Genome-wide dissection of the relationship linking chromatin state and pre-mRNA processing

by

Amandine Bemmo

Department of Human Genetics
Faculty of Medicine
McGill University, Montreal

April 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

Abstract

Motivation: Alternative splicing (AS) can significantly impact cellular function by creating distinct mRNA transcripts from the same gene that have the potential to encode unique proteins, often with distinct or opposing functions. AS is a complex process that involves several, interdependent, layers of regulation. A number of studies suggest that specific histone modifications (HMs) affect pre-mRNA splicing. However, such studies have either been limited to individual genes or have not accounted for the potentially confounding influence of gene expression levels. Splicing is a co-transcriptional process, and in some cases its efficiency has been shown to be affected by the speed of transcription. We hypothesize that the relationship between splicing and transcription is a widespread phenomenon, and results in a transcriptome-wide correlation between gene expression and exon inclusion. This may result in spurious correlations between chromatin marks and splicing.

Results: Using transcriptomic and epigenomic profiles from muscle, monocyte and T cells in humans we demonstrated that the effects of gene expression and the uneven distribution of epigenetic marks across gene bodies can confound the relationship between pre-mRNA processing and epigenetic marks. We subsequently investigated AS events occurring within the same gene, lacking positional biases, in order to correct for these confounders and focus fully on the relationship between exon inclusion, intron retention (IR), and chromatin states. After removing these confounding influences, a number of findings were no longer in agreement with previous studies, namely the differential 5’SS strength between alternative and constitutive exons, differential methylation level between alternative and constitutive exons, and differential CpG content and methylation level between retained and non-retained introns. At the epigenome level, our analysis unprecedentedly revealed that H3K4me3, H3K4me1 and H3K27ac are less enriched in retained introns compared to non-retained introns. Nevertheless, a number of previously reported epigenetic, transcriptomic and sequence features remain significantly associated with exon inclusion and IR.

Conclusion: Here, we remove the effect of gene expression from the relationship linking epigenetic modifications and pre-mRNA splicing. Our data show a distinct coupling between the splicing machinery and the chromatin states independently of the transcription effect. Collectively, our results provide valuable and unique insights into the mechanism of epigenome-mediated alternative splicing, with particular importance for HMs.

Resumé

Motivation : L’épissage alternatif affecte environ 95% des gènes chez les mammifères. C’est l’un des mécanismes les plus importants permettant d’accroître la diversité du transcriptome et du protéome. L’épissage alternatif est un processus complexe impliquant plusieurs niveaux de régulation interdépendants. Selon plusieurs études, les modifications d’histones affectent l’épissage du pre-mARN. Cependant cette évidence repose sur des études qui restreignent leur analyse à des gènes individuels et/ou ne considèrent pas l’influence de l’expression du gène et de la distribution des marques épigénétiques. L’épissage des gènes est un processus qui se déroule parallèlement à la transcription; pour certains gènes, il a été montré que l’efficacité de l’épissage est influencée par la vitesse de transcription du gène. Notre hypothèse est que la relation entre l’épissage et la transcription est un phénomène répandu dans le génome, et donc en découle une corrélation entre l’expression du gène et le niveau d’inclusion des exons. Ceci pourrait engendrer de fausses corrélations dans l’analyse des facteurs qui affectent l’épissage.

Résultats : En utilisant les échantillons de cellules musculaires, de monocytes et lymphocytes T chez l'humain, nous avons montré que le niveau d'expression du gène et la distribution non-uniforme des marques épigénétiques sont des biais pouvant fausser la relation entre l'épissage du pré-mARN et les marques épigénétiques. Par la suite nous avons investigué les événements d’épissage alternatif ayant lieu dans le même gène et occupant les mêmes positions, dans le but de supprimer ces biais et de nous concentrer uniquement sur la relation entre l’inclusion d’exon, la rétention d'intron et les marques épigénétiques. Après avoir supprimé ces biais, un nombre de facteurs antérieurement associés à l'épissage du pré-mARN n'était plus en accord avec les études antérieures; notamment la différence de méthylation et de force du site d'épissage 5’ entre les exons constitutifs et alternatifs, la différence de méthylation et de fréquence de sites CpG entre les introns retenus et les introns non retenus. Notre analyse révèle sans précédent que les modifications d'histones H3K4me3, H3K4me1 et H3K27ac sont plus abondants chez les introns non retenus comparativement aux introns retenus. Néanmoins, un nombre de facteurs précédemment associés à l'inclusion d'exon et la rétention d'introns demeurent significatifs.

Conclusion : Dans cette étude, nous avons démêlé l’effet de l’expression du gène de la relation entre les marques épigénétiques et l’epissage du pré-ARNm. En tout, nos résultats dévoilent différents profils de rétention d’intron, d’occupation de modifications d’histone et de propriété de séquences entre les exons alternatifs et constitutifs. Nos analyses démontrent aussi bien différents profils de modifications d’histone et de propriétés de séquences entre les introns retenus et les introns non retenus. Ceci renforce l’idée que l’épissage est un processus qui ne se déroule pas de façon isolée mais qui est plutôt couplé à la transcription. Notre analyse intégrant le transcriptome et l’épigénome fait ressortir la relation entre les modifications d’histone, la rétention d’intron et l’inclusion de l’exon, qui est une relation dépourvue de l’effet de la transcription.

Acknowledgements

My most sincere gratitude goes to my supervisor Dr. Jacek Majewski. There are no proper words to convey the depth of my regards and respect for you. You provided an opportunity to explore scientific freedom, allowed me to test new ideas, and generously supported my study. Your approachability, passion for science and remarkable attentiveness really help my projects succeed. Thank you for your patience, your advice, and the many hours spent discussing about my project and editing my papers.

I would also like to thank my supervisory committee members, Dr. Tomi Pastinen and Dr. Thomas Duchaine for monitoring project progression. During all my Supervisory Committee meetings they have provided thoughtful and detailed comments.

I would like to thank all my current and former lab mates at McGill Genome center for their inspiring discussion and generous collaboration. I have enjoyed working with all of you. A special thanks goes to, Claudia Kleinman, Warren, Simon Papillon, Maxime Caron, Veronique, Jian, Eric, Matt, Hamid, Najmeh, Somayyeh, Martine and Javad.

This work would not have been possible without the financial support of our funding sources, Canadian Institute for Health Research.

Finally, I would like to thank my family, my mother, brothers and sister for their love and unwavering support.

Contribution of authors and research funding

This thesis is written in traditional format as permitted by the McGill Faculty of Graduate Studies and is comprised of three chapters. The contribution of each author is described below:

Dr Jacek Majewski was the main editor and principal investigator for this study. He contributed the thesis ideas and the ideas related to the statistical aspects. He also contributed to the manuscript and thesis review.

Dr Tomi Pastinen provided for the data from the Epigenome McGill center. He contributed the thesis ideas and the ideas related to the statistical aspects.

Data generation and pre-processing for all chapters were performed by Dr Pastinen's lab.

Dr Pastinen's lab wrote the part of the Material and Methods section about data generation and pre-processing.

I did all the analyses and generated all the figures and tables for all chapters. I wrote the thesis and the manuscript to be published. The manuscript representing a synthesis of the main findings of the thesis has been submitted for publication in the journal *BMC Genomics*.

List of figures

List of tables

List of abbreviations

AS	Alternative Splicing
bp	Base Pair
CE	Cassette Exon
CG	Cytosine-Guanine
ChIP-Seq	CHromatin ImmunoPrecipitation coupled with high throughput SEQuencing
CpG	Cytosine-phosphate-Guanine
DNA	DeoxyriboNucleic Acid
FDR	False Discovery Rate
GERP	Genomic Evolutionary Rate Profiling
GO	Gene Ontology
H3K27ac	Histone H3 Lysine 27 acetylation
H3K27me3	Histone H3 lysine 27 trimethylation
H3K36me3	Histone H3 lysine 36 trimethylation
H3K4me1	Histone H3 lysine 4 monomethylation
H3K4me3	Histone H3 lysine 4 trimethylation
H3K9me3	Histone H3 lysine 9 trimethylation
HM	Histone Modification
IR	Intron Retention
mCpG	methylated Cytosine-phosphate-Guanine

mRNA	messenger RiboNucleic Acid
NMD	Nonsense-mediated mRNA Decay
pre-mRNA	precursor messenger RiboNucleic Acid
Pol II	Polymerase II
RPKM	Reads Per Kilobase per Million mapped reads
RNA	RiboNucleic Acid
3’SS	Three prime Splice Site
5’SS	Five prime Splice Site
T Cells	T lymphocytes
UCSC	University of California Santa Cruz
UTR	UnTranslated Region

List of websites

UCSC: https://genome.ucsc.edu/

Maximum entropy:

http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html

Illustration of gene expression flow: http://www.nature.com/scitable/topicpage/gene-expression-14121669

Chapter 1: Introduction

A gene is a linear sequence of DNA at a specific region in the genome that provides the coded instructions for the synthesis of a functional RNA or protein product. In 1909, the one gene-one protein hypothesis was firstly proposed by the English physician Archibald Garrod. It suggests that each gene codes for a single specific enzyme. In the early 1940s, the evaluation of this concept by the American geneticists Edward Tatum and George Beadle led to the acceptance of this idea [1]. This concept has been revised over time as our knowledge of molecular biology of the gene improves. In 1993, the Nobel Prize in Physiology or Medicine was awarded jointly to Phillip A. Sharp and Richard J. Roberts for the discovery that human genes can be segmented, with exons interrupted by introns. Introns are transcribed but not included in the final protein product as they are excised before translation; the excision of introns and the joining of exons is a process known as pre-mRNA splicing. Further advances in technology and science have opened up new avenues for the investigation and understanding of pre-mRNA processing and its role in the regulation of gene.

Within the last two decades, the entire genome of human [2, 3] and various organisms such as mouse, worm and drosophila [4-6], have been sequenced. One of the surprising findings was the identification of only 30,000 human genes (20,000-25,000 protein-coding genes). In comparison, the nematode *Caenorhabditis elegans* has 19,000 genes and the fruit fly *Drosophila* has 14,000 genes. The realization that human has a comparable number of genes as *Caenorhabditis elegans* and only twice as many genes as the fly raised important questions about the source of organismic complexity. How can the much higher complexity of humans be encoded in only twice the number of genes required by a simple fruit fly? The investigation of human transcriptome indicates that the number of expressed-sequence (mRNA) forms far exceeds the number of genes. This implies that a large portion of the genes have the ability to encode multiple proteins. This process by which a single gene (pre-mRNA) can be spliced in different ways, via the differential usage of splice sites, to produce different mRNA molecules is known as alternative splicing (AS).

AS has emerged as a major source of organismic complexity since it can significantly expand the coding capacity of the genome. It is estimated that nearly 95% of mammalian multiexon genes undergo AS [7-9]. This can lead to *(i)* the creation of molecules with different properties and functions through the addition or deletion of protein domains, or *(ii)* the alteration of mRNA stability resulting in the degradation of the mRNA. Different cells of multicellular organisms share the same DNA code but present heterogeneous structural and functional properties, due to the differential splicing or expression of genes. AS represents a widely acting mode of gene regulation to generate different protein isoforms at different times in development and/or in specific cell or tissue types. AS can have important impact on disease states such as cancer. AS in cancer results in the aberrant expression of transcripts that participate in tumor cell survival, proliferation, invasion, and metastasis [10-13]. Consequently, understanding underlying factors of AS regulation in different biological conditions, cells or tissue types may provide a new view angle of organism complexity, phenotype diversity, and could be applied in therapeutic fields.

In the recent few years technological advances have generated massive amounts of new genome-scale information that has the potential to provide greater insights into mechanisms regulating AS. We still have only just begun to explore those mechanisms. Armed with next-generation sequencing technologies, it is now feasible to identify at multiple levels (e.g DNA, RNA and epigenome levels) genome-wide signals associated with AS. However, there remain many challenges in transforming this amount of information into a more accurate understanding of features associated with AS. One major task is to capture and correct for potential biases hidden in the data to dissect out the contribution of specific features in AS.

1.1 Motivations and objectives

Accumulating evidence indicates that the temporal and spatial regulation of splicing is a multi-layer and complex mechanism in which the splicing machinery interacts with transcription machinery, DNA methylation, histone modifications, and nucleosome positioning. Although the coupling between genetics and epigenetics in the regulation of gene expression has been intensively studied, many fundamental questions about how genetic and epigenetic features interact in concert to influence splicing choices remain unclear. Evidence from AS studies, accumulated over the past decade, indicates that most splicing occurs co-transcriptionally meaning that introns are excised from the pre-mRNA before RNA pol II reaches the end of the gene [14-17]. Therefore the factors that regulate transcription may also affect splicing, and this may introduce confounders when analyzing AS. While intron retention (IR), histone modifications (HM) and methylation have been relatively well characterized in gene expression regulation, increasing evidence supports the association of AS with IR, HMs and methylation. However, the evidence is mostly limited to individual genes and/or does not take into account the potentially confounding relationship between transcription and splicing. These limitations have prompted us to investigate, at the genome scale, the relationship between AS and specific features in muscle and two types of human white blood cells, monocytes and T lymphocytes, by correcting for transcription bias. These features are IR, histone modifications, methylation, and DNA sequence characteristics.

1.2 Outline

This thesis consists of a literature review, three chapters reporting our findings, and a discussion/conclusion that together address the study of association between exon inclusion, IR level, epigenetic marks and sequence features:

- The literature review, chapter 2, summarizes the basic mechanisms of transcription and pre-mRNA processing, describes how these processes are regulated, and relates some of their effects on organism phenotype, diversity and disease. It also describes the prior findings on genetic and epigenetic factors involved in AS regulation.

- The third chapter is dedicated to investigate confounding factors that may influence AS analysis.
- In the fourth chapter, by using an approach correcting for the confounding factors above, we explore at the genome scale how exon inclusion is associated with histone modifications, methylation, IR and sequence features in the vicinity.
- In the fifth chapter, we focus on IR and its underlying factors, while controlling for the confounders above. Particularly, we investigate the genome-wide epigenetic and sequence features that characterize IR.
- The last section, the sixth chapter, is a summary of the main results and a discussion of the future work that is needed to better understand the relation between IR, exon inclusion and epigenetic marks.

Chapter 2: Literature review

2.1 Regulation of Gene Expression

Gene expression is the process by which the information within a gene is used in the synthesis of either a protein or a functional RNA, such as ribosomal RNA and transfer RNA. In mammals, the regulation of protein-coding gene expression requires several steps such as transcription, pre-mRNA splicing, 5’capping and polyadenylation additions, translation and post-translation modification of the protein (Figure 2.1). The first step is the transcription of the gene into a primary RNA transcript (pre-mRNA) that occurs in two main steps: the assembly of the transcription complex at the gene promoter, and the elongation. The second stage is the processing of this pre-mRNA into a mature messenger RNA (mRNA) in three main steps consisting of 5'-capping, splicing and polyadenylation. Then, the mRNA is exported from the nucleus to the cytoplasm where the translation into protein takes place. Subsequently, the protein can be subjected to post-translational modifications such as glycosylation, phosphorylation, and sulfation.

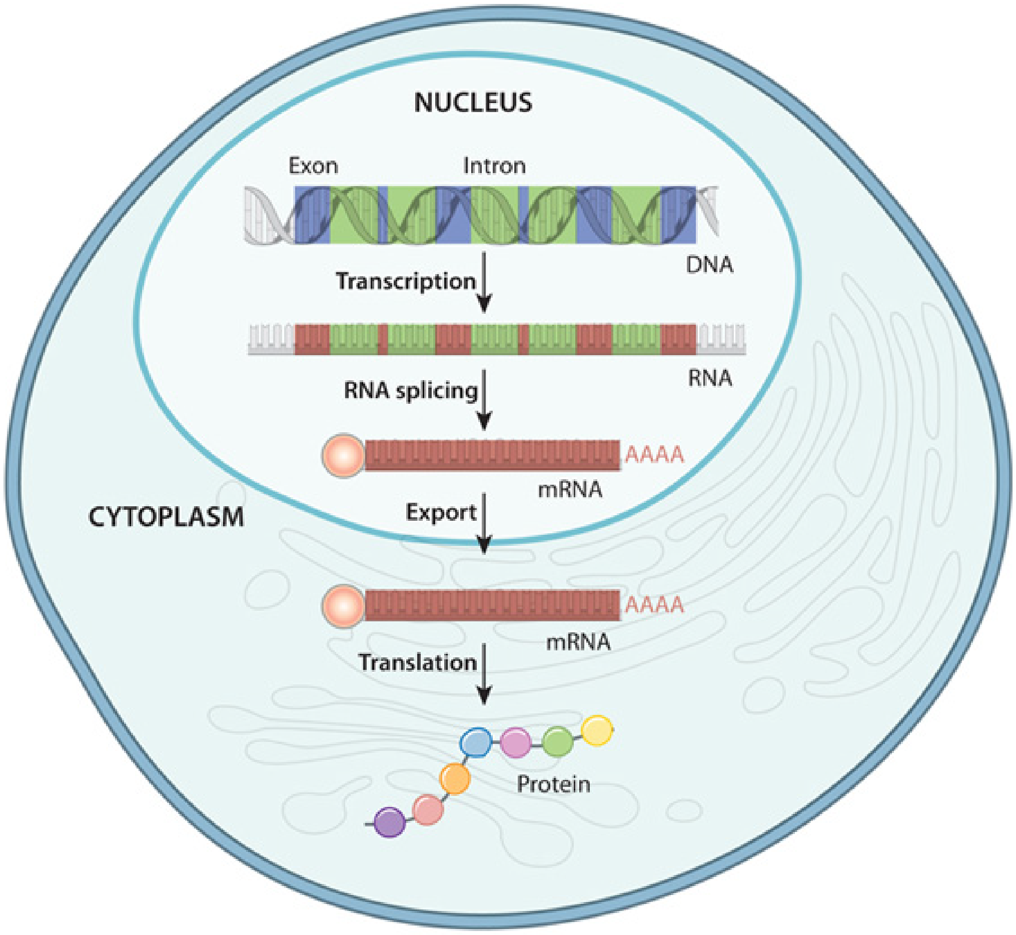


Figure 2.1: Summary of information flow from DNA to protein in eukaryotes.

First, both coding and non-coding regions of DNA are transcribed into pre-mRNA. During the processing of mRNA, introns are excised; the exons are then spliced together. A cap (sphere) and a polyA tail are added to the 5' end and 3' end of the spliced mRNA molecule (red box) respectively. Then the mRNA is exported out of the nucleus. Once in the cytoplasm, the mRNA can be used to generate a protein.

Figure reproduced from http://www.nature.com/scitable/topicpage/gene-expression-14121669.

2.1.1 Transcription

Transcription, the first step of gene expression, is the process by which the information contained in a particular piece of DNA strand is copied into a new molecule of messenger RNA that in turn serves as a template for protein synthesis through translation. Transcription is usually carried out by an enzyme called RNA polymerase II (RNA pol II) coupled to a number of accessory proteins that affect RNA pol II activity by promoting or inhibiting the recruitment of the RNA pol II [18, 19]. In order to recruit RNA pol II to an appropriate transcription site, transcription factors bind to specific DNA sequences: *(i)* promoters, which are almost always located near the starting point for transcription, and *(ii)* enhancers that are located up to 1Mbp upstream and/or downstream of the transcription start site [20]. Together, RNA pol II and the transcription factors form a complex called the transcription initiation complex. Once the transcription initiation complex is assembled on the promoter, transcription starts; the RNA pol II begins RNA synthesis by reading the DNA template and producing a matching complementary antiparallel RNA [21] called the primary mRNA.

2.1.2 Pre-mRNA processing

The newly synthesized pre-mRNA is extensively edited prior to the production of the mRNA ready for translation by the ribosome. Processing of pre-mRNA usually occurs co-transcriptionally, meaning that 5’-capping, intron excision from the pre-mRNA, and polyadenylation tail attachment all occur before the RNA pol II reaches the end of the gene [22-26] (Figure 2.2).

2.1.2.1 5’-capping

As soon as the nascent RNA emerges from the polymerase complex, the 5’ end undergoes a chemical modification with the addition of a 7-Methyl guanine to the triphosphate end of the transcript by three enzymes: a phosphatase, a guanyltransferase and a methylase. The phosphatase removes the triphosphate group of the first nucleotide transcribed, then the guanyltransferase (the capping enzyme) attaches a guanosine via a 5'–5' triphosphate linkage, and the methyltransferase methylates the N7 position of the added guanosine [27]. The 5' cap regulates mRNA export to the cytoplasm [28, 29], prevents of RNA degradation by exonuclease [30-32], promotes 5’ proximal intron excision [33] and regulates translation [34-36].

