and regulation of SUMO proteases**. *Nat Rev Mol Cell Biol* 2012, **13**(12):755-766.
35. Shiio Y, Eisenman RN: **Histone sumoylation is associated with transcriptional repression**. *Proc Natl Acad Sci U S A* 2003, **100**(23):13225-13230.
36. Berger SL: **The complex language of chromatin regulation during transcription**. *Nature* 2007, **447**(7143):407-412.
37. Sims RJ, 3rd, Reinberg D: **Histone H3 Lys 4 methylation: caught in a bind?** *Genes Dev* 2006, **20**(20):2779-2786.
38. Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P *et al*: **A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling**. *Nature* 2006, **442**(7098):86-90.
39. Martin DG, Baetz K, Shi X, Walter KL, MacDonald VE, Wlodarski MJ, Gozani O, Hieter P, Howe L: **The Yng1p plant homeodomain finger is a methyl-histone binding module that recognizes lysine 4-methylated histone H3**. *Mol Cell Biol* 2006, **26**(21):7871-7879.
40. Shi X, Hong T, Walter KL, Ewalt M, Michishita E, Hung T, Carney D, Pena P, Lan F, Kaadige MR *et al*: **ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression**. *Nature* 2006, **442**(7098):96-99.
41. Huang Y, Fang J, Bedford MT, Zhang Y, Xu RM: **Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A**. *Science* 2006, **312**(5774):748-751.
42. Muller J, Kassis JA: **Polycomb response elements and targeting of Polycomb group proteins in Drosophila**. *Curr Opin Genet Dev* 2006, **16**(5):476-484.
43. Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y: **Role of histone H3 lysine 27 methylation in X inactivation**. *Science* 2003, **300**(5616):131-135.
44. Martinez AM, Cavalli G: **The role of polycomb group proteins in cell cycle regulation during development**. *Cell Cycle* 2006, **5**(11):1189-1197.
45. Sparmann A, van Lohuizen M: **Polycomb silencers control cell fate, development and cancer**. *Nat Rev Cancer* 2006, **6**(11):846-856.

46. Simon JA, Kingston RE: **Mechanisms of polycomb gene silencing: knowns and unknowns**. *Nat Rev Mol Cell Biol* 2009, **10**(10):697-708.
47. Ringrose L, Paro R: **Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins**. *Annu Rev Genet* 2004, **38**:413-443.
48. Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G: **Genome regulation by polycomb and trithorax proteins**. *Cell* 2007, **128**(4):735-745.
49. Carey MF, Peterson CL, Smale ST: **Chromatin immunoprecipitation (ChIP)**. *Cold Spring Harb Protoc* 2009, **2009**(9):pdb prot5279.
50. Gilmour DS, Lis JT: **Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes**. *Proc Natl Acad Sci U S A* 1984, **81**(14):4275-4279.
51. Gilmour DS, Lis JT: **In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster**. *Mol Cell Biol* 1985, **5**(8):2009-2018.
52. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR *et al*: **Genomic maps and comparative analysis of histone modifications in human and mouse**. *Cell* 2005, **120**(2):169-181.
53. Boyd KE, Farnham PJ: **Myc versus USF: discrimination at the cad gene is determined by core promoter elements**. *Mol Cell Biol* 1997, **17**(5):2529-2537.
54. Parekh BS, Maniatis T: **Virus infection leads to localized hyperacetylation of histones H3 and H4 at the IFN-beta promoter**. *Mol Cell* 1999, **3**(1):125-129.
55. Bernstein BE, Humphrey EL, Liu CL, Schreiber SL: **The use of chromatin immunoprecipitation assays in genome-wide analyses of histone modifications**. *Methods Enzymol* 2004, **376**:349-360.
56. Sharov V, Kwong KY, Frank B, Chen E, Hasseman J, Gaspard R, Yu Y, Yang I, Quackenbush J: **The limits of log-ratios**. *BMC Biotechnol* 2004, **4**:3.
57. **Sequencing-by-Synthesis: Explaining the Illumina Sequencing Technology** [http://nxseq.bitesizebio.com/articles/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/]
58. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW: **Transforming clinical microbiology with bacterial genome sequencing**. *Nat Rev Genet* 2012, **13**(9):601-612.
59. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nat Rev Genet* 2011, **12**(2):87-98.
60. Martin JA, Wang Z: **Next-generation transcriptome assembly**. *Nat Rev Genet* 2011, **12**(10):671-682.
61. Alkan C, Coe BP, Eichler EE: **Genome structural variation discovery and genotyping**. *Nat Rev Genet* 2011, **12**(5):363-376.
62. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional organization of mammalian genomes**. *Nat Rev Genet* 2011, **12**(1):7-18.
63. Laird PW: **Principles and challenges of genome-wide DNA methylation analysis**. *Nat Rev Genet* 2010, **11**(3):191-203.

64. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K: **Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data**. *Nucleic Acids Res* 2008, **36**(16):5221-5231.
65. Cho I, Blaser MJ: **The human microbiome: at the interface of health and disease**. *Nat Rev Genet* 2012, **13**(4):260-270.
66. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome**. *Nat Rev Genet* 2012, **13**(1):47-58.
67. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome**. *Cell* 2007, **129**(4):823-837.
68. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions**. *Science* 2007, **316**(5830):1497-1502.
69. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP *et al*: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**. *Nature* 2007, **448**(7153):553-560.
70. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A *et al*: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nat Methods* 2007, **4**(8):651-657.
71. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology**. *Nat Rev Genet* 2009, **10**(10):669-680.
72. Schwartz S, Oren R, Ast G: **Detection and removal of biases in the analysis of next-generation sequencing reads**. *PLoS One* 2011, **6**(1):e16685.
73. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Res* 2008, **36**(16):e105.
74. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing**. *Nucleic Acids Res* 2012, **40**(10):e72.
75. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M: **Mapping accessible chromatin regions using Sono-Seq**. *Proc Natl Acad Sci U S A* 2009, **106**(35):14926-14931.
76. Vega VB, Cheung E, Palanisamy N, Sung WK: **Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries**. *PLoS One* 2009, **4**(4):e5241.
77. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M: **Modeling ChIP sequencing in silico with applications**. *PLoS Comput Biol* 2008, **4**(8):e1000158.
78. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Res* 2008, **18**(11):1851-1858.
79. Smith AD, Xuan Z, Zhang MQ: **Using quality scores and longer reads improves accuracy of Solexa read mapping**. *BMC Bioinformatics* 2008, **9**:128.

80. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
81. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
82. Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing**. *PLoS One* 2009, **4**(11):e7767.
83. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
84. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing**. *Brief Bioinform* 2010, **11**(5):473-483.
85. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
86. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
87. Ma B, Tromp J, Li M: **PatternHunter: faster and more sensitive homology search**. *Bioinformatics* 2002, **18**(3):440-445.
88. Ge N, Sen Z, Wai Hong C: **Linear Suffix Array Construction by Almost Pure Induced-Sorting**. In: *Data Compression Conference, 2009 DCC '09: 16-18 March 2009 2009*; 2009: 193-202.
89. Ferragina P, Manzini G: **Opportunistic data structures with applications**. In: *Proceedings of the 41st Annual Symposium on Foundations of Computer Science.* IEEE Computer Society; 2000: 390.
90. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.
91. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
92. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**(9):R137.
93. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ: **FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology**. *Bioinformatics* 2008, **24**(15):1729-1730.
94. Narlikar L, Jothi R: **ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder**. *Methods Mol Biol* 2012, **802**:305-322.
95. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls**. *Nat Biotechnol* 2009, **27**(1):66-75.
96. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data**. *Nat Biotechnol* 2008, **26**(11):1293-1300.

97. Boyle AP, Guinney J, Crawford GE, Furey TS: **F-Seq: a feature density estimator for high-throughput sequence tags**. *Bioinformatics* 2008, **24**(21):2537-2538.
98. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data**. *Nat Methods* 2008, **5**(9):829-834.
99. Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, Chinnaiyan AM: **HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data**. *BMC Bioinformatics* 2010, **11**:369.
100. Song Q, Smith AD: **Identifying dispersed epigenomic domains from ChIP-Seq data**. *Bioinformatics* 2011, **27**(6):870-871.
101. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: **A clustering approach for identification of enriched domains from histone modification ChIP-Seq data**. *Bioinformatics* 2009, **25**(15):1952-1958.
102. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD: **ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions**. *Genome Biol* 2011, **12**(7):R67.
103. Spyrou C, Stark R, Lynch AG, Tavare S: **BayesPeak: Bayesian analysis of ChIP-seq data**. *BMC Bioinformatics* 2009, **10**:299.
104. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL: **A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments**. *BMC Genomics* 2009, **10**:618.
105. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-seq peak detection**. *PLoS One* 2010, **5**(7):e11471.
106. Szalkowski AM, Schmid CD: **Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts**. *Brief Bioinform* 2011, **12**(6):626-633.
107. Nix DA, Courdy SJ, Boucher KM: **Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks**. *BMC Bioinformatics* 2008, **9**:523.
108. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance**. *Bioinformatics* 2007, **23**(21):2881-2887.
109. Ismail N, Jemain AA: **Handling overdispersion with negative binomial and generalized poisson regression models**. *Casualty Actuarial Society Forum* 2007, **Winter**:103-158.
110. Agresti A: **Categorical Data Analysis**, 3 edn: John Wiley & Sons, 2002; 2013.
111. Kharchenko PV, Tolstorukov MY, Park PJ: **Design and analysis of ChIP-seq experiments for DNA-binding proteins**. *Nat Biotechnol* 2008, **26**(12):1351-1359.
112. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS**. *Nat Protoc* 2012, **7**(9):1728-1740.
113. Albert I, Wachi S, Jiang C, Pugh BF: **GeneTrack--a genomic data processing and visualization framework**. *Bioinformatics* 2008, **24**(10):1305-1306.

114. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition**. *Proceedings of the IEEE* 1989, **77**(2):257-286.
115. Viterbi AJ: **Error bounds for convolutional codes and an asymptotically optimum decoding algorithm**. *Information Theory, IEEE Transactions on* 1967, **13**(2):260-269.
116. Cairns J, Spyrou C, Stark R, Smith ML, Lynch AG, Tavare S: **BayesPeak--an R package for analysing ChIP-seq data**. *Bioinformatics* 2011, **27**(5):713-714.
117. Xu H, Wei CL, Lin F, Sung WK: **An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data**. *Bioinformatics* 2008, **24**(20):2344-2349.
118. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.
119. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**(10):R106.
120. **U.S. Chicken Industry History** [http://www.nationalchickencouncil.org/about-the-industry/history/]
121. Churchill AE, Payne LN, Chubb RC: **Immunization against Marek's disease using a live attenuated virus**. *Nature* 1969, **221**(5182):744-747.
122. Rispens BH, van Vloten H, Mastenbroek N, Maas JL, Schat KA: **Control of Marek's disease in the Netherlands. II. Field trials on vaccination with an avirulent strain (CVI 988) of Marek's disease virus**. *Avian Dis* 1972, **16**(1):126-138.
123. Osterrieder N, Kamil JP, Schumacher D, Tischer BK, Trapp S: **Marek's disease virus: from miasma to model**. *Nat Rev Microbiol* 2006, **4**(4):283-294.
124. Asmundson VS, Biely J: **INHERITANCE OF RESISTANCE TO FOWL PARALYSIS (NEUROLYMPHOMATOSIS GALLINARUM): I. DIFFERENCES IN SUSCEPTIBILITY**. *Canadian Journal of Research* 1932, **6**(2):171-176.
125. Hutt FB, Cole RK: **Genetic Control of Lymphomatosis in the Fowl**. *Science* 1947, **106**(2756):379-384.
126. Burgess SC, Davison TF: **Identification of the neoplastically transformed cells in Marek's disease herpesvirus-induced lymphomas: recognition by the monoclonal antibody AV37**. *J Virol* 2002, **76**(14):7276-7292.
127. Burgess SC, Young JR, Baaten BJ, Hunt L, Ross LN, Parcells MS, Kumar PM, Tregaskes CA, Lee LF, Davison TF: **Marek's disease is a natural model for lymphomas overexpressing Hodgkin's disease antigen (CD30)**. *Proc Natl Acad Sci U S A* 2004, **101**(38):13879-13884.
128. Jeurissen SH, Janse EM, Kok GL, De Boer GF: **Distribution and function of non-lymphoid cells positive for monoclonal antibody CVI-ChNL-68.2 in healthy chickens and those infected with Marek's disease virus**. *Vet Immunol Immunopathol* 1989, **22**(2):123-133.

129. Baigent SJ, Ross LJ, Davison TF: **Differential susceptibility to Marek's disease is associated with differences in number, but not phenotype or location, of pp38+ lymphocytes**. *J Gen Virol* 1998, **79 (Pt 11)**:2795-2802.
130. Calnek BW, Schat KA, Ross LJ, Shek WR, Chen CL: **Further characterization of Marek's disease virus-infected lymphocytes. I. In vivo infection**. *Int J Cancer* 1984, **33**(3):389-398.
131. Xing Z, Schat KA: **Expression of cytokine genes in Marek's disease virus-infected chickens and chicken embryo fibroblast cultures**. *Immunology* 2000, **100**(1):70-76.
132. Cantello JL, Anderson AS, Morgan RW: **Identification of latency-associated transcripts that map antisense to the ICP4 homolog gene of Marek's disease virus**. *J Virol* 1994, **68**(10):6280-6290.
133. Parcells MS, Arumugaswami V, Prigge JT, Pandya K, Dienglewicz RL: **Marek's disease virus reactivation from latency: changes in gene expression at the origin of replication**. *Poult Sci* 2003, **82**(6):893-898.
134. Calnek BW: **Marek's disease--a model for herpesvirus oncology**. *Crit Rev Microbiol* 1986, **12**(4):293-320.
135. Schat KA: **Role of the spleen in the pathogenesis of Marek's disease**. *Avian Pathol* 1981, **10**(2):171-182.
136. Baigent SJ, Davison TF: **Development and composition of lymphoid lesions in the spleens of Marek"s disease virus-infected chickens: association with virus spread and the pathogenesis of Marek"s disease**. *Avian Pathology* 1999, **28**(3):287-300.
137. Rouse BT, Wells RJ, Warner NL: **Proportion of T and B lymphocytes in lesions of Marek's disease: theoretical implications for pathogenesis**. *J Immunol* 1973, **110**(2):534-539.
138. Burgess SC, Basaran BH, Davison TF: **Resistance to Marek's disease herpesvirus-induced lymphoma is multiphasic and dependent on host genotype**. *Vet Pathol* 2001, **38**(2):129-142.
139. Haffer K, Sevoian M, Wilder M: **The role of the macrophages in Marek's disease: in vitro and in vivo studies**. *Int J Cancer* 1979, **23**(5):648-656.
140. Barrow AD, Burgess SC, Baigent SJ, Howes K, Nair VK: **Infection of macrophages by a lymphotropic herpesvirus: a new tropism for Marek's disease virus**. *J Gen Virol* 2003, **84**(Pt 10):2635-2645.
141. Djeraba A, Musset E, Bernardet N, Le Vern Y, Quere P: **Similar pattern of iNOS expression, NO production and cytokine response in genetic and vaccination-acquired resistance to Marek's disease**. *Vet Immunol Immunopathol* 2002, **85**(1-2):63-75.
142. Gupta MK, Chauhan HV, Jha GJ, Singh KK: **The role of the reticuloendothelial system in the immunopathology of Marek's disease**. *Vet Microbiol* 1989, **20**(3):223-234.
143. Sharma JM, Coulson BD: **Presence of natural killer cells in specific-pathogen-free chickens**. *J Natl Cancer Inst* 1979, **63**(2):527-531.
144. Sharma JM, Okazaki W: **Natural killer cell activity in chickens: target cell analysis and effect of antithymocyte serum on effector cells**. *Infect Immun* 1981, **31**(3):1078-1085.