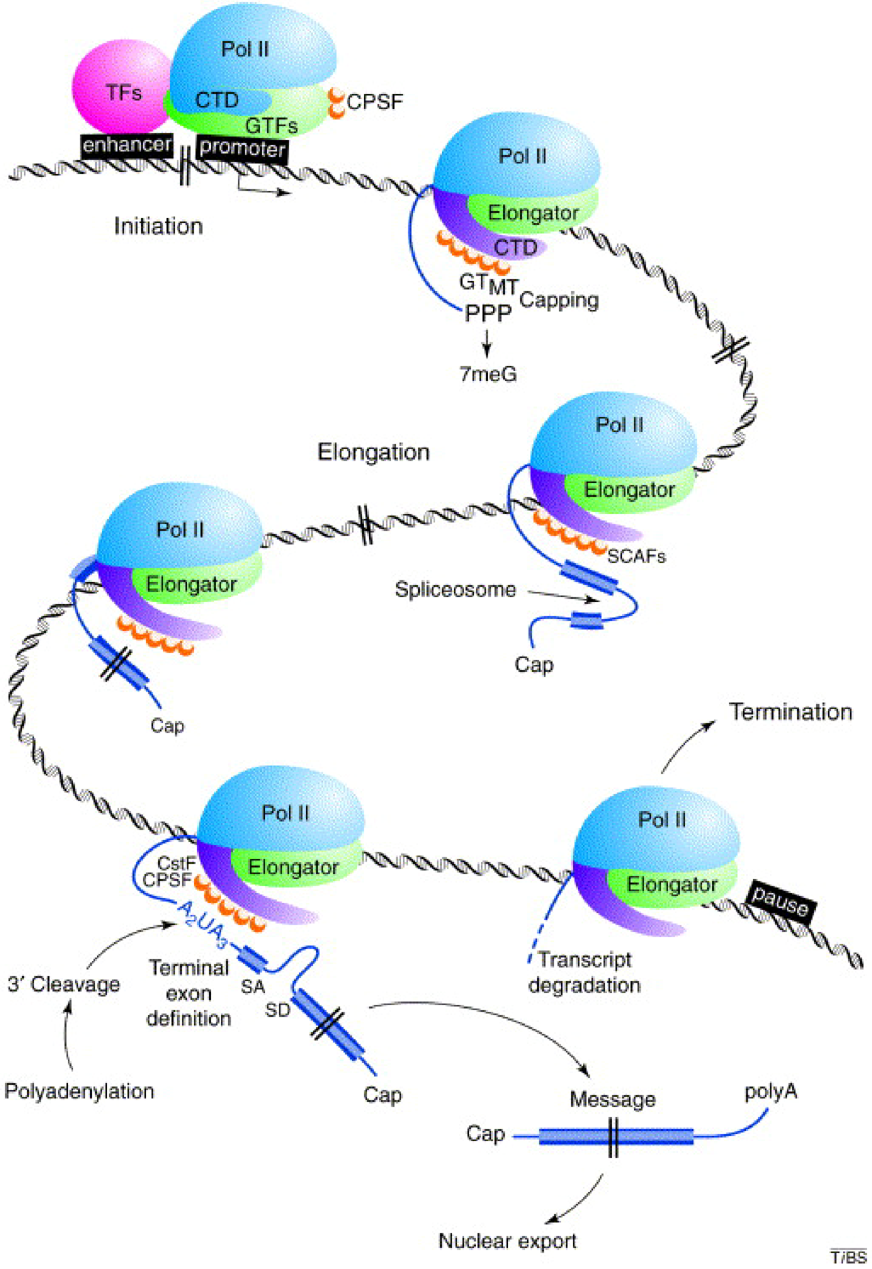


Figure 2.2: Co-transcriptional processing of pre-mRNA.

Schematic illustrating how pre-mRNA processing (5'-capping, splicing and polyadenylation) is associated with the three stages of transcription (initiation, elongation and termination) to form a mRNA. Figure reproduced from Proudfoot et *al* [37].

2.1.2.2 Constitutive splicing

The next RNA-processing reaction to take place on the nascent transcript is the removal of intron sequences from the pre-mRNA and the joining together of exons (Figure 2.3). Canonical splicing is catalyzed by the spliceosome, a large RNA-protein complex of subunits each containing a small nuclear ribonucleoprotein (snRNP) and associated proteins [38-40]. This complex is assembled at the splice sites located at the intron-exon boundaries. Certain signal sequences, at the intron-exon boundaries and within introns, need to be recognized and processed by the spliceosome. Among the requirements are *(i)* a GU, at the donor splice site, located at the 5' end of the intron, *(ii)* an AG, at the acceptor splice site, located at the 3' end of the intron, and *(iii)* a branch site located anywhere from 18 to 40 nucleotides upstream from the ’3end of the intron and containing an adenine that plays a key role in intron removal [41, 42]. The first step of the spliceosome assembly is the binding of the U1 snRNP to the donor splice site; then follows the binding of the U2 snRNP to the branch site. Next, the trimer U4 U5 and U6 snRNPs bind at the intron region, completing the spliceosome assembly. The 5’SS is cut and the 5’end of the intron is attached to the branch site through the pairing of guanine and adenine nucleotides from the 5’ end and the branch point, respectively. The formed structure is called a lariat. The U1 and U4 snRNPs are released. The 3’ SS is cut and the exons connected together. The lariat is released from the remaining part of the spliceosome for degradation [43], and the spliceosome subunits are later dissociated. However, several spliced introns appear to escape the degradation for serving useful functions in the cell [44].

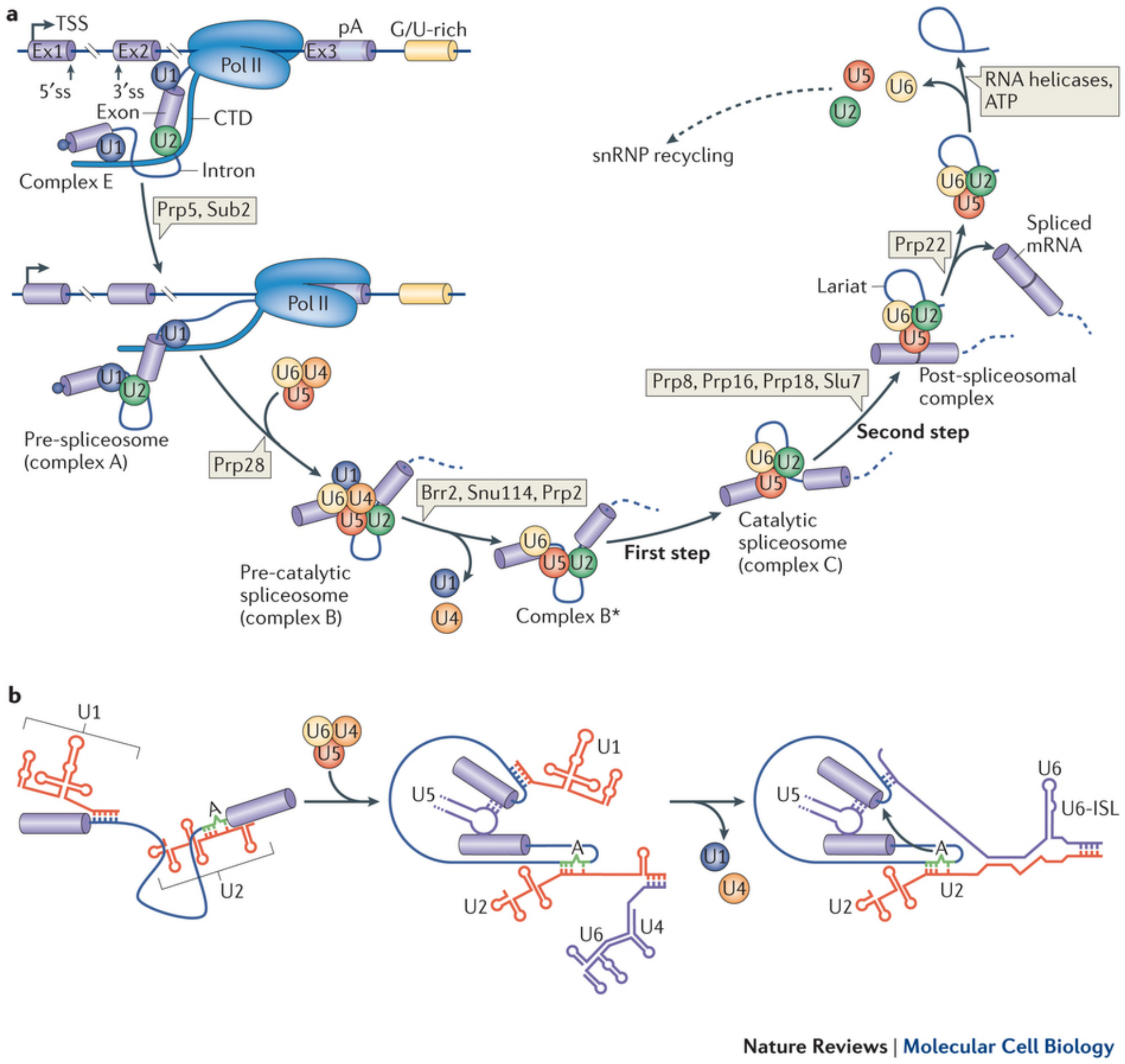


Figure 2.3: Assembly of the spliceosome by the stepwise binding of snRNPs to the pre-mRNA.

(a) Splicing of pre-mRNA takes place in several steps that are catalyzed by small nuclear ribonucleoproteins (snRNPs). (b) During splicing, the catalytic centre of the spliceosome is created by the stepwise rearrangement of RNA–RNA interactions. Figure reproduced from Matera *et al* [40].

A minor category of introns (less than 1% of all introns in the human genome) are processed by a secondary type of spliceosome relying on U12. The splicing process is very similar to the canonical U2-dependent spliceosome. The main differences between U2- and U12-type introns are the 5’ splice site and branch site sequences [45]. The U1, U2, U4 and U6 found in the U2-dependent spliceosome are substituted by four different snRNPs U11, U12, U4atac and U6atac respectively, in the U12-dependent mechanism [46-48].

2.1.2.3 3’ end polyadenylation

A chain of adenine bases, called a polyadenylation tail, is attached to the 3' end of the mRNA. The polyadenylation addition is catalyzed by the polyadenylate polymerase enzyme, which recognizes the conserved upstream sequence AAUAAA and the downstream G/U-rich sequence as signals for the tail addition [49, 50]. In the process, the 3' end of the synthesized RNA is cut by a multicomplex protein [51] and a stretch of RNA that has only adenine bases is synthesized at the 3' end of the RNA [52]. The polyadenylation tail influences the nuclear export, translation, and stability of mRNA. Most eukaryotic RNA transcripts are polyadenylated, with the exception of some genes, such as histone genes [53].

2.1.3 Translation

Once the mRNA is exported to the cytoplasm, the ribosome complex (composed of several ribosomal RNA molecules and a number of proteins) synthesizes the protein by reading the mRNA sequence based on the genetic code [54], a set of combinations of tri-nucleotides (codons) each corresponding to a specific amino acid or stop signal (Figure 2.4). The ribosome complex binds to the mRNA start and passes along in one codon at the time. Each time a new codon moves into the site, a new transfer RNA (tRNA) brings in an amino acid. The complementary matching of the three nucleotides on the tRNA (called the anti-codon) and the three nucleotides on the mRNA (codon), ensures the correct sequence of amino acid. Subsequently the amino acid is transferred to the growing amino acid chain. Elongation continues until all the codons are read; the termination occurs when the ribosome reaches a stop codon (UAA, UAG, and UGA). At this point the ribosomal complex is disassembled and the protein is released in the cell [55].

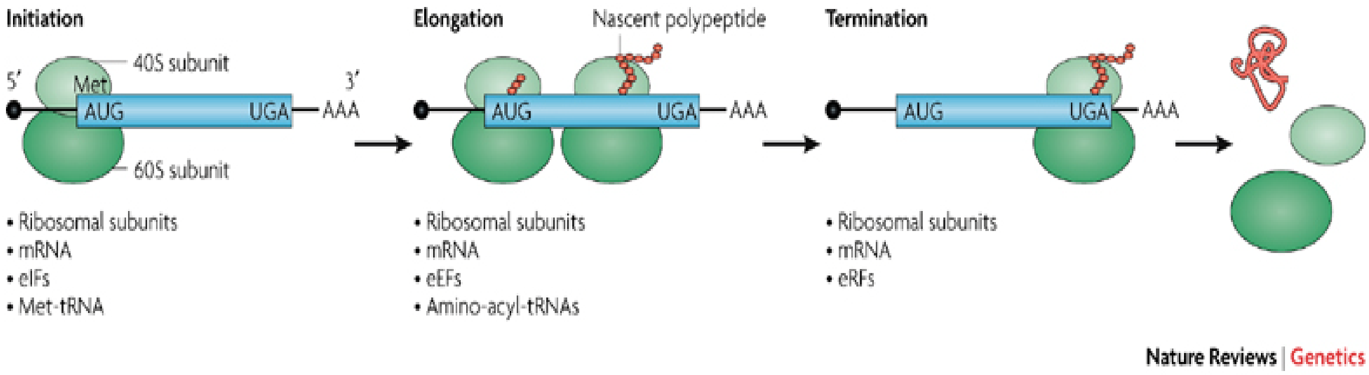


Figure 2.4: Translation process of protein synthesis.

Figure reproduced from Scheper *et al* [56].

2.2 Control of gene expression

Different cells of multicellular organisms share the same DNA code but present heterogeneous structural and functional properties, due to differential regulation of genes. Gene regulation is important as it increases the flexibility and adaptability of an organism by enabling a cell to express a protein when it is needed, for example a cell- or tissue-specific expression at a specific development phase, cellular differentiation or morphogenesis. At any step of gene regulation, a cell can use a wide range of mechanisms to modulate gene expression.

2.2.1 Regulation at the DNA level

At the DNA level, gene expression can be influenced through epigenetic modifications. As examples, *(1)* phosphorylation, acetylation or methylation of histones which can modify the DNA structure by influencing DNA packing [57], and *(2)* DNA methylation at gene-promoter regions to activate or silence gene [58, 59].

2.2.2 Regulation at the RNA level

Much of the gene expression control in eukaryotes is performed through transcriptional and post-transcriptional regulation. Examples are *(1)* the influence of RNA pol II activity by regulatory transcription factors that enhance or repress RNA pol II activity, *(2)* alternative transcription start or alternative transcription end (Figure 2.5 parts f and g), *(3)* the modification of canonical splicing through the usage of differential splice sites (Figure 2.5 parts a-e), *(4)* the modulation of mRNA transportation in the cytoplasm by controlling access or efficiency of transport channel (receptors in the interior of the pores) of the mRNA, *(5)* the modulation of the degree of mRNA degradation, *(6)* trans-acting regulators such as microRNA binding to mRNA 3' UTR to silence gene expression, *(7)* the influence of mRNA stability through the modification of 5' capping and polyadenylation tail.

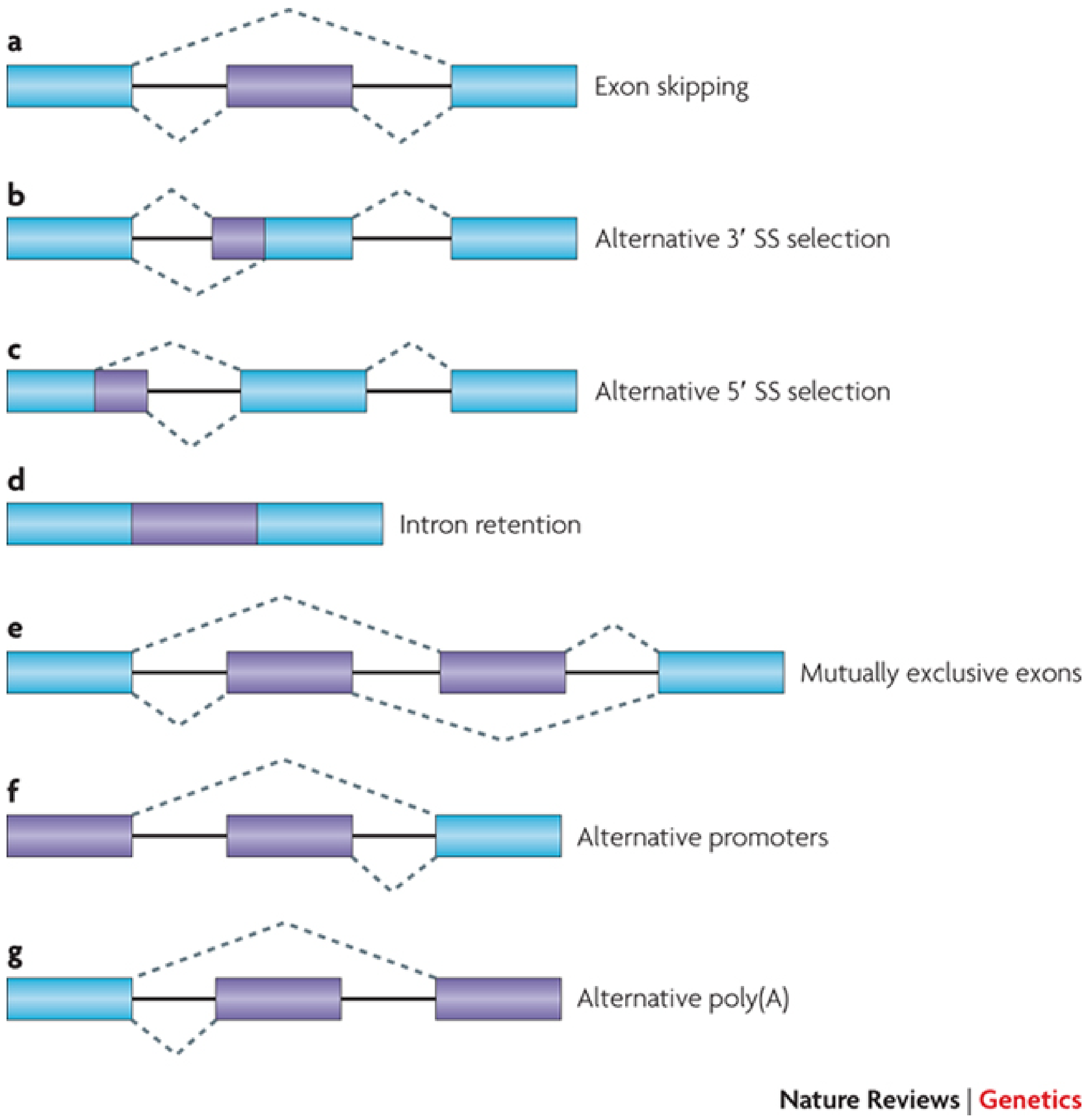


Figure 2.5: Types of AS and alternative transcription

Gene structure is described as exons (colored rectangles) connected by introns (solid black lines). Dashed lines indicate splicing options. In each model, constitutive exons are blue and alternative regions are purple. Figure reproduced from Keren *et al* [60].

2.2.3 Translational and post-translational regulation

These mechanisms include, *(1)* the modification of protein synthesis from the mRNA template, for example the availability of any protein or amino acid required for the protein synthesis [61]. For example the inhibition of translation initiation and protein synthesis in response to specific stress signals [61-63]; *(2)* post-translation modifications of the synthesized protein [64] influencing its activity such as phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, lipidation and proteolysis.

2.3 Co-transcriptional splicing

Remarkable progress made in the field of gene expression regulation has consistently revealed increasing levels of complexity. Splicing regulation is in fact more complex and displays extensive cross-talk between several layers of regulation. Evidence so far indicates that splicing is not only regulated through trans-acting splicing factors but also by processes involving the transcription machinery. Transcription and splicing were thought to be independent events for many years. But in fact most splicing occurs co-transcriptionally meaning that introns are excised from the pre-mRNA before the RNA pol II reaches the end of the gene [14-17]. Splicing is regulated co-transcriptionally at multiple levels in which the chromatin structure, DNA methylation, histone modifications and nucleosome positioning are key features that are involved in splicing regulation. Two non exclusive mechanisms have been suggested to explain the coupling between transcription and splicing machinery: recruitment coupling in which splicing is affected by the recruitment of splicing factors to the transcription apparatus, and kinetic coupling in which the splicing is modulated by the speed of RNA pol II elongation.

2.3.1 Recruitment coupling model

In this mode, the transcription machinery recruits splicing factors to transcription sites. The RNA polymerase II carboxy-terminal domain (Pol II CTD) plays a major role in functionally coupling transcription with pre-mRNA processing reactions namely 5’ capping, 3’ processing, and AS [23, 24, 65]. Pol II CTD offers a flexible landing pad for various transcription and splicing factors and facilitates their recruitment on the nascent pre-mRNA [66]. Splicing factors may associate with the polymerase via the hyperphosphorylated Pol II CTD [67]. An example of how the transcription complex can influence AS through the recruitment of splicing factors is the Pol II CTD which mediates the inhibitory effect of the serine/arginine-rich splicing factor 3 (SRSF3) on the inclusion of the cassette exon 33 in fibronectin mRNA [68]. An example of splicing factor recruitment that does not rely on pol II CTD has been reported for the thermogenic activator PGC-1, which promotes the inclusion of fibronectin exon 25 into the mRNA only if it can bind the promoter of the gene [69].

2.3.2 Kinetic coupling model

This model proposes that RNA pol II-mediated elongation rate influences the regulation of AS by affecting the pace at which splice signals are exposed to the splicing factors in the nascent pre-mRNA during transcription. For example, a high RNA pol II elongation rate due to a strong upstream promoter or an open chromatin structure will increase the possibility that weak 3' SS around cassette exons will be outcompeted by an already transcribed strong 3' SS downstream, leading to the skipping of alternative exons [70-72]. In comparison, if the RNA pol II elongation rate is low, enough time will be provided for the splicing machinery to recognize any weak 3' SS resulting in the inclusion of alternative exons [73-77]. For example Kadener *et al.* demonstrated that slow transcription of the fibronectin gene (FN1) results in the inclusion of the fibronectin extra domain 1 (ED1) exon which is preceded by a weak 3' SS. However, this exon was excluded when the transcription elongation rate was higher [78]. Another example is the DNA-binding protein CTCF (involved in targeting gene insulators) which promotes the inclusion of the cassette exon 5 in CD45 by binding to a target site in the downstream intron of exon 5, thus creating a barrier to RNA Pol II elongation[79]. Interestingly, DNA methylation of this intronic site prevents CTCF binding, releases RNA Pol II, and reverses the inclusion of the cassette exon 5 splicing [80].

2.3.3 Nucleosome positioning

Analyses of chromatin have shown that chromatin structure influences splicing choices by both histone modification and nucleosome positioning [81, 82]. The nucleosome is the fundamental packaging unit of the chromatin. Each nucleosome consists of a 147-bp segment of double-stranded DNA wrapped around a single histone octamer unit. Mapping of nucleosome locations at the genome-wide level, in human as well as in caenorhabditis elegans and drosophila melanogaster, revealed that nucleosomes are not randomly located but instead positioned preferentially on exons compared to introns, and thus can enhance exon recognition during co-transcriptional splicing [83-87]. Nucleosome enrichment on exons was found to positively correlate with RNA pol II accumulation indicating a reduced rate of elongation in these regions. In the proposed mechanism, DNA tightly packaged into nucleosome acts as a barrier to RNA pol II elongation; this transcriptional pausing might further promote splicing factor recruitment resulting in higher exon inclusion into the mRNA [88-90].

2.3.4 Histone modifications

DNA is wrapped around a molecule of eight proteins grouped together called histone. DNA and histone together is called chromatin. Histones can undergo covalent modifications through the addition of chemical groups to the tails. There are more than 50 possible histone tail modifications. The protein modifications influence binding affinity among histones and between histones and DNA; consequently, their presence can modulate the chromatin state. Characteristic patterns of histone modifications have been associated with active or repressed chromatin. Histone modification level over upstream or core promoter regions is predictive of gene expression levels [91]. Some histone modifications are abundant at RNA pol II transcription start sites, particularly H3K4me3 [91]. Specific histone modifications are associated with active transcription (for example H3K36me3, H3K4me2, H3K4me3 and H3K9ac) whereas others are associated with transcriptional silencing (for example, H3K9me2, H3K9me3 and H3K27me3). Moreover, some histone modifications, such as H3K36me3, H3K79me1, H2BK5me1, H3K27me1, H3K27me2, are more enriched over internal exons compared to introns and are clearly correlated with exon expression [83, 85, 92-95]. Among these modifications, H3K36me3 is the most strongly associated with exonic regions and correlates with the level of exon inclusion and gene expression [85, 93].

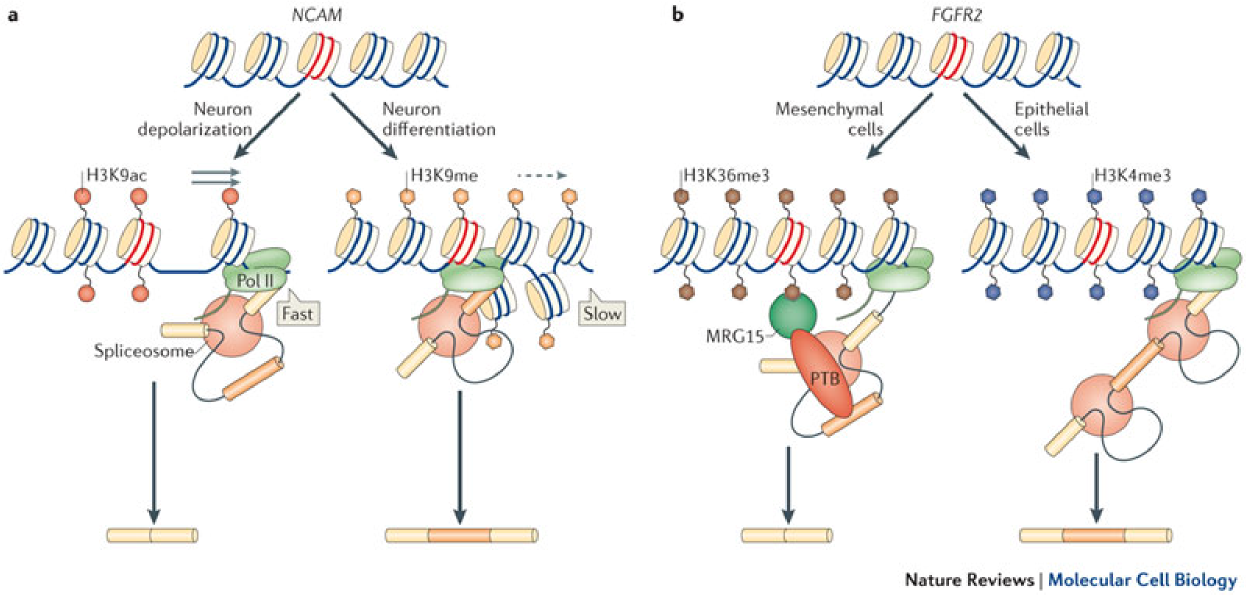


Figure 2.6: Two alternative mechanisms by which chromatin may influence alternative splicing.

Alternative splicing decisions are affected by the nature of histone marks that are deposited on the chromatin around a gene in response to external stimuli or to the differentiation state of the cell. **a |** Example of how histone modifications can affect kinetic coupling between transcription and alternative splicing. Neuron depolarization triggers intragenic acetylation of histone 3 at Lys9 (H3K9ac) and a subsequent increase in RNA polymerase II (Pol II)-mediated elongation; this favours skipping of neural cell adhesion molecule (NCAM) exon 18. Conversely, neuron differentiation promotes inclusion of exon 18 in NCAM through H3K9 methylation (H3K9me), causing a reduction in Pol II-mediated elongation (Ref. 57 [71] and I.E.S. and A.R.K., unpublished observations). **b |** Histone modifications can affect alternative splicing through a recruitment coupling mechanism. In mesenchymal cells, intragenic H3K36 trimethylation (H3K36me3) at the FGFR2 (fibroblast growth factor receptor 2) locus recruits the negative splicing factor PTB through the adaptor protein MRG15, and this results in exclusion of an alternative exon. Conversely, inclusion of this FGFR2 exon is increased in epithelial cells in which levels of H3K36me3 are lower compared with H3K4me3, which reduces MRG15 recruitment [96].

Figure reproduced from Kornblihtt et *al* [97].

Histone modifications are believed to contribute to AS regulation. An example of how a histone modification can affect kinetic coupling between transcription and splicing is at the NCAM (neural cell adhesion molecule) gene locus. Membrane depolarization of neuronal cells triggers the accumulation of intragenic H3K9ac, promoting an open chromatin structure that increases RNA pol II-mediated elongation and subsequently the skipping of exon 18 (Figure 2.6.a left). Conversely, during neuron differentiation, the silencing histone marks H3K9me2 and H3K27me3 are enriched in NCAM gene body causing a reduction of RNA pol II elongation and a higher inclusion of exon 18 (Figure 2.6.a right) [71]. Histone modifications are differentially distributed with respect to exon-intron architecture of genes. For instance, the trimethylated Lys36 of histone H3 (H3K36me3) was found to be generally enriched in constitutive exons but less prominent at alternative exons [98]. H3K36me3 has been shown at the FGFR2 (fibroblast growth factor receptor 2) locus to regulate alternative splicing for which exon IIIb is included in PNT2 prostate cells whereas exon IIIb is skipped in mesenchymal stem cells (MSCs). H3K36me3 through the recruitment of the negative splicing complex MRG15-PTB leads to the exclusion of exon IIIb in MSCs (Figure 2.6.b left). Conversely, exon IIIb inclusion in epithelial cells correlates with a higher level of H3K4me3 (compared to H3K36me3) which reduces the MRG15 recruitment (Figure 2.6.b right) [96]. Interestingly, several PTB-dependant alternatively spliced exons in other genes (e.g TPM1, TPM2 and PKM2) showed similar splicing-specific HM patterns between PNT2 and MSC in the same study. Additionally, H3K4me3 and H3K9me3 have been reported to affect splicing events in CHD1 [99] and multiple exon skipping in CD44 [100].

2.3.5 DNA methylation

Chromatin regions have distinct profiles of DNA methylation. DNA methylation is known to regulate transcription trough gene promoter marking and also is thought to influence exon recognition via co-transcriptional splicing. DNA methylation may modulate co-transcriptional splicing via the actions of methyl-binding domain proteins (MBDs). MBDs recruit histone-modifying enzymes to alter the neighboring chromatin, which could impair the RNA Pol II elongation rate [101, 102]. DNA methylation may occur with higher frequency in exons versus introns in human [94, 103-105] as well as in other species [95, 106-108]. It has been proposed that alternative exons have lower levels of DNA methylation compared to constitutive exons [104, 109]. However, recently Singer *et al* claimed that DNA methylation difference between introns and exons is biased by the non-uniform distribution of CpG associated with varying conservation rates between region types [110]. Thus, it is not yet clearly defined whether methylation plays a role in co-transcriptional splicing.

2.3.6 The role of introns in gene regulation

Introns are segments of a gene between exons, which are transcribed but do not participate in the production of the final protein product as they are excised from transcripts through pre-mRNA splicing, resulting in the formation of mRNA. Nevertheless, elevated sequence conservation among homologous introns of closely related species indicates functional constraints on intronic sequences during evolution [111]. Many studies have now showed a considerable range of biological functions carried out by introns. At the level of DNA, genomic introns are involved in transcription initiation, transcript termination and chromatin organization. Several studies found specific intronic DNA signals, mainly within the 5'-most introns, that regulate transcription initiation. These intronic DNA signals are enhancers [112-115], silencers [115-117] or others elements that modulate the function of the main upstream promoter [118, 119]. It was found throughout a wide range of species that the first introns, and particularly those in the 5’ UTR, are significantly longer than more distal introns within genes [120]. The accepted explanation of this finding is that these introns are longer because they harbor certain *cis* regulatory motifs related to transcription initiation [121]. It has been also shown that the 3’ -most intron carry functional DNA signals. The functional coupling between splicing, and the 3’ -most intron has been demonstrated in vitro [122]. Introns may also play a role in chromatin organization via their participation in the preferential nucleosome positioning on exons; in the proposed mechanism, introns harbor sequence elements near their ends which act as nucleosome disfavoring elements, pushing the nucleosomes away toward the exons [83].

At the transcriptome level, intron retention (IR) is involved in gene expression regulation. Coupled with nonsense-mediated mRNA decay (NMD), IR triggers the fine-tune regulation of gene expression [123, 124]. Increased IR level may reduce gene expression level through a bidirectional cross-regulation mechanism between localized RNA pol II accumulation and impaired splicing factor recruitment [124]. In the proposed mechanism, low level of expression results in the reduction of splicing factor recruitment to nascent transcripts promoting in turn high IR level and localized pausing of RNA pol II over retained introns; decreased RNA pol II elongation may further amplify IR level and ultimately transcript turnover. Despite numerous studies demonstrating the role of IR in gene expression regulation, little is known about how IR level might be involved in the regulation of exon inclusion.

2.4 Summary of the literature review

This literature review highlights the importance of transcription and pre-mRNA processing in the regulation of gene expression. Transcription and pre-mRNA processing work in concert to modulate the quantity and the nature of mRNA isoform a gene will produce. Remarkable progress made in the field has increased our understanding of the multiple, interdependent, mechanisms regulating AS. Chromatin compaction, transcriptional factors, RNA pol II elongation rates, histone modifications and nucleosome positioning, have emerged as key players in AS regulation. Although much excitement has been generated by those recent findings, there is a need for more careful genome-scale investigation of epigenetic and sequence features associated with pre-mRNA processing. Specifically, we need to address two main questions: *(i)* do the findings from focused experimentation on individual genes generalize to the genome scale, and *(ii)* does the relationship between epigenome modifications and pre-mRNA processing still hold after correcting for the correlation between transcription and splicing?

Chapter 3: Confounding effects of gene expression and epigenetic mark distribution in the analysis of splicing

3.1 Introduction

Evidence for coupling between transcription and splicing indicates that for some genes the speed of transcription and/or the activity of transcription factors affect the efficiency of splicing [72-74, 77]. This implies that factors, epigenetic marks for instance, that affect transcription could also affect splicing. Given that epigenetic marks are generally correlated with gene expression levels [91, 125-129], we argue that the relationship between splicing and histone modifications can be confounded by the relation between gene expression and histone modifications. Despite the landscape of insights in the field, this issue has not been fully resolved by prior studies analyzing the relationship between splicing and epigenetic marks or other variables of interest. Our goal is to determine whether the relationship between transcription levels and pre-mRNA processing efficiency – which has been described in some targeted studies – can be detected at a genome-wide level. If so, this interdependence has the potential to confound the relationship between splicing and other variables of interest, such as epigenetic modifications, and would need to be accounted for in further analyses.

In the following analyses, we used a subset of the McGill epigenome mapping centre (EMC) dataset spanning numerous projects that aim to integrate sequence-based variation with multiple levels of epigenetic and transcriptional regulation across the genome of human tissues and animal disease models [130]. This subset consists of (*i)* nine muscle samples and *(ii)* two purified blood cell populations obtained from 28 normal Swedish individuals, for a total of 37 CD14$^-$ CD4$^+$ T-cell samples and 37 CD14$^+$ monocyte samples. Transcriptome (total RNA sequencing), methylation (whole genome bisulfite sequencing) and six HMs (ChIP-Seq) profiles were collected from these samples. The six HMs consist of: trimethylated histone H3 at lysine 36 (H3K36me3), associated with transcribed regions [131]; monomethylated histone H3 at lysine 4 (H3K4me1), associated with enhancer regions [132]; trimethylated histone H3 at lysine 4 (H3K4me3), associated with promoter regions [132]; acetylated Histone H3 at lysine 27 (H3K27ac), associated with increased activation of enhancer and promoter regions [133-135]; H3 lysine 27 trimethylation (H3K27me3), associated with polycomb repression [131]; and H3 lysine 9 trimethylation (H3K9me3), associated with heterochromatin regions [136]. This comprehensive dataset provides great opportunities to study the basic relationships between the epigenome and the transcriptome. It is important to note that total RNA-seq - the type of data used here as well as in most current transcriptomic studies - measures both the pool of nascent partially processed transcripts and mature polyadenylated RNA. This has particular impact on the analysis of IR, as one cannot discriminate between high IR level as the consequence of incomplete or delayed removal and introns that are retained in the mature mRNA. Since it is known that total RNA-Seq detects a considerable amount of partially processed heteronuclear pre-mRNA [137], and that productive IR events are rare, we expect that, when profiling IR we are predominantly detecting introns which have not been completely removed. In the present chapter, we show that *(i)* the inter-correlation between levels of gene expression, splicing and epigenetic marks and *(ii)* the distribution of epigenetic marks across gene bodies, have confounding effects that need to be corrected when analyzing the association between splicing and variables of interest.

3.2 Methods

3.2.1 Data

Data profiling DNA methylation, six histone modifications (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3 and H3K9me3), and gene expression in 37 T lymphocytes, 37 monocytes and nine muscle human cells were obtained from the McGill Epigenome Mapping Center [130]. The methylation data is derived from whole genome bisulfite sequencing, the histone modification data is from chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-Seq) experiments, and the mRNA data is from high-throughput sequencing of total RNA.

3.2.2 Data pre-processing and filtering

Sequence data were generated using the Illumina HiSeq 2000/2500 and were processed using Illumina's CASAVA 1.8 software. Data preprocessing and primary filtering of ChIP-Seq, RNA-seq and whole genome bisulfite-seq are described below (Appendix Table 0.1 - Table 0.3). Sequencing data are available at the website of the McGill Epigenomics Mapping Centre (EMC): http://epigenomesportal.ca/.

3.2.2.1 Chip-Seq

Adaptor sequences and low quality score bases (Phred score < 30) were first trimmed using Trimmomatic v. 0.22 [138], and reads less than 32 bp long were then discarded. The resulting reads were mapped to the human reference genome (GRCh37/hg19) using bwa v. 0.6.1 [139]. Peaks were called with MACS v. 2.0.10.07132012 [140] using input IP as a control.

3.2.2.2 Whole genome bisulfate sequencing

The pipeline was implemented as previously described in Johnson's paper [141]. Adaptor sequences and low quality score bases (Phred score < 30) were first trimmed using Trimmomatic v. 0.22 [138], and reads less than 32 bp long were then discarded. Filtered reads and the reference genome undergo C->T and G->A nucleotide conversions before alignment. Reads were aligned per sequencing lane to the reference genome using bwa v. 0.6.1 [139]. Overlapping 3' read ends were clipped using nxtgen-utils v. 0.12.2 [142]. Lane bam files were merged and duplicates marked using Picard v. 1.77 [143]. Multiple filtering steps were performed using nxtgen-utils: duplicates were removed, only properly paired mates with mapping qualities > 20 were kept, reads with more than 2% mismatches were discarded and only mates with proper orientation were kept. An mpileup file was generated using samtools v. 0.1.18 [144]. Methylation calls were obtained using nxtgen-utils from the mpileup file. BisSNP v0.82.2 [145] was run on the filtered bam files.

3.2.2.3 RNA-Seq

Adaptor sequences and low quality score bases (Phred score < 30) were first trimmed using Trimmomatic v. 0.22 [138], and reads less than 32 bp long were then discarded. The resulting reads were mapped to the human reference genome (GRCh37/hg19) using Tophat v. 2.0.10 [146] and bowtie v. 2.1.0 [147]. We used the bedtools suite [148] to obtain raw mapping-read counts or average read count per nucleotide on whole genes, exons, introns and other features of interest, based on the gencode v19 gene annotation which was downloaded from the UCSC genome browser [149]. Transcript isoforms from single gene were merged using bedtools into 38175 consensus genes. Differential gene expression analyses were done using DESeq [150].

3.2.3 Exon inclusion estimation

We discarded exons shorter than 20 bp because of difficulties in aligning reads to short exons. To characterize exon inclusion, we considered any exon with the two consecutive flanking exons (upstream-exon|exon|downstream-exon, defined as an exon trio, Figure 3.1) and the exon-exon splice junctions derived from the splicing of these exons. Read coverage spanning only exon junctions was used. For each exon trio, we counted the number of RNA-Seq reads supporting the inclusion of the exon (reads mapping to the upstream-exon|exon junction and to the exon|downstream-exon junction) and the number of reads supporting the skipping of the exon (reads mapping to the upstream-exon|downstream-exon junction). The read coverages from the two junctions supporting the inclusion were averaged. The exon inclusion ratio was defined as the inclusion read-coverage divided by the total reads (exclusion read-coverage + inclusion read-coverage). The value of this ratio is between 0 and 1; the closer to 0 the lower the exon inclusion.

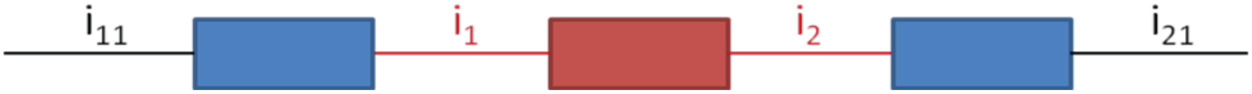


Figure 3.1: Exon trio design.

Exon-trio design for any exon of interest (red box). When comparing IR levels in the vicinity of the exon (i_1, i_2) with IR levels more distant (i_{11}, i_{12}), i_1 and i_2 levels were averaged as well as i_{11} and i_{21} levels.

3.2.4 Estimation of IR level

Each putative intron was delineated by the adjacent 5’ and 3’ exons. We selected introns with no overlap with annotated exons either in the same or a different annotated gene. A minimum overlap of 20 nt was required between counted reads and each intron involved to avoid misalignment reads with most part mapping to the adjacent exons. IR level was estimated by the average of read count per nucleotide of the intron divided by the sum of the average of read count per nucleotide of exonic regions of the host gene and the average of read count per nucleotide of the intron.

3.2.5 Estimation of the histone modification levels of genes

HM levels in gene were determined using the RPKM metric, i.e, Reads Per Kilobase of transcript per Million mapped reads [151], the ChIP-Seq RPKM signals were normalized (ratio) by the RPKM of corresponding control input DNA. Input DNA is the DNA that has been cross-linked and sonicated but without any specific immunoprecipitation.

3.2.6 Estimation of CpG density and DNA methylation levels

The percentage of CpG sites was calculated at each 5-bp by dividing the number of CpG by the total number of input sequences. Methylation level of the cytosine at each CpG site was measured as the ratio of methylated read-coverage to total read-coverage. The normalized DNA methylation level across multiple sites was determined by the ratio of methylated CpG to CpG density.

3.3 Results

3.3.1 Exon inclusion and IR levels are correlated with gene expression level

To examine the extent to which alternative splicing is coupled with transcription, we tested whether there is a transcriptome-wide correlation between gene expression and exon inclusion or IR. Using genome-wide gene expression data in monocytes and T cells, we compared exon inclusion and IR levels between the top 1000 highly and lowly expressed genes (nominal expression cut-off ≥ 10 read counts), simply referred to as highly and lowly expressed genes. Exon-inclusion levels were estimated using exon-trio models, which consist of the exon of interest and the two flanking exons upstream and downstream (Figure 3.1; see methods). We found, in general, that exons that belong to highly expressed genes have higher levels of inclusion than exons that belong to lowly expressed genes (monocytes P = 6.81×10^{-7} and T cells P = $5.10\ \times10^{-4}$, one-sided Student's t-test) (Figure 3.2-A). As expected, lowly expressed genes are more enriched in alternative exons (exon inclusion level ≤ 0.75) compared to highly expressed genes (monocytes P = 3.81×10^{-4} and T cells P = 2.39×10^{-3}, one-sided Fisher exact test).