145. Garcia-Camacho L, Schat KA, Brooks R, Jr., Bounous DI: **Early cell-mediated immune responses to Marek's disease virus in two chicken lines with defined major histocompatibility complex antigens**. *Vet Immunol Immunopathol* 2003, **95**(3-4):145-153.
146. Bumstead N: **Genomic mapping of resistance to Marek's disease**. *Avian Pathology* 1998, **27**(sup1):S78-S81.
147. Neulen ML, Gobel TW: **Chicken CD56 defines NK cell subsets in embryonic spleen and lung**. *Dev Comp Immunol* 2012, **38**(3):410-415.
148. Omar AR, Schat KA: **Syngeneic Marek's disease virus (MDV)-specific cell-mediated immune responses against immediate early, late, and unique MDV proteins**. *Virology* 1996, **222**(1):87-99.
149. Liu JL, Lin SF, Xia L, Brunovskis P, Li D, Davidson I, Lee LF, Kung HJ: **MEQ and V-IL8: cellular genes in disguise?** *Acta Virol* 1999, **43**(2-3):94-101.
150. Hansen MP, van Zandt JN, Law GRJ: **Differences in susceptibility to Marek's disease in chickens carrying two different B blood group alleles**. *Poult Sci* 1967, **46**:1268.
151. Pazderka F, Longenecker B, Law GJ, Stone H, Ruth R: **Histocompatibility of chicken populations selected for resistance to Marek's disease**. *Immunogenetics* 1975, **2**(1):93-100.
152. Omar AR, Schat KA: **Characterization of Marek's disease herpesvirus-specific cytotoxic T lymphocytes in chickens inoculated with a non-oncogenic vaccine strain of MDV**. *Immunology* 1997, **90**(4):579-585.
153. Kaufman J, Volk H, Wallny HJ: **A "minimal essential Mhc" and an "unrecognized Mhc": two extremes in selection for polymorphism**. *Immunol Rev* 1995, **143**:63-88.
154. Bumstead N, Sillibourne J, Rennie M, Ross N, Davison F: **Quantification of Marek's disease virus in chicken lymphocytes using the polymerase chain reaction with fluorescence detection**. *J Virol Methods* 1997, **65**(1):75-81.
155. Fredericksen T, Longenecker B, Pazderka F, Gilmour D, Ruth R: **A T-cell antigen system of chickens: Ly-4 and Marek's disease**. *Immunogenetics* 1977, **5**(1):535-552.
156. Gilmour D, Brand A, Donnelly N, Stone H: **Bu-1 andTh-1, two loci determining surface antigens of B or T lymphocytes in the chicken**. *Immunogenetics* 1976, **3**(1):549-563.
157. Vallejo RL, Pharr GT, Liu HC, Cheng HH, Witter RL, Bacon LD: **Non-association between Rfp-Y major histocompatibility complex-like genes and susceptibility to Marek's disease virus-induced tumours in 6(3) x 7(2) F2 intercross chickens**. *Anim Genet* 1997, **28**(5):331-337.
158. Yonash N, Bacon LD, Witter RL, Cheng HH: **High resolution mapping and identification of new quantitative trait loci (QTL) affecting susceptibility to Marek's disease**. *Anim Genet* 1999, **30**(2):126-135.
159. Liu HC, Cheng HH, Tirunagaru V, Sofer L, Burnside J: **A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping**. *Animal Genetics* 2001, **32**(6):351-359.