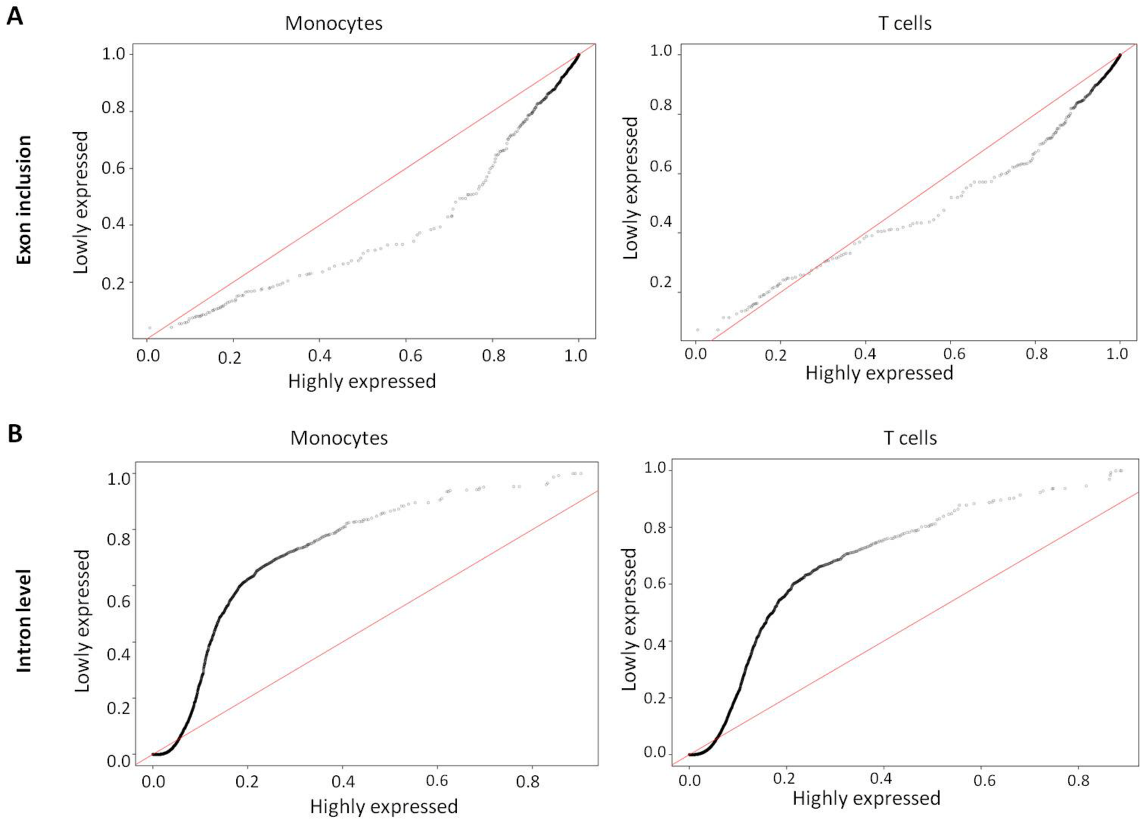


Figure 3.2: Quantile-quantile-plots comparing the distribution of levels of exon inclusion and IR between the top 1000 highly and lowly expressed genes.

(A) The figures were generated based on the exon inclusion levels of the 1000 highly and lowly expressed genes. Each point corresponds to the same quantile for each data set and shows the exon inclusion level at highly expressed genes versus lowly expressed genes at that quantile. The QQ-plots show that levels of exon inclusion within highly expressed genes (*x* axis) are stronger compared to lowly expressed genes (*y* axis). (B) The figures were generated based on the IR levels of the 1000 highly and lowly expressed genes. Each point corresponds to the same quantile for each data set and shows the IR level at highly expressed genes versus lowly expressed genes at that quantile. The QQ-plots show that retention levels of intron within highly expressed genes (*x* axis) are weaker compared to lowly expressed genes (*y* axis).

Moreover, differentially included exons between monocytes and T cells (inclusion level difference ≥ 0.1 and FDR corrected t-test P < 0.05) showed a positive correlation between exon inclusion difference and gene expression fold-change (r = 0.14 and P = 3.84×10^{-4}, Pearson correlation) (Figure 3.3). We next performed a similar analysis on the estimated levels of IR (see methods). Consistent with previous reports [124], we found that lowly expressed genes have significantly higher IR levels compared to highly expressed genes (monocytes $P = 9.21\times10^{-82}$ and T cells $P = 6.89\times10^{-88}$, one-sided Student's t-test) (Figure 3.2-B), and are more enriched in retained introns (IR level ≥ 0.1; monocytes $P = 1.41\times10^{-196}$ and T cells $P = 1.31\times10^{-91}$, one-sided Fisher exact test). These results support a model that transcription level does indeed influence splicing outcomes.

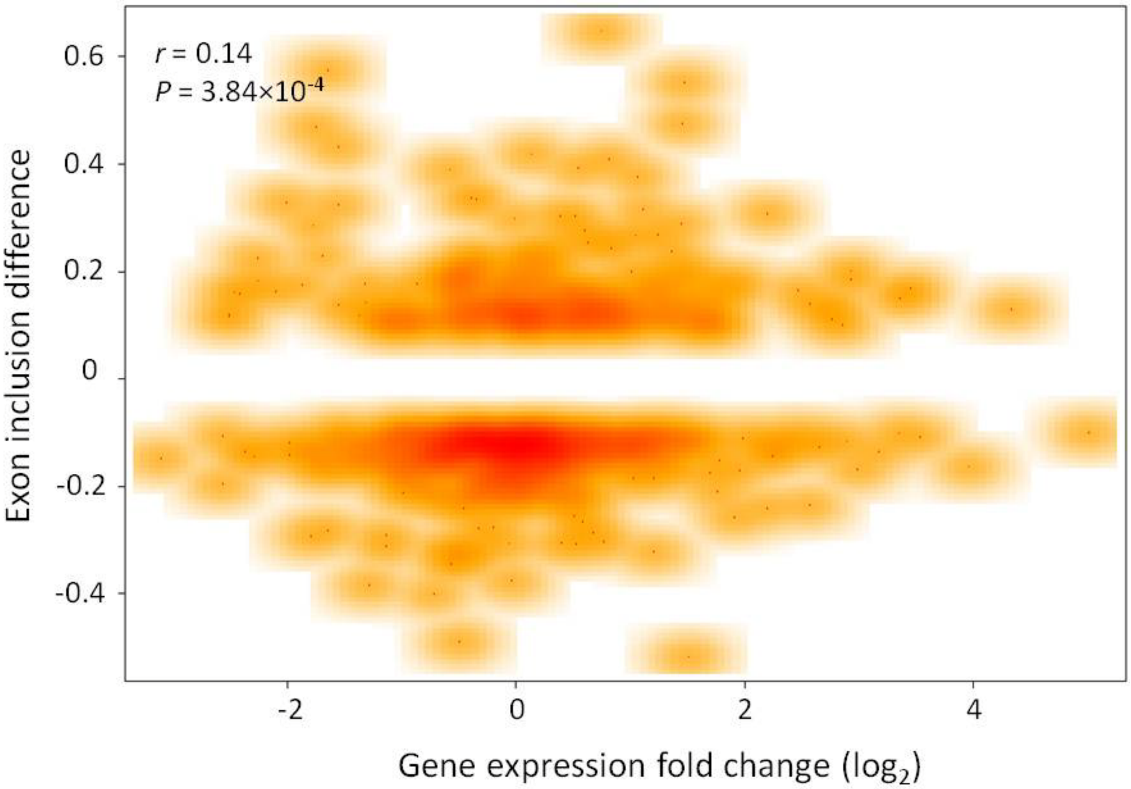


Figure 3.3: Gene expression variation and exon inclusion variation are positively correlated. Scatter plots comparing exon inclusion difference (*y* axis) to gene expression fold-change (*x* axis) of the differentially included exons between monocytes and T cells. The spearman correlation coefficient and its significance (p-value) are given on the top.

3.3.2 Genome-wide correlation between exon inclusion level and neighboring intron retention level

We extended our analysis on a genome-wide scale by examining the relationship between exon inclusion level and the retention level of neighboring introns. Specifically, we compared the retention level of introns flanking the exon with the retention level of introns more distant (Figure 3.1), using the student test in monocytes and T cells separately. Exons were divided into four categories according to their inclusion level: inclusion level ≤ 0.5; 0.5 ≤ inclusion level < 0.9; 0.9 ≤ inclusion level < 1; and inclusion level = 1. The same analysis was performed for each category on the means (across biological replicates) of IR levels flanking the exon versus the means of IR levels more distant. The fold-changes in IR levels were also computed. We found that alternative exons were strongly associated with higher retention level of surrounding introns compared to constitutive exons (Table 3.1; Figure 3.4). We observed that as exon inclusion ratio increases, the fold-change between the IR levels at introns flanking the exon and introns located more distantly decreased (Figure 3.5). These observations indicate that IR occurs at the proximity of alternative exons and is negatively correlated with the exon inclusion level.

Table 3-1: Summary of the genome-wide comparison of IR levels adjacent to exons with IR levels more distant.

Range of exon inclusion level[1]	Monocytes[2]			T cells[3]		
	Fold-change ($\log_2$)[4]	Student t-test p-value[5]	Sample size[6]	Fold-change ($\log_2$)[4]	Student t-test p-value[5]	Sample size[6]
]0-0.5[	0.60	6.43×10^{-4}	68	0.64	2.16×10^{-4}	71
[0.5-0.9[	0.42	4.42×10^{-12}	514	0.39	1.24×10^{-11}	487
[0.9-1[	0.11	1.68×10^{-14}	11337	0.11	1.19×10^{-16}	10439
[1]	-0.16	3.68×10^{-25}	11497	-0.14	8.35×10^{-26}	12535

For each range of exon inclusion[1] and cell type (monocytes[2] and T cells[3]), we indicated the $\log_2$ fold-change between the retention level of introns adjacent to exons and the retention level of introns located more distantly[4], the significance (t-test p-value) of differences in retention levels between adjacent introns and introns located more distantly[5], and the sample size of the t-test[6]. Ranges of exon inclusion level:]0-0.5[, exon inclusion level < 0.05; [0.5-0.9[, 0.5 ≤ exon inclusion level < 0.9; [0.9-1[, 0.9 ≤ exon inclusion level < 1; [1], exon inclusion level =1.

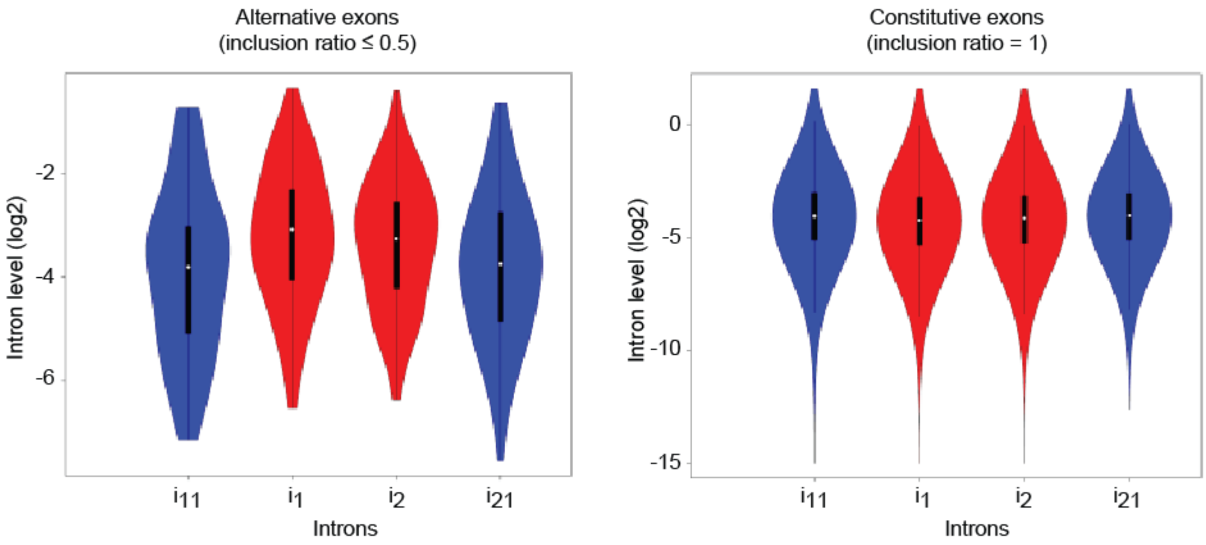


Figure 3.4: Distribution of IR levels in the neighborhood of alternative exons compared to constitutive exons.

Distribution of IR levels in $\log_2$ scale (*y* axis) in the neighborhood of alternative exons (inclusion level ≤ 0.5) compared to constitutive exons (inclusion level = 1). Each point on the *x* axis corresponds to an intron proximal to the exon of interest (see Figure 3.1).

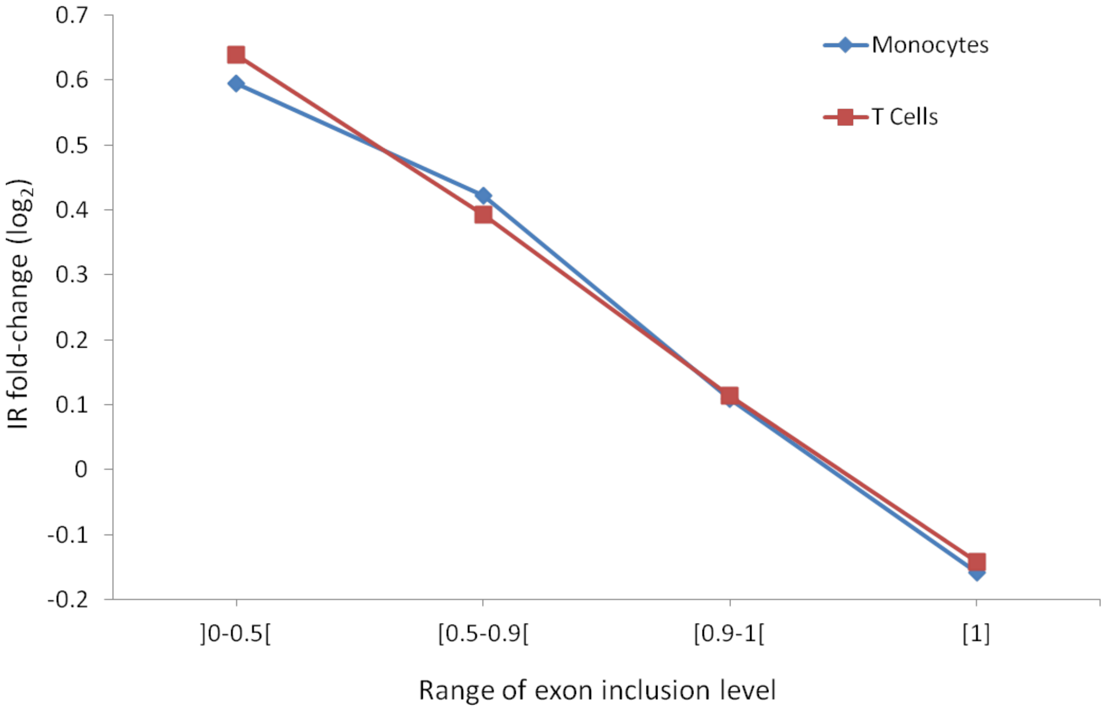


Figure 3.5: Fold change between the retention level of introns adjacent to the exon and the retention level of introns that are located more distantly.

Fold change between the retention level of introns adjacent to the exon (i_1, i_2) and the retention level of introns that are located more distantly (i_{11}, i_{21}). The *x* axis indicates the range of exon inclusion level:]0-0.5[, exon inclusion level < 0.05; [0.5-0.9[, 0.5 ≤ exon inclusion level < 0.9; [0.9-1[, 0.9 ≤ exon inclusion level < 1; [1], exon inclusion level = 1. The *y* axis is the log2 fold-change between the retention level of introns adjacent to the exon and the retention level of introns that are located more distantly.

3.3.3 Histone modifications are associated with gene expression level and are not uniformly distributed across gene bodies

We next investigated how HMs are related to the regulation of gene expression and intron retention. We gathered ChIP-seq data for six HMs (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3 and H3K9me3) in muscle, monocytes, and T cells. We first assessed the distribution of each HM within the top 1000 highly and lowly expressed genes and all genes (gene bodies, 1000-bp upstream transcription start sites and 1000-bp downstream transcription end sites). As previously described, highly expressed genes carry higher levels of H3K36me3, H3K4me1, H3K4me3, H3K27ac while they carry lower levels of H3K27me3 and H3K9me3 (Figure 3.6, Figure 3.7 and Figure 0.1-Figure 0.4). This pattern is consistent across samples and cell types. Also, levels of H3K36me3 in the first exon are low and increase significantly in the second and subsequent internal exons [152, 153]. Accordingly, we found that H3K36me3 increases from the 5’ end to 3’ end of the gene. H3K4me1, H3K4me3, H3K27ac are enriched at promoter regions – consistent with their known role as marks of transcriptionally-active promoters and/or enhancers - and sharply decreased from the transcription start site.

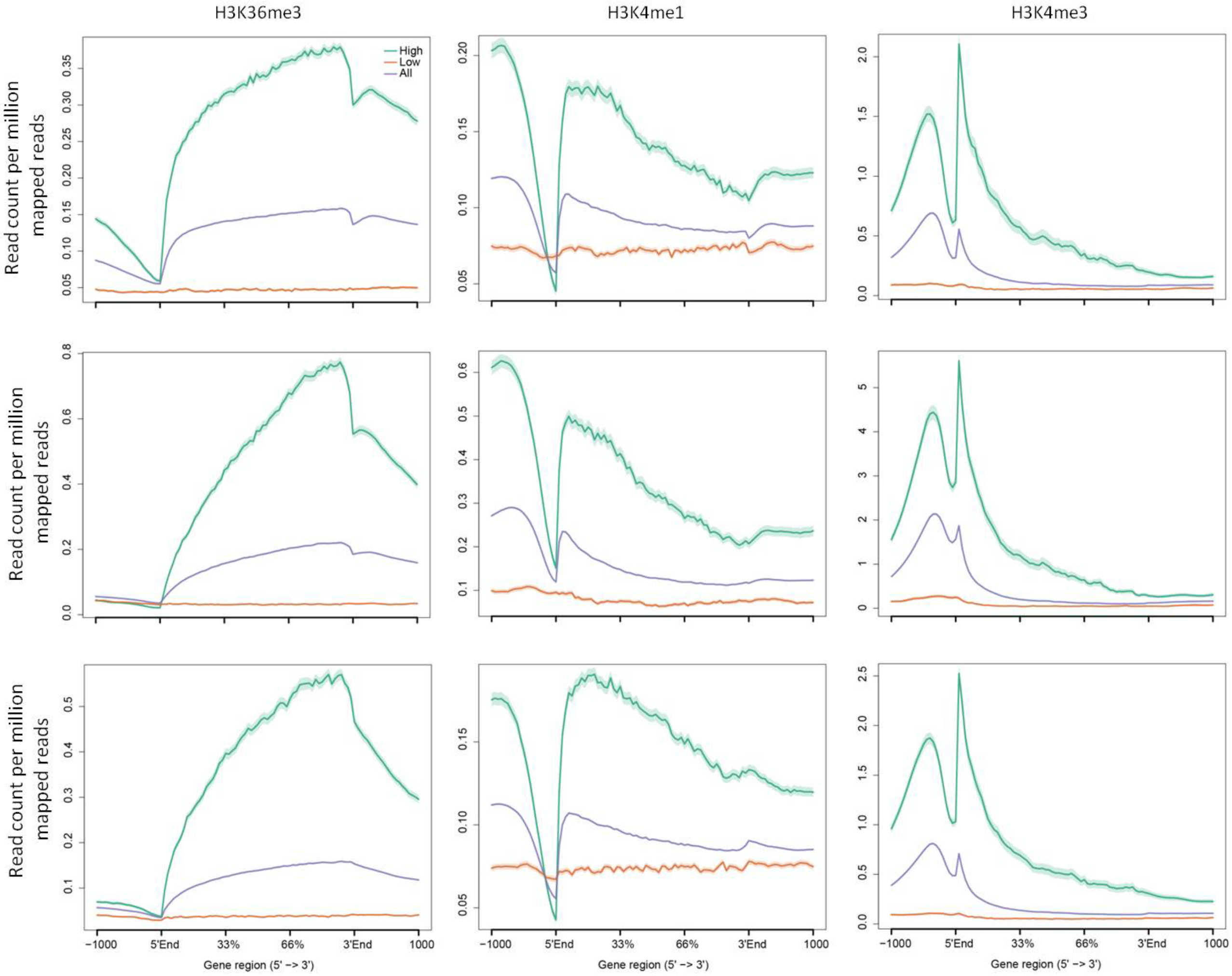


Figure 3.6: Composite plots of patterns of H3K36me3, H3K4me1 and H3K4me3 across genic regions, in monocytes.

H3K36me3, H3K4me1 and H3K4me3 are plotted across genic regions (*x* axis), of either the 1000 highly expressed genes (green), the 1000 lowly expressed genes (orange) or all genes (purple), in three monocyte samples. Data represented as read count per million mapped reads (*y* axis).