160. Kaiser P, Underwood G, Davison F: **Differential cytokine responses following Marek's disease virus infection of chickens differing in resistance to Marek's disease**. *J Virol* 2003, **77**(1):762-768.
161. Abdul-Careem MF, Hunter BD, Parvizi P, Haghighi HR, Thanthrige-Don N, Sharif S: **Cytokine gene expression patterns associated with immunization against Marek's disease in chickens**. *Vaccine* 2007, **25**(3):424-432.
162. Smith J, Sadeyen JR, Paton IR, Hocking PM, Salmon N, Fife M, Nair V, Burt DW, Kaiser P: **Systems analysis of immune responses in Marek's disease virus-infected chickens identifies a gene involved in susceptibility and highlights a possible novel pathogenicity mechanism**. *J Virol* 2011, **85**(21):11146-11158.
163. Kumar S, Kunec D, Buza JJ, Chiang HI, Zhou H, Subramaniam S, Pendarvis K, Cheng HH, Burgess SC: **Nuclear Factor kappa B is central to Marek's disease herpesvirus induced neoplastic transformation of CD30 expressing lymphocytes in-vivo**. *BMC Syst Biol* 2012, **6**:123.
164. Mitra A, Liu G, Song J: **A genome-wide analysis of array-based comparative genomic hybridization (CGH) data to detect intra-species variations and evolutionary relationships**. *PLoS One* 2009, **4**(11):e7978.
165. Team RDC: **R: A Language and Environment for Statistical Computing**; 2008.
166. Feng J, Liu T, Zhang Y: **Using MACS to Identify Peaks from ChIP-Seq Data**. In: *Current Protocols in Bioinformatics.* vol. 34: John Wiley & Sons, Inc.; 2011: 2.14.11-12.14.14.
167. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A *et al*: **The UCSC Genome Browser database: update 2011**. *Nucleic Acids Res* 2010, **39**(Database issue):D876-882.
168. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A *et al*: **Ensembl BioMarts: a hub for data retrieval across taxonomic space**. *Database (Oxford)* 2011, **2011**:bar030.
169. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2011**. *Nucleic Acids Res* 2011, **39**(Database issue):D800-806.
170. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
171. Torrence C, Compo G: **A Practical Guide to Wavelet Analysis**. *Bulletin of the American Meteorological Society* 1998, **79**(1):61-78.
172. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.
173. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function**. *Genome Res* 2003, **13**(9):2129-2141.

174. Caldwell RB, Kierzek AM, Arakawa H, Bezzubov Y, Zaim J, Fiedler P, Kutter S, Blagodatski A, Kostovska D, Koter M *et al*: **Full-length cDNAs from chicken bursal lymphocytes to facilitate gene function analysis**. *Genome Biol* 2005, **6**(1):R6.
175. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome**. *Nat Biotechnol* 2010, **28**(8):817-825.
176. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al*: **Mapping and analysis of chromatin state dynamics in nine human cell types**. *Nature* 2011, **473**(7345):43-49.
177. Weishaupt H, Sigvardsson M, Attema JL: **Epigenetic chromatin states uniquely define the developmental plasticity of murine hematopoietic stem cells**. *Blood* 2010, **115**(2):247-256.
178. Flanagan JM: **Host epigenetic modifications by oncogenic viruses**. *Br J Cancer* 2007, **96**(2):183-188.
179. Knight JS, Lan K, Subramanian C, Robertson ES: **Epstein-Barr virus nuclear antigen 3C recruits histone deacetylase activity and associates with the corepressors mSin3A and NCoR in human B-cell lines**. *J Virol* 2003, **77**(7):4261-4272.
180. Tsai CN, Tsai CL, Tse KP, Chang HY, Chang YS: **The Epstein-Barr virus oncogene product, latent membrane protein 1, induces the downregulation of E-cadherin gene expression via activation of DNA methyltransferases**. *Proc Natl Acad Sci U S A* 2002, **99**(15):10084-10089.
181. Epstein MA, Achong BG, Churchill AE, Biggs PM: **Structure and development of the herpes-types virus of Marek's disease**. *J Natl Cancer Inst* 1968, **41**(3):805-820.
182. Okazaki W, Purchase HG, Burmester BR: **Protection against Marek's disease by vaccination with a herpesvirus of turkeys**. *Avian Dis* 1970, **14**(2):413-429.
183. Davison F, Nair V: **Use of Marek's disease vaccines: could they be driving the virus to increasing virulence?** *Expert Rev Vaccines* 2005, **4**(1):77-88.
184. Gimeno IM: **Marek's disease vaccines: a solution for today but a worry for tomorrow?** *Vaccine* 2008, **26 Suppl 3**:C31-41.
185. Witter RL: **Increased virulence of Marek's disease virus field isolates**. *Avian Dis* 1997, **41**(1):149-163.
186. Liu HC, Cheng HH, Tirunagaru V, Sofer L, Burnside J: **A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping**. *Anim Genet* 2001, **32**(6):351-359.
187. Liu HC, Kung HJ, Fulton JE, Morgan RW, Cheng HH: **Growth hormone interacts with the Marek's disease virus SORF2 protein and is associated with disease resistance in chicken**. *Proc Natl Acad Sci U S A* 2001, **98**(16):9203-9208.