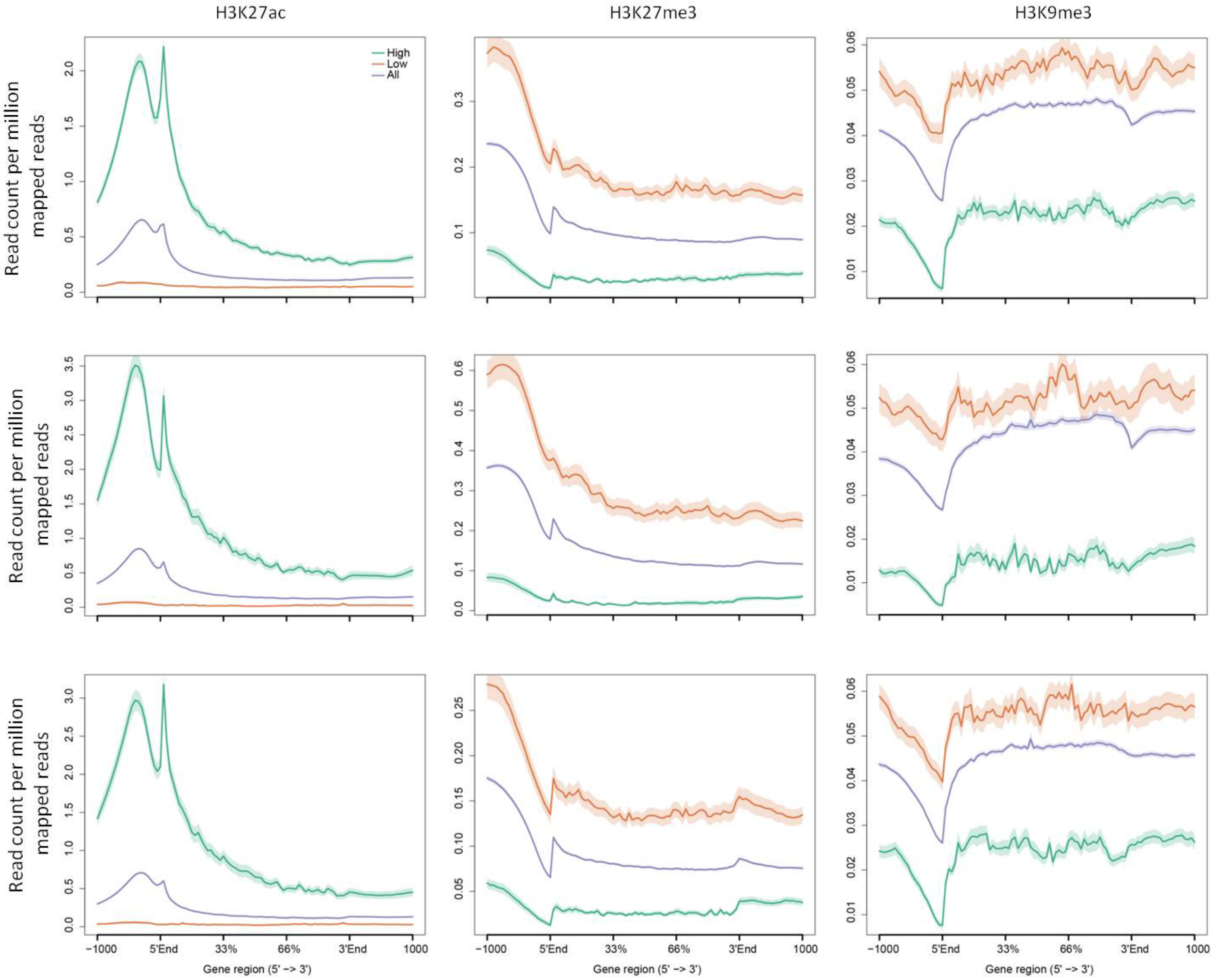


Figure 3.7: Composite plots of patterns of H3K27ac, H3K27me3 and H3K9me3 ChIP-seq signals across genic regions, in monocytes.

H3K27ac, H3K27me3 and H3K9me3 are plotted across genic regions (*x* axis), of either the 1000 highly expressed genes (green), the 1000 lowly expressed genes (orange) or all genes (purple), in three monocyte samples. Data represented as read count per million mapped reads (*y* axis).

To assess the association between gene expression and HMs, we performed gene-by-gene correlations across cell types between mRNA level of expressed genes and HM levels in the entire gene (see method). The summary of significant correlations (FDR corrected $P < 0.05$ and $|r| \geq 0.3$) is reported in Table 3.2. Interestingly, levels of H3K36me3, H3K4me3, H3K4me1 are positively correlated with gene expression (Table 3.2), but are negatively correlated with intron retention levels (Table 3.3), which suggest that these HMs are important for the regulation of both gene expression and intron removal. Given that gene expression and IR levels are negatively correlated, the influence of gene expression needs to be removed to target only the association between IR level and these HMs.

Table 3-2: Correlation of gene expression level with histone modification level

HM[1]	# Positive correlation[2]	# negative correlation[3]	Exact Binomial test p-value [4]
H3K36me3	606	89	2.06×10^{-95}
H3K4me3	449	340	1.18×10^{-04}
H3K4me1	909	132	3.04×10^{-143}
H3K27ac	14	22	2.43×10^{-01}
H3K27me3	25	37	1.62×10^{-01}
H3K9me3	35	31	7.12×10^{-01}

The number of significant positive correlations[2] and negative correlations[3] are indicated for each histone modification[1]. Differences in the occurrence of [2] and [3] were assessed with exact binomial test[4].

Table 3-3: Correlation of gene-intron level with histone modification level.

HM[1]	# positive correlation[2]	# negative correlation[3]	Exact binomial test pvalue [4]
H3K36me3	80	563	2.39×10^{-90}
H3K4me3	276	406	7.27×10^{-07}
H3K4me1	136	870	1.49×10^{-131}
H3K27ac	13	19	3.77×10^{-01}
H3K27me3	46	24	1.15×10^{-02}
H3K9me3	41	29	1.88×10^{-01}

The number of significant positive correlations[2] and negative correlations[3] are indicated for each histone modification[1]. Differences in the occurrence of [2] and [3] were assessed with exact binomial test[4]

3.3.4 CpG, HM and methylation levels of exons and introns are biased by their position in the gene

The distribution of intron positions in expressed genes showed that introns that are flanked by non-coding exons tend to be located in 5'UTR, hence close to the 5' end of the gene, compared to introns that are flanked by protein-coding exons (monocytes $P = 5.12\times10^{-128}$, T cells $P = 2.30\times10^{-111}$ and muscle $P = 9.29\times10^{-96}$; Mann-Whitney U test) (Figure 3.8). Furthermore, we found that introns flanked by non-coding exons show higher retention level compared to introns flanked by coding exons (Figure 3.9). This pattern is consistent with previous studies which report that IR level is higher in UTR regions [154] .

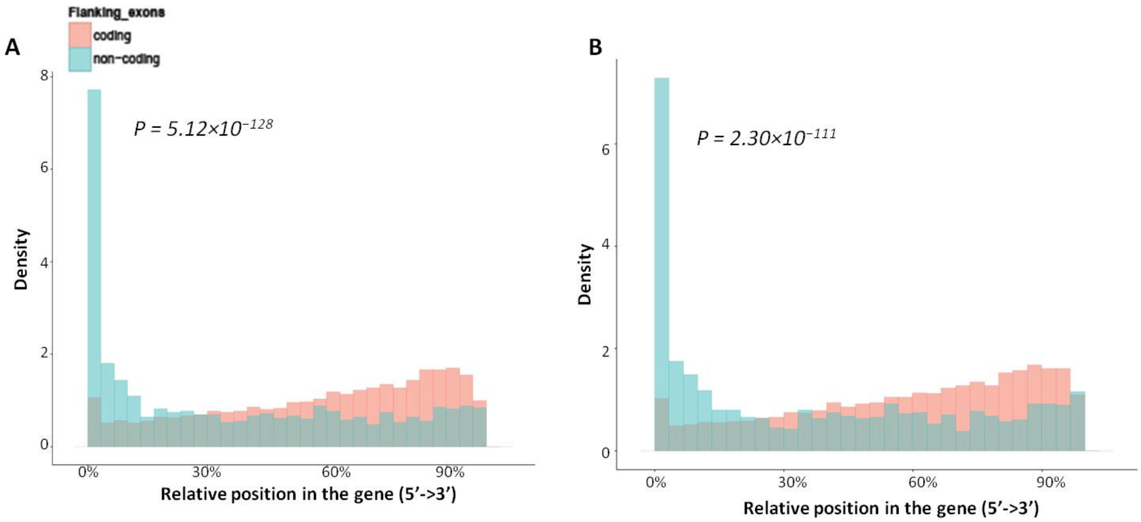


Figure 3.8: Distribution of intron positions in the gene.

Distribution of the relative positions of introns flanked by coding exons (pink) and introns flanked by non-coding exons (blue), in the gene, in (A) monocytes and (B) T cells. The significance (p-value) of the Mann-Whitney U test testing for differences in the relative positions is shown on top.

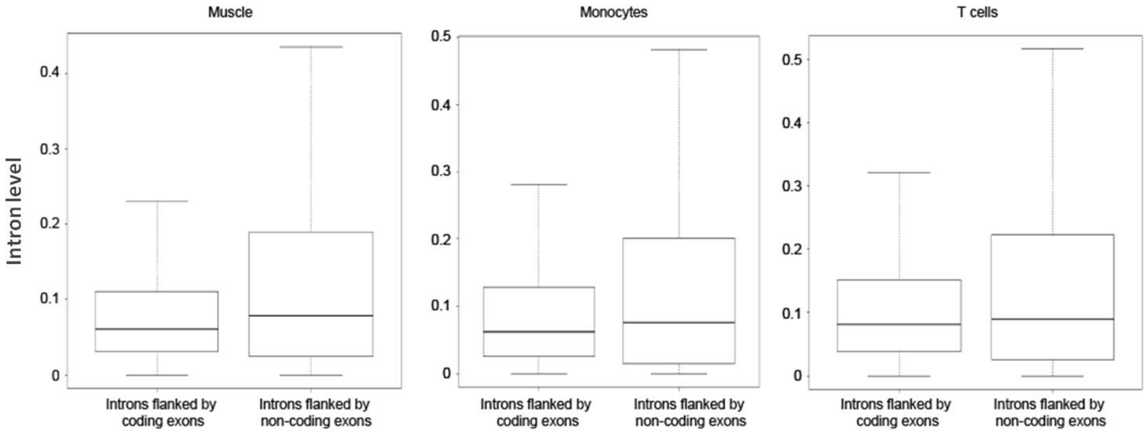


Figure 3.9: Distribution of IR levels in muscle, monocytes and T cells.

The *x* axis represents the intron category (intron flanked by coding exons or intron flanked by non-coding exons) and the *y* axis indicates the IR level.

To assess the influence of intron/exon positions on CpG and methylation distribution, we compared CpG and methylation between introns flanked by coding exons and introns flanked by non-coding exons originating from expressed genes. As expected, introns flanked by non-coding exons tend to harbor significantly higher CpG levels since they are closer to promoters which are normally CpG-rich (Figure 3.10-A - Figure 3.12-A). In the same line, introns flanked by non-coding exons have significantly lower methylation levels (mCpG/CpG) since promoters of expressed genes are lowly methylated (Figure 3.10-C - Figure 3.12-C).

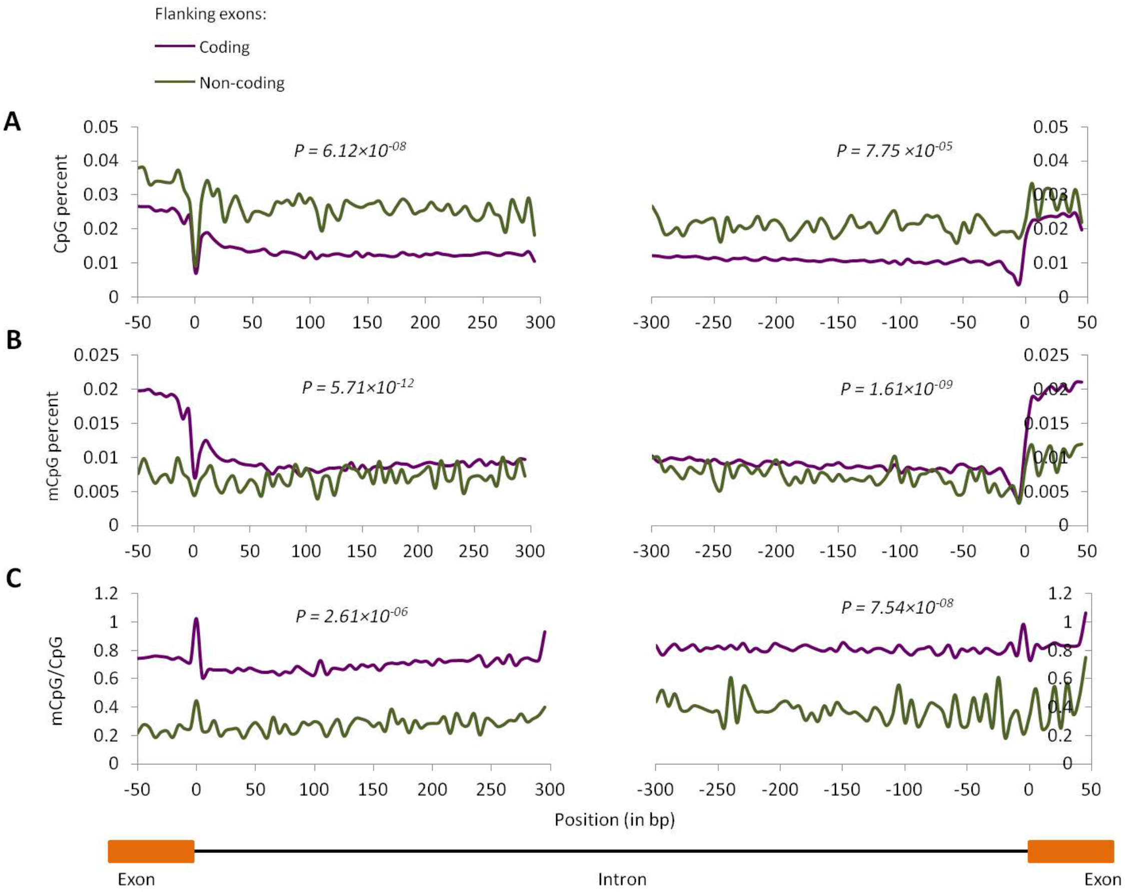


Figure 3.10: Differential CpG and methylation profiles between introns flanked by coding exons and introns flanked by non-coding exons, in monocytes.

(A) CpG density, (B) methylation levels, and (C) normalized-methylation levels in introns flanked by coding exons (purple) and introns flanked by non-coding exons (green). The significance (p-value) of the Mann-Whitney U test testing for differences in feature level between introns flanked by coding exons and introns flanked by non-coding exons is shown on top.

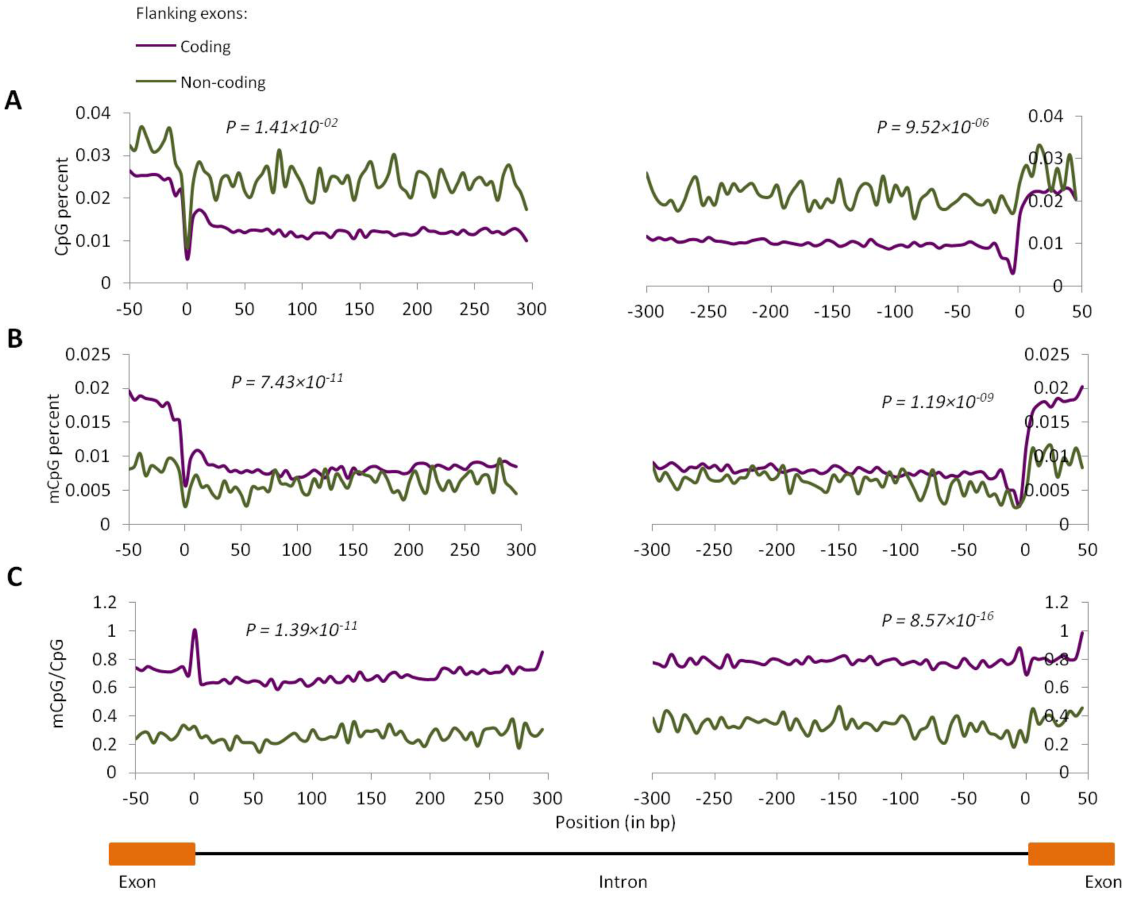


Figure 3.11: Differential CpG and methylation profiles between introns flanked by coding exons and introns flanked by non-coding exons, in muscle.

(A) CpG density, (B) methylation levels, and (C) normalized-methylation levels in introns flanked by coding exons (purple) and introns flanked by non-coding exons (green). The significance (p-value) of the Mann-Whitney U test testing for differences in feature level between introns flanked by coding exons and introns flanked by non-coding exons is shown on top.

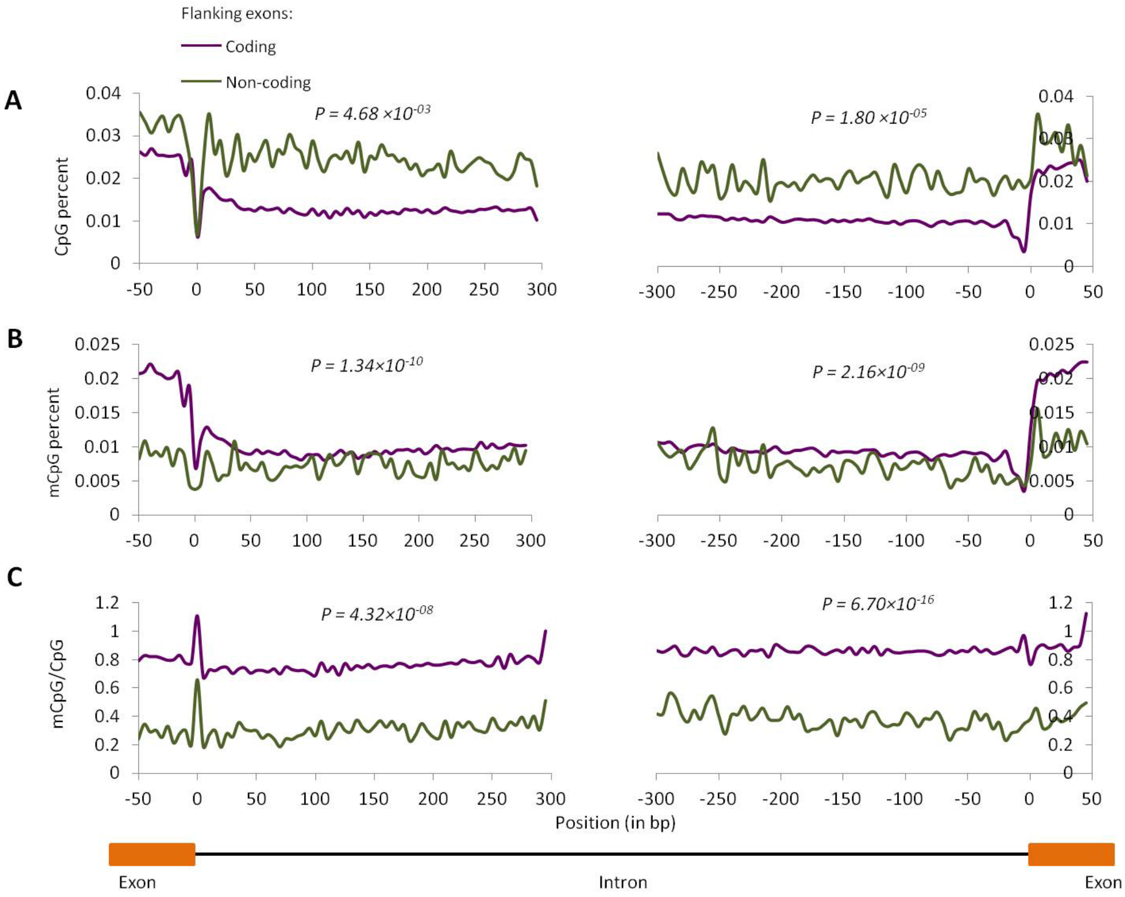


Figure 3.12: Differential CpG and methylation profiles between introns flanked by coding exons and introns flanked by non-coding exons, in T cells.