188. Wain HM, Toye AA, Hughes S, Bumstead N: **Targeting of marker loci to chicken chromosome 16 by representational difference analysis**. *Anim Genet* 1998, **29**(6):446-452.
189. Yunis R, Jarosinski KW, Schat KA: **Association between rate of viral genome replication and virulence of Marek's disease herpesvirus strains**. *Virology* 2004, **328**(1):142-150.
190. Zehner ZE, Paterson BM: **Characterization of the chicken vimentin gene: single copy gene producing multiple mRNAs**. *Proc Natl Acad Sci U S A* 1983, **80**(4):911-915.
191. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome**. *Cell* 2007, **129**(4):823-837.
192. Smith J, Sadeyen JR, Paton IR, Hocking PM, Salmon N, Fife M, Nair V, Burt DW, Kaiser P: **Systems Analysis of Immune Responses in Marek's Disease Virus Infected Chickens Identifies a Gene Involved in Susceptibility and Highlights a Possible Novel Pathogenicity Mechanism**. *J Virol* 2011, **85**(21):11146-11158.
193. McCoy KD, Le Gros G: **The role of CTLA-4 in the regulation of T cell immune responses**. *Immunol Cell Biol* 1999, **77**(1):1-10.
194. Curran MA, Montalvo W, Yagita H, Allison JP: **PD-1 and CTLA-4 combination blockade expands infiltrating T cells and reduces regulatory T and myeloid cells within B16 melanoma tumors**. *Proc Natl Acad Sci U S A* 2010, **107**(9):4275-4280.
195. Heidari M, Sarson AJ, Huebner M, Sharif S, Kireev D, Zhou H: **Marek's disease virus-induced immunosuppression: array analysis of chicken immune response gene expression profiling**. *Viral Immunol* 2010, **23**(3):309-319.
196. Fingleton B: **Matrix metalloproteinases: roles in cancer and metastasis**. *Front Biosci* 2006, **11**:479-491.
197. Rath NC, Parcells MS, Xie H, Santin E: **Characterization of a spontaneously transformed chicken mononuclear cell line**. *Vet Immunol Immunopathol* 2003, **96**(1-2):93-104.
198. Chen C, Li H, Xie Q, Shang H, Ji J, Bai S, Cao Y, Ma Y, Bi Y: **Transcriptional profiling of host gene expression in chicken liver tissues infected with oncogenic Marek's disease virus**. *J Gen Virol* 2011, **92**:2724-2733.
199. Ioannidis P, Kottaridi C, Dimitriadis E, Courtis N, Mahaira L, Talieri M, Giannopoulos A, Iliadis K, Papaioannou D, Nasioulas G *et al*: **Expression of the RNA-binding protein CRD-BP in brain and non-small cell lung tumors**. *Cancer Lett* 2004, **209**(2):245-250.
200. Ioannidis P, Mahaira L, Papadopoulou A, Teixeira MR, Heim S, Andersen JA, Evangelou E, Dafni U, Pandis N, Trangas T: **8q24 Copy number gains and expression of the c-myc mRNA stabilizing protein CRD-BP in primary breast carcinomas**. *Int J Cancer* 2003, **104**(1):54-59.
201. Kato T, Hayama S, Yamabuki T, Ishikawa N, Miyamoto M, Ito T, Tsuchiya E, Kondo S, Nakamura Y, Daigo Y: **Increased expression of insulin-like**

growth factor-II messenger RNA-binding protein 1 is associated with tumor progression in patients with lung cancer. *Clin Cancer Res* 2007, **13**(2 Pt 1):434-442.
202. Ross J, Lemm I, Berberet B: **Overexpression of an mRNA-binding protein in human colorectal cancer**. *Oncogene* 2001, **20**(45):6544-6550.
203. Berger A, Santic R, Hauser-Kronberger C, Schilling FH, Kogner P, Ratschek M, Gamper A, Jones N, Sperl W, Kofler B: **Galanin and galanin receptors in human cancers**. *Neuropeptides* 2005, **39**(3):353-359.
204. Rauch I, Kofler B: **The galanin system in cancer**. *Exs* 2010, **102**:223-241.
205. Sanz LA, Chamberlain S, Sabourin JC, Henckel A, Magnuson T, Hugnot JP, Feil R, Arnaud P: **A mono-allelic bivalent chromatin domain controls tissue-specific imprinting at Grb10**. *Embo J* 2008, **27**(19):2523-2532.
206. Rodriguez J, Munoz M, Vives L, Frangou CG, Groudine M, Peinado MA: **Bivalent domains enforce transcriptional memory of DNA methylated genes in cancer cells**. *Proc Natl Acad Sci U S A* 2008, **105**(50):19809-19814.
207. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**(1):1-13.
208. Chan HM, La Thangue NB: **p300/CBP proteins: HATs for transcriptional bridges and scaffolds**. *J Cell Sci* 2001, **114**(Pt 13):2363-2373.
209. Thiel G, Cibelli G: **Regulation of life and death by the zinc finger transcription factor Egr-1**. *J Cell Physiol* 2002, **193**(3):287-292.
210. Scharnhorst V, Menke AL, Attema J, Haneveld JK, Riteco N, van Steenbrugge GJ, van der Eb AJ, Jochemsen AG: **EGR-1 enhances tumor growth and modulates the effect of the Wilms' tumor 1 gene products on tumorigenicity**. *Oncogene* 2000, **19**(6):791-800.
211. Huang RP, Liu C, Fan Y, Mercola D, Adamson ED: **Egr-1 negatively regulates human tumor cell growth via the DNA-binding domain**. *Cancer Res* 1995, **55**(21):5054-5062.
212. Buscaglia C, Calnek BW: **Maintenance of Marek's disease herpesvirus latency in vitro by a factor found in conditioned medium**. *J Gen Virol* 1988, **69 (Pt 11)**:2809-2818.
213. Deng X, Li X, Shen Y, Qiu Y, Shi Z, Shao D, Jin Y, Chen H, Ding C, Li L *et al*: **The Meq oncoprotein of Marek's disease virus interacts with p53 and inhibits its transcriptional and apoptotic activities**. *Virol J* 2010, **7**:348.
214. Wu Z-Z, Sun N-K, Chao CCK: **Knockdown of CITED2 using short-hairpin RNA sensitizes cancer cells to cisplatin through stabilization of p53 and enhancement of p53-dependent apoptosis**. *Journal of Cellular Physiology* 2011, **226**(9):2415-2428.
215. Phan RT, Dalla-Favera R: **The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells**. *Nature* 2004, **432**(7017):635-639.
216. Schat KA, Chen CL, Calnek BW, Char D: **Transformation of T-lymphocyte subsets by Marek's disease herpesvirus**. *J Virol* 1991, **65**(3):1408-1413.
217. Yu Y, Luo J, Mitra A, Chang S, Tian F, Zhang H, Yuan P, Zhou H, Song J: **Temporal Transcriptome Changes Induced by MDV in Marek's Disease-**