(A) CpG density, (B) methylation levels, and (C) normalized-methylation levels in introns flanked by coding exons (purple) and introns flanked by non-coding exons (green). The significance (p-value) of the Mann-Whitney U test testing for differences in feature level between introns flanked by coding exons and introns flanked by non-coding exons is shown on top.

Similarly, introns flanked by coding exons have significantly higher levels of H3K36me3 (Figure 3.13) but significantly lower levels of H3K4me1, H3K4me3 and H3K27ac (Figure 3.14) as the consequence of the non-uniform distribution of HMs across gene bodies. Our results show that the position of introns or exons can influence their CpG density and also their level of epigenetic marks.

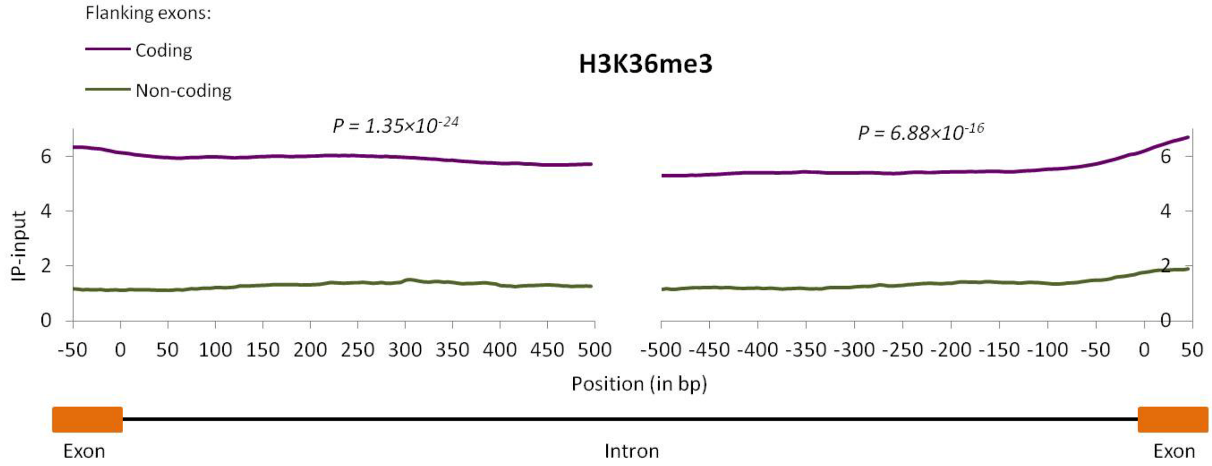


Figure 3.13: Differential H3K36me3 profiles between introns flanked by coding exons and introns flanked by non-coding exons, in T cells.

Average ChIP-Seq signals of H3K36me3 in the first and last 500bp of introns flanked by coding exons (purple line) and introns flanked by non-coding exons (green line). The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows. The significance (p-value) of the Mann-Whitney U test testing for differences in H3K36me3 level between introns flanked by coding exons and introns flanked by non-coding exons is shown on top.

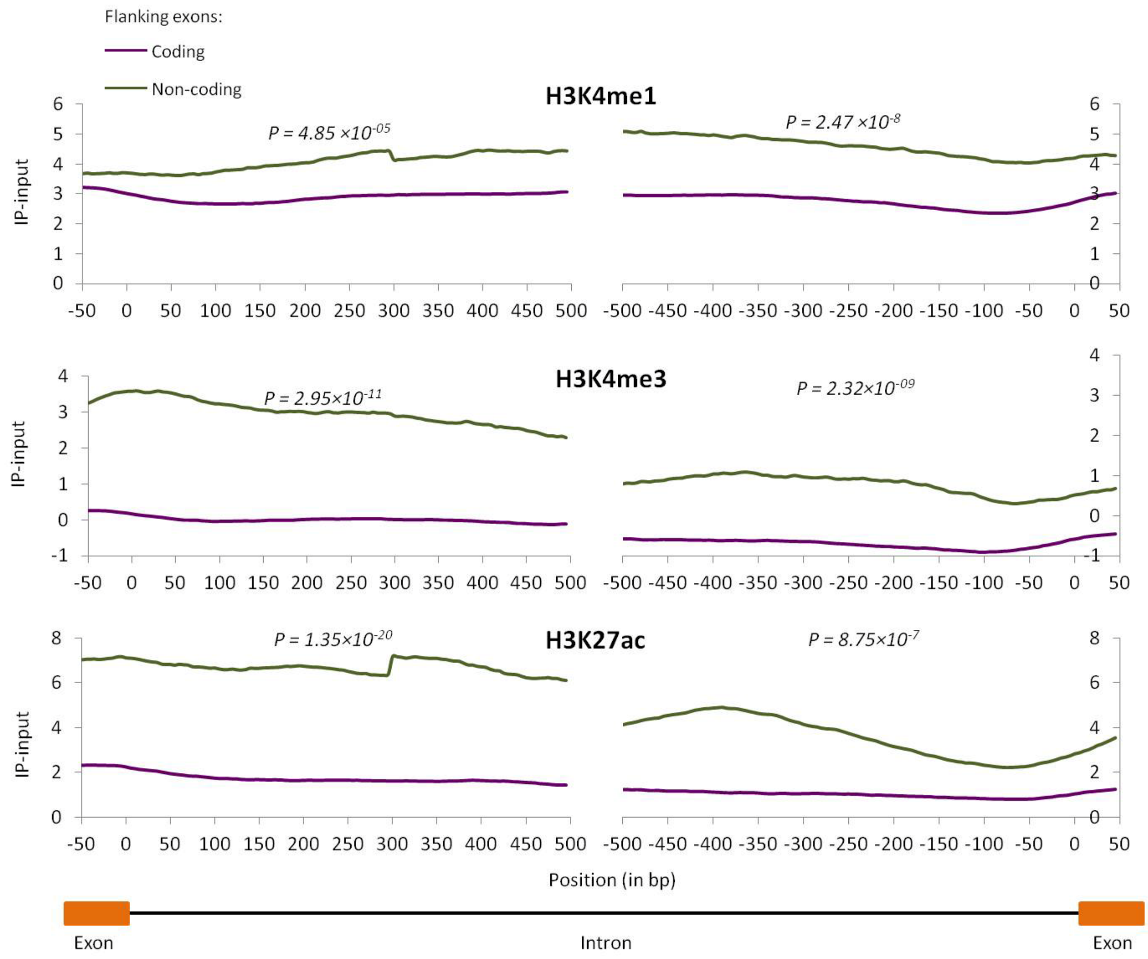


Figure 3.14: Differential H3K4me1, H3K4me3, H3K27ac profiles between introns flanked by coding exons and introns flanked by non-coding exons, in T cells.

Average ChIP-Seq signals of H3K4me1, H3K4me3, H3K27ac in the first and last 500bp of introns flanked by coding exons (purple line) and introns flanked by non-coding exons (green line). The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows. The significance (p-value) of the Mann-Whitney U test testing for differences in HM level between introns flanked by coding exons and introns flanked by non-coding exons is shown on top.

3.4 Correcting for confounding effects of gene expression levels and epigenetic mark distribution

Given the complex interplay between gene expression, HMs, splicing and IR level, it is essential to remove the direct effect of gene expression in order to study any potential effect of epigenome modifications on splicing and IR. We address this issue by employing a matched case-control approach. Specifically, when comparing the properties of alternative to constitutive exons, we select pairs of alternative and constitutive exons originating from the same gene. The same strategy is used to compare retained vs. non-retained introns. This approach ensures that within each such discordant pair both units experience the same transcriptional conditions on average in the cell population but differ only at the pre-mRNA processing level. In addition, since patterns of HMs and methylation are variable across the gene body, it is essential to avoid positional biases of cases and controls when comparing HMs and methylation between cases and controls. Therefore, to address this issue in our analysis of differential exon inclusion, we focused only on internal protein coding exons and we ensured that in general there is no difference in distribution of positions between alternative exons and constitutive exons (monocytes P = 0.37, muscle P = 0.07 and T cells P = 0.14, Mann-Whitney U test) (Figure 3.15).

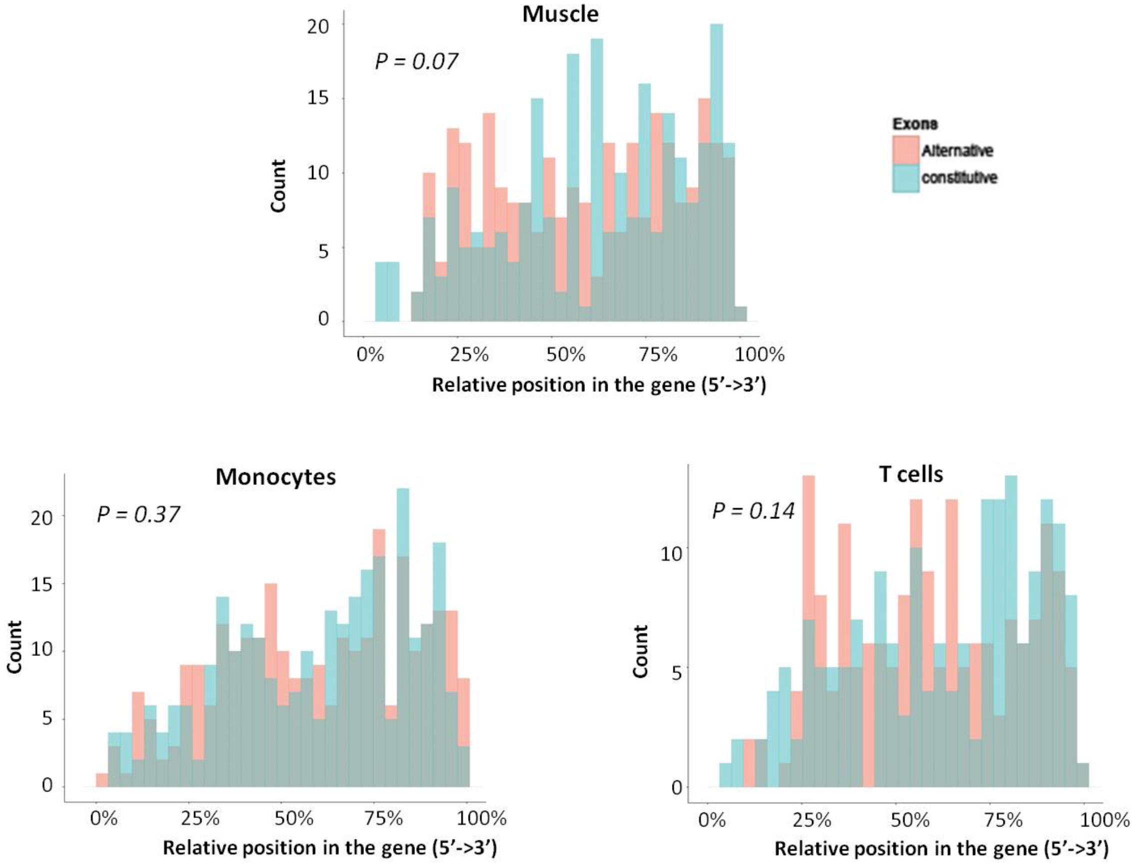


Figure 3.15: Distribution of exon positions in the gene.

Distribution of the relative positions of alternative (pink) and constitutive (blue) exons, in the gene, in muscle, monocytes and T cells. The significance (p-value) of the Mann-Whitney U test testing for differences in the relative positions is shown on top.

In our analysis of IR, we investigated introns flanked by non-coding exons (from UTRs or non-coding genes) and introns flanked by protein-coding exons separately since they have distinct profiles of epigenetic marks. We also ensured that within each category of introns, retained and non-retained introns do not have statistically significant differential positions in the gene: although the distributions of the positions of retained introns and non-retained introns are different, these differences are not large enough to be statistically significant (Mann-Whitney U test), and hence should have little effect on the downstream analyses (Figure 3.16 and Figure 3.17). Since the silencing marks H3K27me3 and H3K9me3 are not present in expressed genes at detectable levels (Figure 3.7), we didn't investigate them in the subsequent analyses. To minimize sample heterogeneity, we restricted subsequent analyses to the subset of samples that have matching profiles of RNA-Seq, methylation, H3K36me3, H3K4me1, H3K4me3 and H3K27ac. This subset consists of two monocyte, nine muscle and 14 T cell samples.

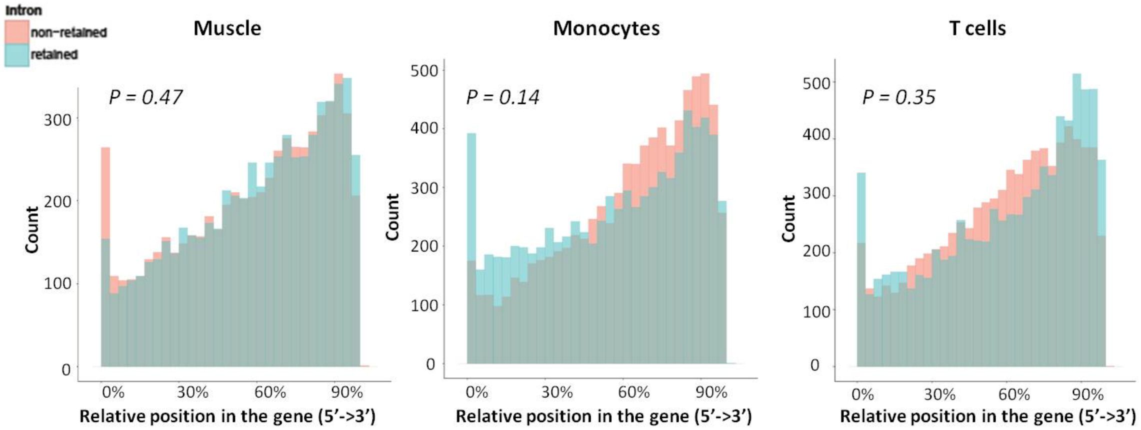


Figure 3.16: Distribution of the positions of retained and non-retained introns flanked by coding exons.

Distribution of the relative positions of non-retained (pink) and retained (blue) introns in the gene, in muscle, monocytes and T cells. The significance (p-value) of the Mann-Whitney U test testing for differences in the relative positions is shown on top.

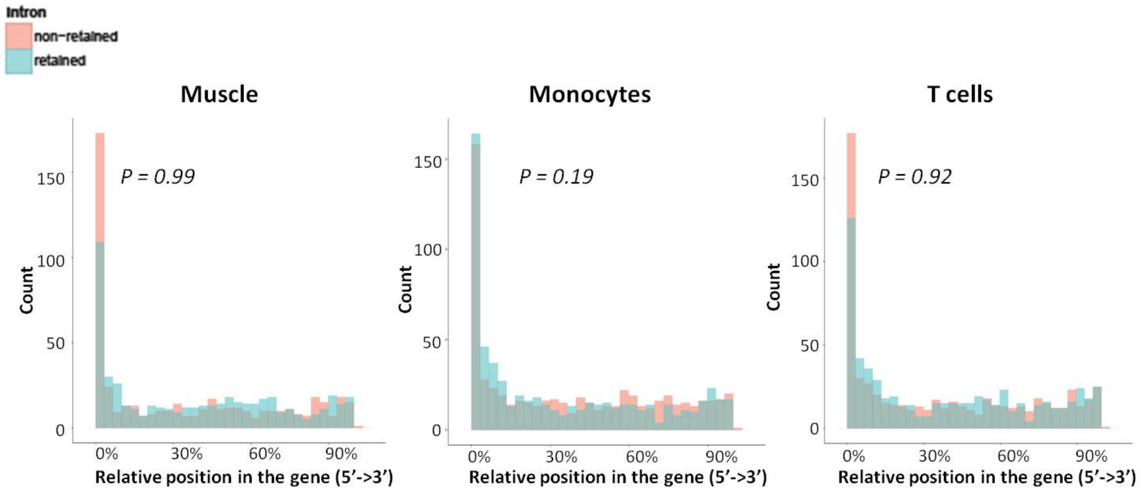


Figure 3.17: Distribution of the positions of retained and non-retained introns flanked by non-coding exons.

Distribution of the relative positions of non-retained (pink) and retained (blue) introns in the gene, in muscle, monocytes and T cells. The significance (p-value) of the Mann-Whitney U test testing for differences in the relative positions is shown on top.

3.5 Conclusion

Collectively our results suggest there is a relationship between IR, exon inclusion and gene expression. Consequently, to target only the association between exon inclusion, IR level and epigenetic marks, the influence of gene expression needs to be removed. Additionally, when comparing epigenetic marks between alternative and constitutive exons or between retained and non-retained introns, it is important to avoid epigenetic mark differences which are driven by positional effects.

Chapter 4: Association of exon inclusion with intron retention and epigenetic marks

4.1 Introduction

Alternative splicing (AS) is a widespread process by which one gene can be spliced differently to generate distinct mRNAs, leading to protein isoforms that may have distinct functions. AS events affect nearly 95% of mammalian genes [7-9], and the deregulation of this process affects the progression of various human diseases and cancers [155-158]. Cassette exon represents the most common AS event [154, 159-161]. Splicing is regulated by the interplay of cis-regulatory sequences and trans-acting factors, and, is coupled to transcription [14-17], which suggests that factors (such as IR and HMs) that regulate transcription may also influence AS. There is growing evidence that IR level is an important contributor to the regulation of gene expression in normal physiological functions [123, 162-165] as well as pathologies [166]. In contrast, relatively little is known about whether IR level is involved in the regulation of exon splicing. It has also been argued that chromatin remodeling, particularly the H3K36me3 mark, which is strongly associated with exonic sequences and positively correlated with exon inclusion level [96, 98], plays a part in the regulation of alternative splicing. Moreover, it is believed that DNA methylation may influence splice-site recognition, with higher methylation levels found in exons compared to introns in humans [94, 103-105] and other species [95, 106-108]. It has been further proposed that alternatively spliced regions may have lower methylation levels compared to constitutively spliced regions [104, 167].

Although there has been a considerable amount of research carried out to generate those hypotheses, in view of the confounding correlations discussed in the previous chapter, we postulate that there is a need for more careful genome-scale investigation of epigenetic and sequence features associated with pre-mRNA processing. Specifically, we need to address two main questions: (1) what remains of the association between exon inclusion and epigenetic features after the removal of the influence of transcription? (2) How intron retention is associated with exon inclusion regulation? We took advantage of the comprehensive transcriptome and epigenome data to identify large-scale differences in RNA splicing process. We used two monocyte, nine muscle and 14 T cell samples, each having matching whole-transcriptome, methylation and four HM profiles (H3K36me3, H3K4me1, H3K4me3 and H3K27ac). Here, we focused on exon inclusion, IR level in the vicinity, and the underlying epigenetic and sequence features that might be associated. Specifically, we measured the association between exon inclusion and IR level, HMs, methylation and sequence features. After correcting for gene expression and HM distribution, a number of previously identified features specific to exon inclusion were no longer significant: conversely to early studies, alternative and constitutive exons do not display differential methylation levels and 5’SS strength. However, our corrective approach confirmed a number of previously reported features associated with exon splicing. Accordingly, we found that alternative exons are associated with a higher retention level of flanking introns which further exhibited a higher evolutionary constraint compared to introns flanking constitutive exons. Interestingly, retained introns are predictive of alternative exons. Alternative exons show weaker 3’ SS. At the epigenome level, H3K36me3 levels are lower in alternative exons.

4.2 Materials and methods

4.2.1 Data

DNA methylation, four histone modifications (H3K4me3, H3K4me1, H3K27ac and H3K36me3) and gene expression in 14 T lymphocyte, two monocyte and nine muscle human cell samples.

4.2.2 Exon inclusion filtering

Exon inclusion level was estimated as described in the methods of chapter 3. We discarded UTRs exons and non-coding transcripts from the analyses. Our set of candidate alternative or constitutive exons is defined as internal exons that are not first, second, second to last, or last in any of the transcripts of the gencode v19 annotation. Exon inclusion ratios were then filtered according to the following critera: *(i)* the sum of inclusion and exclusion read-coverage must be equal or higher than 10; *(ii)* upstream and downstream exons must be constitutive (exon inclusion ratio > 0.75) with an average of read count per nucleotide equal or greater than 5, and *(iii)* a balanced count of reads mapping to the upstream-exon|exon and exon|downstream-exon junctions to discriminate alternative start or end events: given the total number of RNA-Seq reads supporting the inclusion of an exon (reads mapping to the upstream-exon|exon junction and to the exon|downstream-exon junction), the difference between the percentage of reads mapping to the upstream-exon|exon junction and the percentage of reads mapping to the exon|downstream-exon junction must be ≤ 75%. Using these criteria, we identified 7725 exon trios belonging to 2674 genes in muscle, 5064 exon trios belonging to 2400 genes in T cells, and 4902 exon trios belonging to 2647 genes in monocytes.

4.2.3 Paired alternative and constitutive exons within the same gene

We limited our analyses only to exons having flanking introns longer than 50 bp, and the flanking introns do not overlap any annotated exons. Exons were divided into two categories according to the magnitude of the exon inclusion level: alternative exons with inclusion ratio ≤ 0.75, and constitutive exons with 0.95 ≤ inclusion ratio ≤ 1. In each gene we selected one alternative exon and one constitutive, leaving us with a total of 248, 179 and 276 matching pairs of exons in muscle, T cells and monocytes, respectively.

4.2.4 Estimation and visualization of DNA methylation, HMs, CpG levels and conservation

The density of CpG and methylation levels were estimated as described in the methods of chapter 3. For HMs, DNA-input-subtracted ChIP-Seq read coverages were averaged at any 5-bp window across regions of interest. These averages were plotted directly without smoothing. To determine conservation, we used the Genomic Evolutionary Rate Profiling (GERP) scores [168, 169] downloaded from the UCSC platform [170, 171]. GERP identifies constrained elements in multiple alignments of 35 mammals to the human genome reference sequence (hg19) by quantifying substitution deficits at single nucleotide resolution. These deficits represent the predicted substitutions if the site was under neutral selection, but did not occur because the site has been under functional constraint. Sites with positive GERP scores ≥ 2 are considered to be under purifying selection, while sites with lower scores, including negative values, are considered as evolving neutrally.

4.2.5 Statistical test to compare HM, CpG, mCpG, mCpG/CpG and conservation levels between alternative and constitutive exons

To test the significance of the differences in HMs (ChIP-Seq read coverage minus the corresponding input-DNA coverage), CpG, mCpG, mCpG/CpG and conservation (GERP score) between alternative and constitutive exons within the same gene, for each exon pair we first calculated the mean of any feature in the 100-bp region which is centered on splice sites. Then, the resulting means value pairs were subjected to a paired Mann-Whitney U test.

4.2.6 Analysis of splice-site strength

The maxEntScan [172] was used to score splice site strength based on the calculation of maximum entropy for 3'SS and 5'SS. The maximum entropy takes into account adjacent as well as non-adjacent dependencies between nucleotide sequences of 9-mer (for 5'SS) or 23-mer (for 3'SS) corresponding to donor's and acceptor's flanking sequences. It assigns log-odd ratios to candidate human splice sites to be a true splice site using a maximum entropy score model. The higher the score, the higher the probability that the splice site is an actual one. Apart from maximum entropy (ME), the position-specific weight matrix (WM) and inhomogeneous 1st order Markov models (MM) were used concurrently to score the splice site strength.

4.3 Results

4.3.1 Exon skipping is associated with retention of adjacent introns

We aimed to investigate the association between exon inclusion and IR level. For each of the paired alternative and constitutive exons (see method), we compared the level of introns that are adjacent to the exon against the level of introns that are located more distantly (Figure 3.1). We observed an excess of alternative exons exhibiting a significantly higher IR in the neighborhood (t-test 0.05-level FDR *P*-value < 0.05, and the fold-change between the levels of adjacent introns and distant introns ≥ 2) (Figure 4.1-A). A representative example of such relationship is shown in Figure 4.1-B. We next determined the probability that retained introns and alternative exons are adjacent by random chance. We exploited the hypergeometric distribution to calculate *(i)* the probability of an alternative exon to be flanked by a retained intron more than constitutive exons do by chance, and *(ii)* the probability of a retained intron to be flanked by an alternative exon more than non-retained introns do by chance. We compared alternative and constitutive exons, within the same gene, for the enrichment of at least one flanking retained intron (IR level ≥ 0.1). We found that alternative exons are more likely to be directly flanked by a retained intron compared to constitutive exons (monocytes P = 2.52×10^{-02}, odds ratio (OR) = 1.50; muscle P = 5.79×10^{-02}, OR = 1.42; T cells P = 4.35×10^{-02}, OR= 1.67; one-sided Fisher exact test). Conversely, the comparison of retained and non-retained introns within the same gene revealed that retained introns are not more likely to be directly flanked by an alternative exon compared to non-retained introns (muscle P = 0.17, OR = 1.03 ; monocytes P = 0.25 , OR = 1.02 ; T cells P = 0.35 , OR = 1.01; one-sided Fisher exact test). This result indicates that retained introns are predictive of alternative exons, but alternative exons are not predictive of retained introns. Therefore, exons observed with a retained flanking intron commonly undergo AS; this indicates that IR in the vicinity of exons may have an influence on exon splicing.

A

	Muscle[1]				T cells[2]			
Exon type	(i$_1$,i$_2$) > (i$_{11}$,i$_{21}$)[3]	(i$_1$,i$_2$) < (i$_{11}$,i$_{21}$)[4]	Total[5]	Binomial p-value[7]	(i$_1$,i$_2$) > (i$_{11}$,i$_{21}$)[3]	(i$_1$,i$_2$) < (i$_{11}$,i$_{21}$)[4]	Total[5]	Binomial p-value[5]
Alternative exons	101	8	248	1.27×10^{-21}	61	21	179	1.13×10^{-05}
Constitutive exons	34	23	248	1.85×10^{-01}	33	20	179	9.84×10^{-02}

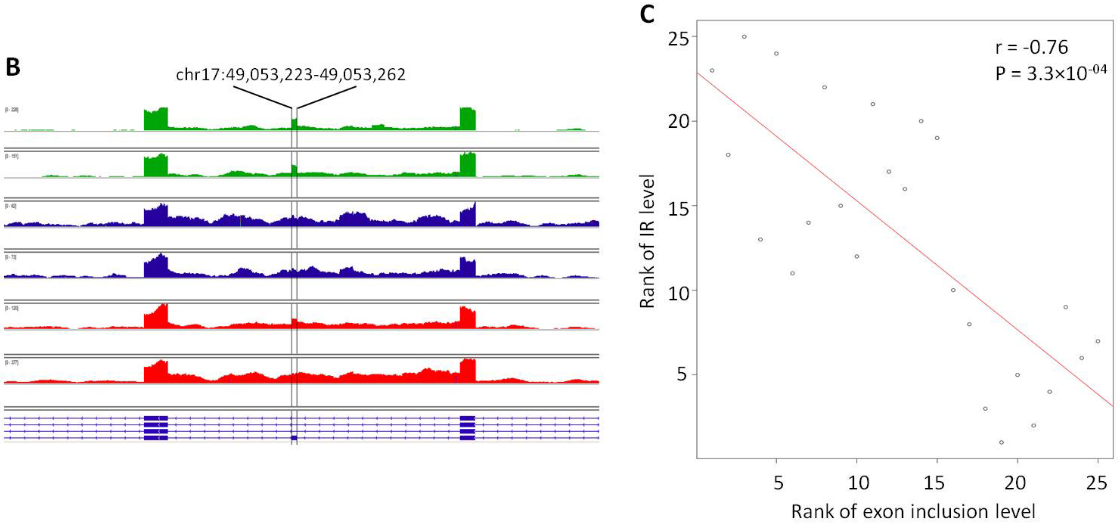


Figure 4.1: Association between exon inclusion and IR level in the vicinity.

(A)Summary of the comparison of the IR level adjacent to the exon and the IR level more distant, using a t-test in muscle and T cells separately. For each exon category and cell type (muscle[1] and T cells[2]), we indicated the number of comparisons where the retention level of introns flanking the exon is significantly higher than the retention level of introns more distant[3], the number of significant comparisons where the retention level of introns flanking the exon is lower than the retention level of introns more distant[4], and the total number of exons tested[5]. Differences in the occurrence of [3] and [4] were assessed with the exact binomial test[6]. (B) IGV (Integrative Genome Viewer [173]) screenshot depicting an example of higher retention level of introns flanking an alternative exon (exon 24, highlighted region) of gene SPAG9. RNA-seq reads mapping to SPAG9 are shown for two muscle (green), two T cell (blue) and two monocyte (red) samples. The exon 24 of *SPAG9* gene is included at 0.48, 0.12 and 0.30 in muscle, T cells and monocytes respectively;

IR levels within the trio exon23-exon24-exon25 are significantly (Student's t-test FDR-corrected P < 0.05) higher than IR levels flanking the trio (IR fold-change: muscle = 4.04, T cells = 2.23 and monocytes = 3.04). (C) Spearman correlation between the inclusion ratio (*x* axis) of exon 24 of SPAG9 and the retention level of flanking introns (average of the retention levels of introns 23 and 24; *y* axis). The regression line is shown in red. Data from muscle, monocytes and T cells were combined in this analysis.

We next measured the strength of association between the exon inclusion level and the retention level of adjacent introns, using the Spearman correlation in monocytes, muscle and T cells combined. 98 out of 361 exons showed a significant correlation (FDR-corrected $P < 0.05$) with a predominance of negative correlations (65 exons (66%), binomial test $P = 8.00\times10^{-4}$). This indicates that the more an exon is included, the lower the retention level of flanking introns. This is illustrated by exon 24 of gene *SPAG9* which shows a significant anti-correlation (Spearman correlation $P = 3.3\times10^{-04}$; r = -0.76) between its level of inclusion and the retention level of adjacent introns (Figure 4.1-C).

4.3.2 Alternative and constitutive exons do not have distinct CpG and DNA methylation levels

To investigate the potential effect of methylation on AS, we determined methylation levels in the 50 bp on either side of splice sites of alternative and constitutive exons; we calculated average normalized DNA methylation as the ratio of methylated CpG (mCpG) to CpG in order to control for CpG abundance since higher methylation levels can be a direct result of a higher frequency of CpG dinucleotides. Both alternative exons and constitutive exons showed a higher level of CpG (Figure 4.2-A, Figure 4.3-A and Figure 4.4-A) and mCpG (Figure 4.2-B, Figure 4.3-B and Figure 4.4-B) in exonic regions compared to intronic regions as proposed by previous studies [94, 95, 103-108]; but the difference in mCpG is lost when we normalized methylation values by CpG content to correct for sequence composition bias in our DNA methylation estimation (Figure 4.2-C, Figure 4.3-C and Figure 4.4-C). Interestingly, we observed that alternative exons and constitutive exons did not show differential DNA methylation (monocytes 3'SS *P* = 0.23, 5'SS *P* = 0.61; muscle 3'SS *P* = 0.33, 5'SS *P* = 0.17; T Cells 3'SS *P* = 0.43, 5'SS *P* = 0.28; paired Mann-Whitney U test) (Figure 4.2-C, Figure 4.3-C and Figure 4.4-C). Thus, after correcting for CpG abundance, positional effect of exons, and gene expression levels, our results do not support an association between CpG, DNA methylation and alternative splicing.

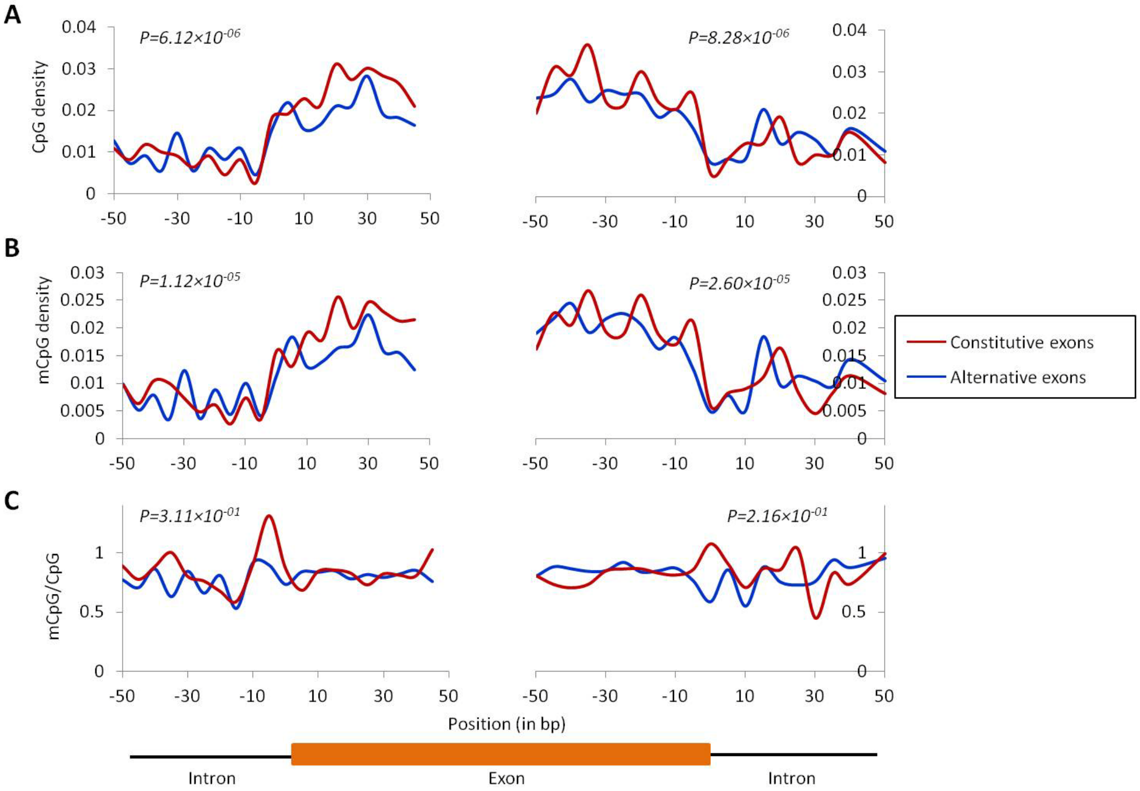


Figure 4.2: CpG and methylation profiles of alternative and constitutive exons in monocytes.

(A) CpG density, (B) methylation levels, and (C) normalized-methylation levels in ±50 bp (5-bp bin) around splice sites of alternative exons (inclusion level ≤ 0.75; blue line) and constitutive exons (inclusion level between 0.95 and 1; red line). The Mann-Whitney U test p-value testing for differences in feature level between exonic and intronic regions is shown on top.

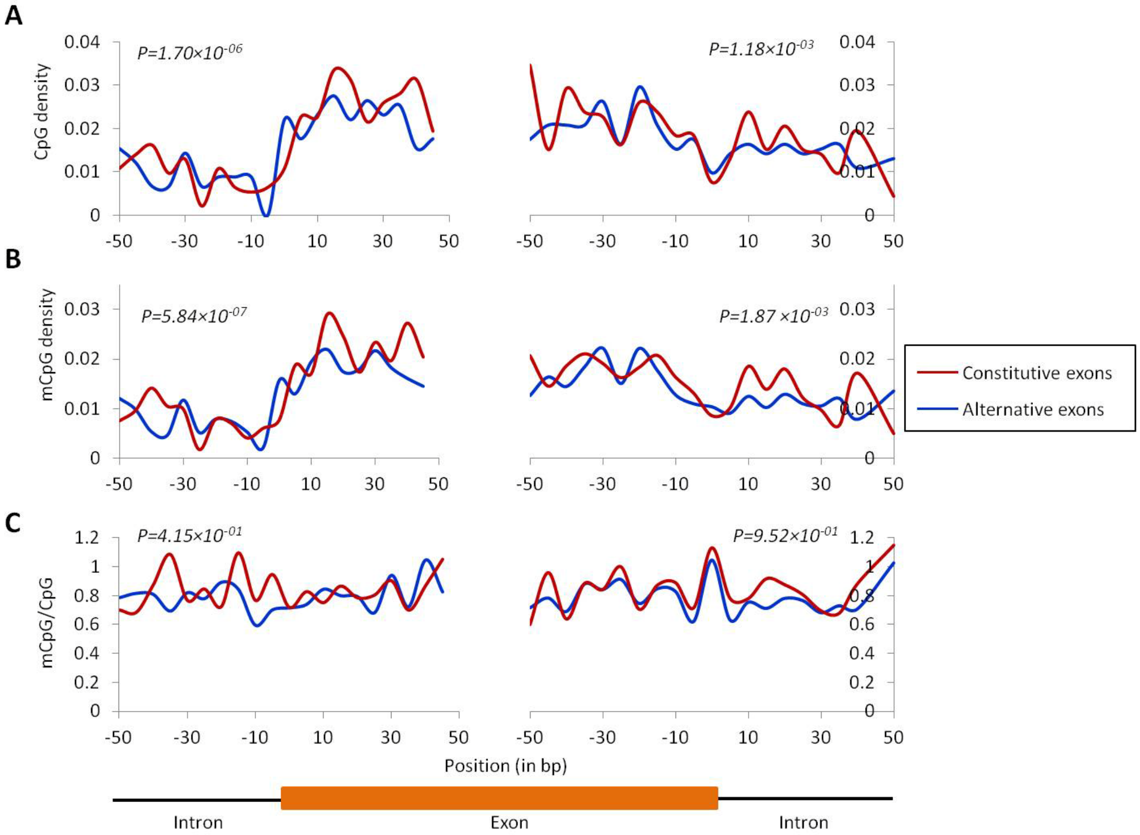


Figure 4.3: CpG and methylation profiles of alternative and constitutive exons in T cells.

CpG density, (B) methylation levels, and (C) normalized-methylation levels in ±50 bp (5-bp bin) around splice sites of alternative exons (inclusion level ≤0.75; blue line) and constitutive exons (inclusion level between 0.95 and 1; red line). The Mann-Whitney U test p-value testing for differences in feature level between exonic and intronic regions is shown on top.

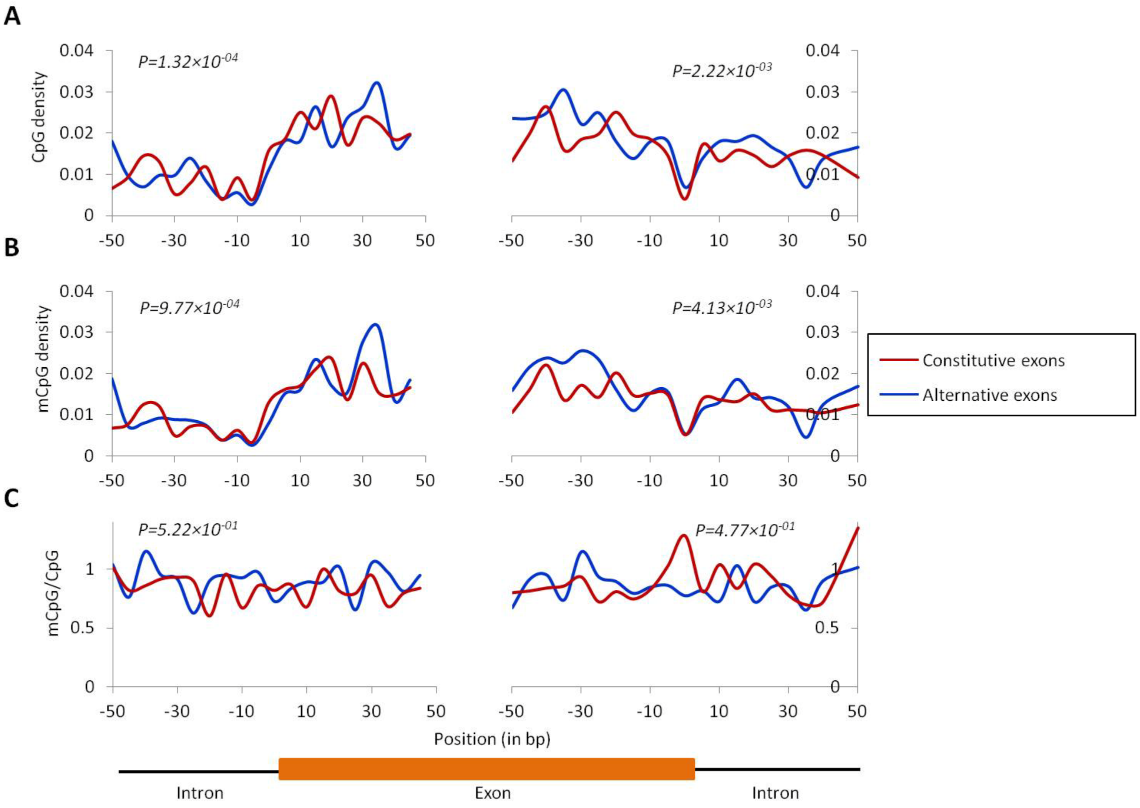


Figure 4.4: CpG and methylation profiles of alternative and constitutive exons in muscle.

CpG density, (B) methylation levels, and (C) normalized-methylation levels in ±50 bp (5-bp bin) around splice sites of alternative exons (inclusion level≤ 0.75; blue line) and constitutive exons (inclusion level between 0.95 and 1; red line). The Mann-Whitney U test p-value testing for differences in feature level between exonic and intronic regions is shown on top.

4.3.3 Matched case-control approach validates a number of features previously associated with exon inclusion

4.3.3.1 Alternative exons have lower H3K36me3 occupancy than constitutive exons

To investigate the potential association of HMs with AS, we compared ChIP-seq read signals of HMs (H3K36me3, H3K4me3, H3K4me1, and H3K27ac) between paired alternative and constitutive exons within the same gene. Among the HMs that were assessed, only H3K36me3 showed differential enrichment between alternative and constitutive exons. H3K36me3 is significantly enriched across constitutive exons compared to alternative exons in muscle, T cells and monocytes ($P < 0.05$, paired Mann-Whitney U test) (Figure 4.5). Moreover, H3K36me3 is more abundant in exonic regions compared to intronic regions for both alternative and constitutive exons, which is consistent with previous reports highlighting the strong association of H3K36me3 with exonic regions [98]. Altogether, these findings indicate, that H3K36me3 is positively associated with exon inclusion regulation even after accounting for confounding effects of transcription levels.