Resistant and -Susceptible Inbred Chickens. *BMC Genomics* 2011, **12**(1):501.
218. Morimura T, Ohashi K, Kon Y, Hattori M, Sugimoto C, Onuma M: **Apoptosis and CD8-down-regulation in the thymus of chickens infected with Marek's disease virus**. *Arch Virol* 1996, **141**(11):2243-2249.
219. Luo J, Mitra A, Tian F, Chang S, Zhang H, Cui K, Yu Y, Zhao K, Song J: **Histone methylation analysis and pathway predictions in chickens after MDV infection**. *PLoS One* 2012, **7**(7):e41849.
220. Parcells MS, Lin SF, Dienglewicz RL, Majerciak V, Robinson DR, Chen HC, Wu Z, Dubyak GR, Brunovskis P, Hunt HD *et al*: **Marek's disease virus (MDV) encodes an interleukin-8 homolog (vIL-8): characterization of the vIL-8 protein and a vIL-8 deletion mutant MDV**. *J Virol* 2001, **75**(11):5159-5173.
221. Niikura M, Kim T, Hunt HD, Burnside J, Morgan RW, Dodgson JB, Cheng HH: **Marek's disease virus up-regulates major histocompatibility complex class II cell surface expression in infected cells**. *Virology* 2007, **359**(1):212-219.
222. Abdul-Careem MF, Hunter BD, Lee LF, Fairbrother JH, Haghighi HR, Read L, Parvizi P, Heidari M, Sharif S: **Host responses in the bursa of Fabricius of chickens infected with virulent Marek's disease virus**. *Virology* 2008, **379**(2):256-265.
223. Schat KA, Calnek BW, Fabricant J: **Influence of the bursa of Fabricius on the pathogenesis of Marek's disease**. *Infect Immun* 1981, **31**(1):199-207.
224. Mustonen L, Alinikula J, Lassila O, Nera K-P: **Bursa of Fabricius**. In: *eLS*. John Wiley & Sons, Ltd; 2001.
225. Burg RW, Feldbush T, Morris CA, Maag TA: **Depression of thymus-and bursa-dependent immune systems chicks with Marek's disease**. *Avian Dis* 1971, **15**(4):662-671.
226. Mitra A, Luo J, Zhang H, Cui K, Zhao K, Song J: **Marek's disease virus infection induces widespread differential chromatin marks in inbred chicken lines**. *BMC Genomics* 2012, **13**:557.
227. Smith E, Shilatifard A: **The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes**. *Mol Cell* 2010, **40**(5):689-701.
228. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D84-90.
229. R Development Core Team: **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing; 2008.
230. Wickham H: **ggplot2: Elegant Graphics for Data Analysis**: Springer; 2009.
231. Buza JJ, Burgess SC: **Modeling the proteome of a Marek's disease transformed cell line: a natural animal model for CD30 overexpressing lymphomas**. *Proteomics* 2007, **7**(8):1316-1326.
232. Labbaye C, Testa U: **The emerging role of MIR-146A in the control of hematopoiesis, immune function and cancer**. *J Hematol Oncol* 2012, **5**:13.

233. Schulte LN, Westermann AJ, Vogel J: **Differential activation and functional specialization of miR-146 and miR-155 in innate immune sensing**. *Nucleic Acids Res* 2013, **41**(1):542-553.
234. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, Takahashi T: **A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation**. *Cancer Res* 2005, **65**(21):9628-9632.
235. Sonkoly E, Pivarcsi A: **microRNAs in inflammation**. *Int Rev Immunol* 2009, **28**(6):535-561.
236. Sassen S, Miska EA, Caldas C: **MicroRNA: implications for cancer**. *Virchows Arch* 2008, **452**(1):1-10.
237. Yao Y, Zhao Y, Xu H, Smith LP, Lawrie CH, Watson M, Nair V: **MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs**. *J Virol* 2008, **82**(8):4007-4015.
238. Tian F, Luo J, Zhang H, Chang S, Song J: **MiRNA expression signatures induced by Marek's disease virus infection in chickens**. *Genomics* 2012, **99**(3):152-159.
239. Sandhu SK, Volinia S, Costinean S, Galasso M, Neinast R, Santhanam R, Parthun MR, Perrotti D, Marcucci G, Garzon R *et al*: **miR-155 targets histone deacetylase 4 (HDAC4) and impairs transcriptional activity of B-cell lymphoma 6 (BCL6) in the Emu-miR-155 transgenic mouse model**. *Proc Natl Acad Sci U S A* 2012, **109**(49):20047-20052.
240. Bornachea O, Santos M, Martinez-Cruz AB, Garcia-Escudero R, Duenas M, Costa C, Segrelles C, Lorz C, Buitrago A, Saiz-Ladera C *et al*: **EMT and induction of miR-21 mediate metastasis development in Trp53-deficient tumours**. *Sci Rep* 2012, **2**:434.
241. Boyerinas B, Park SM, Hau A, Murmann AE, Peter ME: **The role of let-7 in cell differentiation and cancer**. *Endocr Relat Cancer* 2010, **17**(1):F19-36.
242. Bhattacharjya S, Nath S, Ghose J, Maiti GP, Biswas N, Bandyopadhyay S, Panda CK, Bhattacharyya NP, Roychoudhury S: **miR-125b promotes cell death by targeting spindle assembly checkpoint gene MAD1 and modulating mitotic progression**. *Cell Death Differ* 2012.
243. Kappelmann M, Kuphal S, Meister G, Vardimon L, Bosserhoff AK: **MicroRNA miR-125b controls melanoma progression by direct regulation of c-Jun protein expression**. *Oncogene* 2012.
244. Yu D, Tan AH, Hu X, Athanasopoulos V, Simpson N, Silva DG, Hutloff A, Giles KM, Leedman PJ, Lam KP *et al*: **Roquin represses autoimmunity by limiting inducible T-cell co-stimulator messenger RNA**. *Nature* 2007, **450**(7167):299-303.
245. Ma L, Teruya-Feldstein J, Weinberg RA: **Tumour invasion and metastasis initiated by microRNA-10b in breast cancer**. *Nature* 2007, **449**(7163):682-688.
246. Silber J, Lim DA, Petritsch C, Persson AI, Maunakea AK, Yu M, Vandenberg SR, Ginzinger DG, James CD, Costello JF *et al*: **miR-124 and miR-137**

inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. *BMC Med* 2008, **6**:14.

247. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA *et al*: **Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage**. *Cell* 2012, **151**(1):206-220.
248. Paschos K, Allday MJ: **Epigenetic reprogramming of host genes in viral and microbial pathogenesis**. *Trends Microbiol* 2010, **18**(10):439-447.
249. Roulston A, Marcellus RC, Branton PE: **Viruses and apoptosis**. *Annu Rev Microbiol* 1999, **53**:577-628.
250. Pikarsky E, Porat RM, Stein I, Abramovitch R, Amit S, Kasem S, Gutkovich-Pyest E, Urieli-Shoval S, Galun E, Ben-Neriah Y: **NF-kappaB functions as a tumour promoter in inflammation-associated cancer**. *Nature* 2004, **431**(7007):461-466.
251. Jarosinski KW, Njaa BL, O'Connell P H, Schat KA: **Pro-inflammatory responses in chicken spleen and brain tissues after infection with very virulent plus Marek's disease virus**. *Viral Immunol* 2005, **18**(1):148-161.
252. Honda R, Tanaka H, Yasuda H: **Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53**. *FEBS Lett* 1997, **420**(1):25-27.
253. Truman LA, Ford CA, Pasikowska M, Pound JD, Wilkinson SJ, Dumitriu IE, Melville L, Melrose LA, Ogden CA, Nibbs R *et al*: **CX3CL1/fractalkine is released from apoptotic lymphocytes to stimulate macrophage chemotaxis**. *Blood* 2008, **112**(13):5026-5036.
254. Nishimoto N, Ogata A, Shima Y, Tani Y, Ogawa H, Nakagawa M, Sugiyama H, Yoshizaki K, Kishimoto T: **Oncostatin M, leukemia inhibitory factor, and interleukin 6 induce the proliferation of human plasmacytoma cells via the common signal transducer, gp130**. *J Exp Med* 1994, **179**(4):1343-1347.
255. Zhou S, Kurt-Jones EA, Cerny AM, Chan M, Bronson RT, Finberg RW: **MyD88 intrinsically regulates CD4 T-cell responses**. *J Virol* 2009, **83**(4):1625-1634.
256. Reis ST, Pontes-Junior J, Antunes AA, Dall'Oglio MF, Dip N, Passerotti CC, Rossini GA, Morais DR, Nesrallah AJ, Piantino C *et al*: **miR-21 may acts as an oncomir by targeting RECK, a matrix metalloproteinase regulator, in prostate cancer**. *BMC Urol* 2012, **12**:14.
257. Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY: **miR-21-mediated tumor growth**. *Oncogene* 2007, **26**(19):2799-2803.