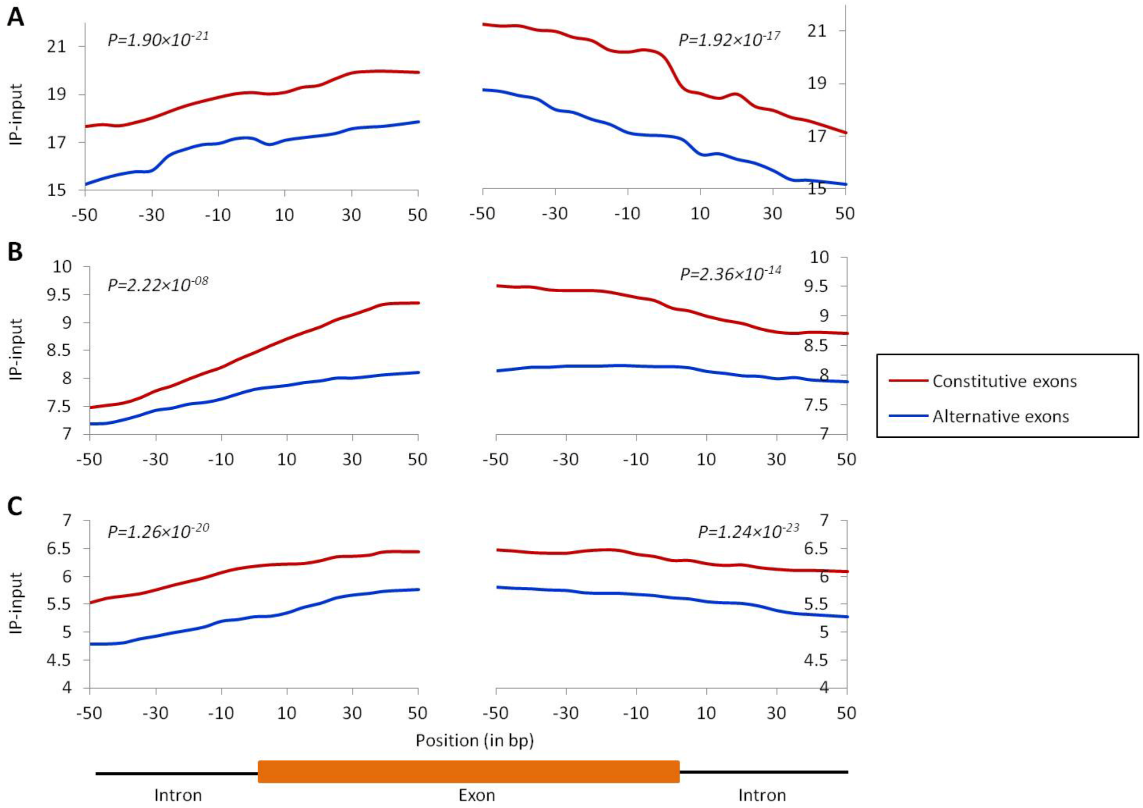


Figure 4.5: Pattern of H3K36me3 at regions of splice sites of alternative and constitutive exons.

Distribution of H3K36me3 at regions of splice sites of alternative exons (inclusion level ≤ 0.75; blue line) and constitutive exons (inclusion level between 0.95 and 1; red line) in (A) monocytes, (B) muscle and (C) T cells. H3K36me3 ChIP-seq signals within a 50bp region flanking either side of the splices sites. The *x* axis is the position in bp relative to the splice site, while the *y* axis is the input-subtracted average ChIP-Seq fragment density at each 5-bp. The Mann-Whitney U test p-value testing for differences in H3K36me3 between alternative and constitutive exons is shown on top.

4.3.3.2 Alternative exons show higher conservation of surrounding intron sequences

To gain insight into the association of conservation and AS, we next examined the extent of conservation between alternative and constitutive exons using the Genomic Evolutionary Rate Profiling (GERP) score (see method). Interestingly, introns flanking alternative exons exhibit significantly higher evolutionary constraint compared to introns flanking constitutive exons ($P < 0.05$, paired Mann-Whitney U test) (Figure 4.6). However, exonic portions (downstream 3'SS and upstream 5'SS) do not display differences in conservation constraints between alternative and constitutive exons. A higher conservation of introns in the vicinity of alternative exons has been reported previously [174, 175] and may indicate the conservation of signals essential in AS decisions (inclusion or exclusion of the exon).

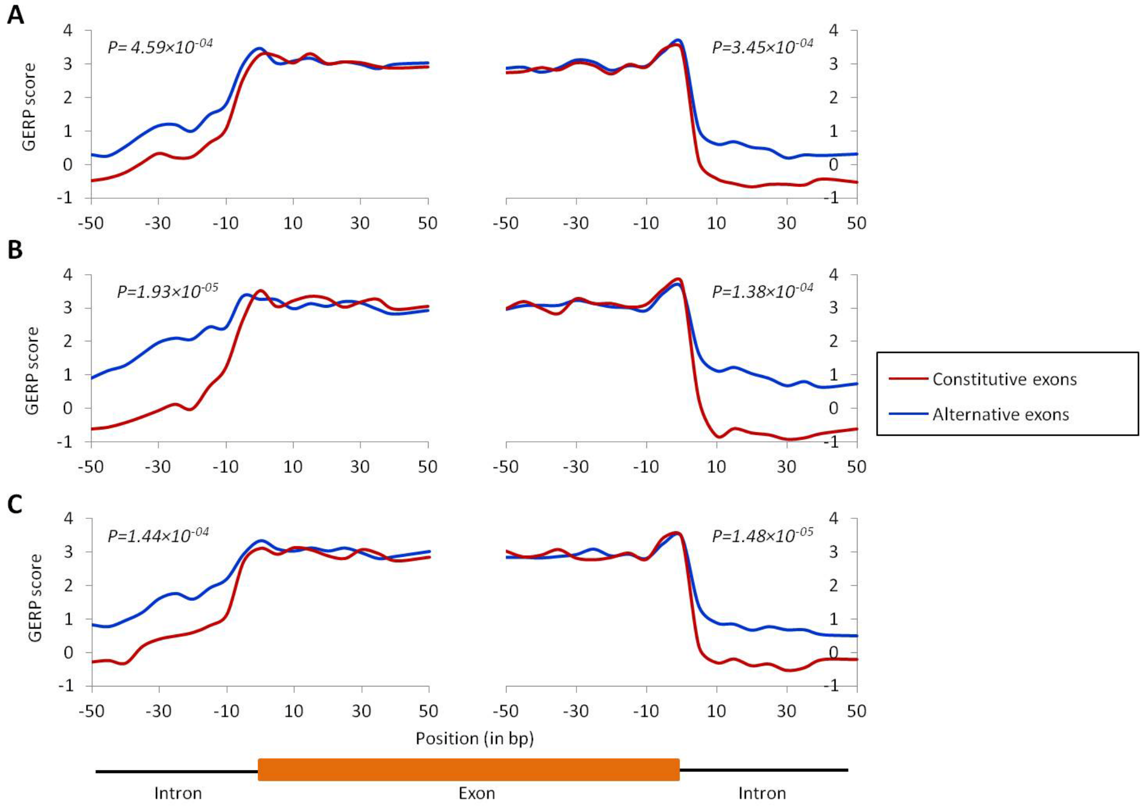


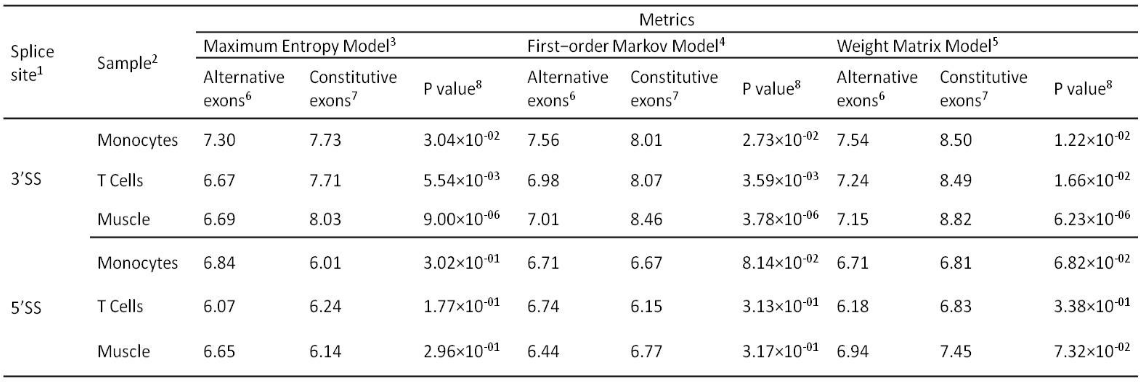
Figure 4.6: Alternative exons show higher conservation of flanking intronic sequences than constitutive exons.

Pattern of conservation ±50 bp around splice sites of alternative and constitutive exons in (A) monocytes, (B) muscle and (C) T cells. The *x* axis is the position in bp relative to the splice site, while the *y* axis is the average conservation (GERP scores; 5-bp bin). The Mann-Whitney U test p-value testing for differences in conservation between alternative and constitutive exons is shown on top.

4.3.3.3 Alternative exons have weaker 3’SS but not 5’SS

To measure the influence of splice site strength in exon inclusion, for each exon group, the 5’SS and 3’SS strengths were predicted using the maximum entropy modeling [172] (see method). We found that 3’SS of alternative exons are significantly weaker than the ones of constitutive exons in muscle, monocytes and T Cells (Table 4.1). However, we found no significant difference for 5’SS. Thus our finding partially reinforces the theory that weak splice sites are required for alternative exons [176].

Table 4-1: Constitutive exons have stronger 3’SS strengths compared to alternative exons.



Splice site[1]	Sample[2]	Metrics								
		Maximum Entropy Model[3]			First–order Markov Model[4]			Weight Matrix Model[5]		
		Alternative exons[6]	Constitutive exons[7]	P value[8]	Alternative exons[6]	Constitutive exons[7]	P value[8]	Alternative exons[6]	Constitutive exons[7]	P value[8]
3’SS	Monocytes	7.30	7.73	3.04×10^{-02}	7.56	8.01	2.73×10^{-02}	7.54	8.50	1.22×10^{-02}
	T Cells	6.67	7.71	5.54×10^{-03}	6.98	8.07	3.59×10^{-03}	7.24	8.49	1.66×10^{-02}
	Muscle	6.69	8.03	9.00×10^{-06}	7.01	8.46	3.78×10^{-06}	7.15	8.82	6.23×10^{-06}
5’SS	Monocytes	6.84	6.01	3.02×10^{-01}	6.71	6.67	8.14×10^{-02}	6.71	6.81	6.82×10^{-02}
	T Cells	6.07	6.24	1.77×10^{-01}	6.74	6.15	3.13×10^{-01}	6.18	6.83	3.38×10^{-01}
	Muscle	6.65	6.14	2.96×10^{-01}	6.44	6.77	3.17×10^{-01}	6.94	7.45	7.32×10^{-02}

Maximum Entropy Model[3], First–order Markov Model[4] and Weight Matrix Model[4] methods were used to compare the splice-site strengths between constitutive and alternative exons. The average of splice site strength scores is indicated for each splice site[1], cell type[2] and exon category[6,7]. Differences in splice site strength between alternative and constitutive exons were assessed with the paired one-sided Mann-Whitney U test[8].

Transition statement: Intron retention as a layer of gene regulation

Early studies have shown that IR contributes to the negative regulation of gene expression levels [123, 124]. At the level of pre-mRNA processing, numerous studies have reported that *(i)* introns exhibiting high retention levels are removed at a slower rate compared to other introns, and *(ii)* introns flanking alternative exons are removed more slowly [14, 177-179]. Interestingly, we found in the previous chapter that exon inclusion level is negatively correlated with IR level in the neighborhood. This prompted us to focus on introns and accurately investigate the sequence and epigenetic features that are associated with the regulation of intron retention while controlling for the confounders indentified in chapter 3.

Chapter 5: Epigenetic and sequence features associated with intron retention

5.1 Introduction

Splicing consists of the excision of intron RNA sequences between exons. During splicing, introns are normally removed and exons are joined in a mRNA precursor. Introns constitute about 25% of the human genome [3]. Increased intron retention (IR) has generally been regarded as a consequence of mis-splicing [180-184]. However, IR is emerging as an important contributor to gene regulation. Transcriptome-wide high IR has been associated with many cancers [82, 166, 185]. Wong *et al* reported in mouse bone marrow cells the use of increased IR by NMD in order to regulate the mRNA levels of genes involved in immune response during granulocytic differentiation [123]. Transcriptome-wide investigation of IR level across several cell and tissue types from human and mouse reveals surprising abundance of IR, with 35% of multiexon genes containing intron(s) with ≥50% retention in at least one cell type [124].

In contrast, relatively little is known about the factors associated with IR regulation, and there is a need to accurately investigate features associated with IR at the genome scale. In this chapter, we mainly aimed to explore features that are associated with IR after correcting for the correlation between IR and transcription. We measured the association between IR and HMs, methylation and sequences features in the same dataset as chapter 4. We corrected for gene expression influence by analyzing intron retention events within the same gene. Additionally, we analyzed separately introns having differential positions in the gene. Our corrective approach doesn’t support the previously reported association between IR and CpG content as well as methylation level. At the epigenetic level, we unprecedentedly found that H3K4me1, H3K4me3 and H3K27ac are less predominant in retained introns compared to non-retained introns. Nevertheless, a number of previously reported features specific to retained introns still hold true after correcting for the confounders. These features include short intron lengths, weaker flanking splice sites, short flanking exons and lower H3K36me3 occupancy levels. By correcting for a number of confounding variables, we improved the measured effects as well as the reliability of these features associated with IR.

5.2 Materials and methods

5.2.1 Data

Same dataset as chapter 4: DNA methylation, four histone modifications (H3K4me3, H3K4me1, H3K27ac and H3K36me3), and gene expression in 14 T lymphocyte, two monocyte and nine muscle human cells.

5.2.2 IR filtering

IR levels were estimated as described in the methods of chapter 4. IR values were then filtered according to the following criterion: to ensure *(i)* sufficient read evidence for IR level detection and *(ii)* enough coverage for the precision and resolution of the estimation of IR levels, the sum of the average read count per nucleotide of the intron and the average read count per nucleotide of exonic regions of the host gene should be equal or higher than 10. This gave rise to 90558, 91007 and 88720 introns that are flanked by coding exons in muscle, monocytes and T cells respectively. Similarly, we identified 12406, 12412, and 12324 introns that are flanked by non-coding exons in muscle, monocytes and T cells respectively.

5.2.3 Paired retained and non-retained introns within the same gene

To build the set of paired retained (IR level≥ 0.1) and non-retained (IR level < 0.05) introns, we selected retained and non-retained introns originating from the same gene according to the following criteria: *(1)* The two introns should have comparable lengths $\frac{\min(i1,i2)}{\max(i1,i2)}$ ≥ 0.1, where i_1 is the length of the retained intron and i_2 the length of the non-retained intron. *(2)* If an intron appears several times in pairs, we choose the pair with the highest IR level difference between the retained and the non-retained introns. In the set of introns flanked by coding exons, we obtained 8068, 8154 and 6143 matching intron pairs in monocytes, T cells and muscle, respectively. As for the set of introns flanked by non-coding exons we obtained 626, 625 and 504 matching intron pairs in monocytes, T cells and muscle, respectively.

5.2.4 Statistical test to compare HM, CpG, mCpG, mCpG/CpG and conservation levels between retained and non-retained introns

As described in the methods of chapter 4. GERP score, CpG, methylation and HM levels were estimated in 500-bp upstream and 50-bp downstream 3’SS, and 50-bp upstream and 500-bp downstream 5’SS.

5.2.5 Visualization of DNA methylation, HMs, CpG levels and conservation

Features were estimated as described in the methods of chapter 4. *(i)* GERP score, CpG and methylation levels were plotted in 300-bp upstream and 50-bp downstream 3'SS, and 50-bp upstream and 300-bp downstream 5'SS. *(ii)* HMs were plotted in 500-bp upstream and 50-bp downstream 3'SS, and 50-bp upstream and 500-bp downstream 5'SS.

5.2.6 Analysis of splice-site strength

As described in the methods of chapter 4.

5.2.7 Identification of differential intron retention

We selected introns that are retained (IR level ≥ 0.1) in at least one cell type; this gave rise to 88802 introns which belong to 9108 genes for the set of introns flanked by coding exons. Similarly, we obtained 11695 introns which belong to 5318 genes for the set of introns flanked by non coding exons. To investigate differential intron retention among muscle, monocytes and T cells, we performed cell-type pairwise comparisons of IR levels by applying for each intron a student t-test, then followed by a FDR correction at α=0.05 level. An adjusted p-value < 0.05 (after FDR correction) and a minimum IR fold-change of 2 between at least two cell types were applied to consider the significance of differentially retained introns.

5.2.8 Gene Ontology enrichment analysis of differentially retained introns

We retrieved the host genes of introns -flanked by coding exons- that were differentially retained between ≥ 2 cell types. These genes were compared against all protein-coding genes using the PANTHER resource [186, 187] to test for differential enrichment of gene ontology (GO) biological processes.

5.3 Results

5.3.1 Intron retention is widespread in the transcriptome

Our method of cDNA preparation for sequencing retains a detectable fraction of immature transcripts which could comprise either full-length pre-mRNA molecules or nascent transcripts. Notably, 38% of all mapped sequence reads were located in introns (Figure 5.1). This proportion is similar to previously reported rates [137, 188] and indicates a large number of immature transcripts. Thus, we were able to obtain a considerable amount of information about the levels of heteronuclear pre-mRNAs from which the introns have not been completely removed.

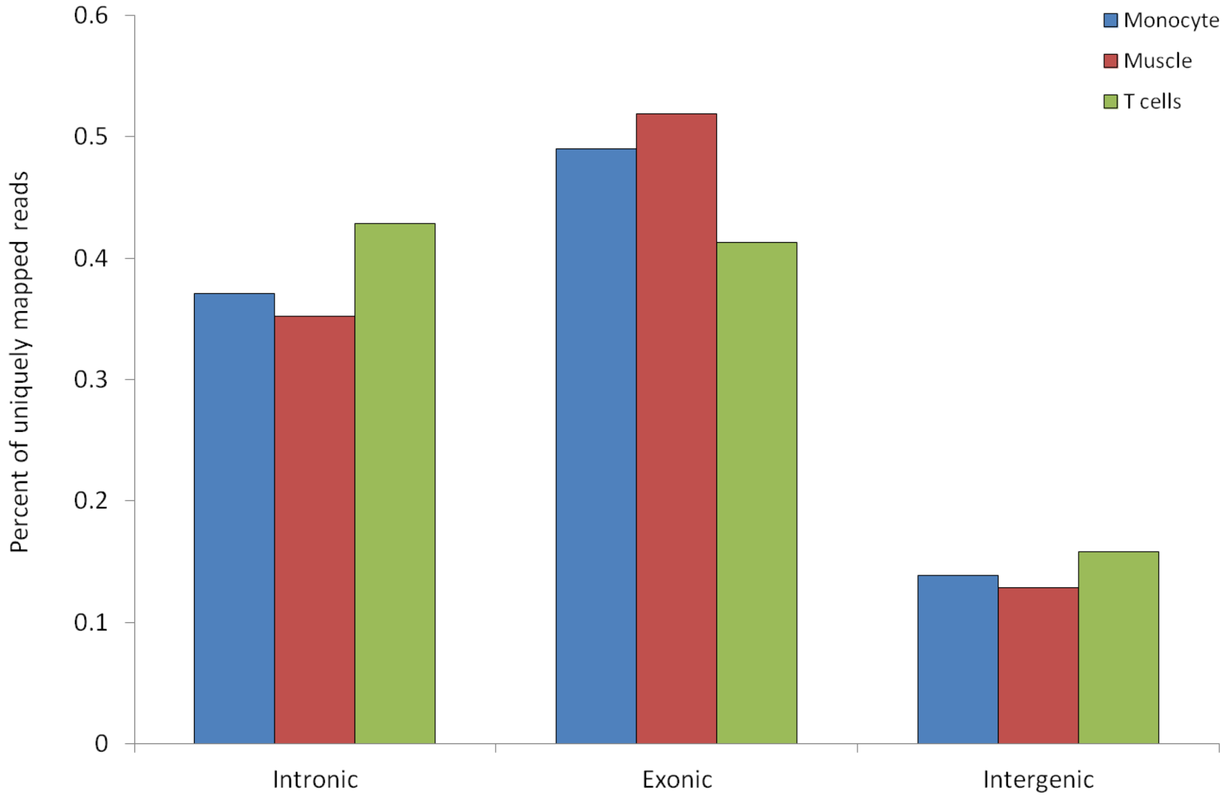


Figure 5.1: A large proportion of RNA-Seq reads map to intronic regions.

Percentage of reads mapping to intronic, exonic and intergenic regions is shown for each cell type (on the *y* axis).

We separately analyzed introns flanked by protein-coding exons and introns flanked by non-coding exons (from UTRs and non-coding genes) due to differences in position in the gene. The level of IR was estimated as described in the methods of chapter 3. IR level estimates were filtered for detectable introns (see methods). We defined the candidate retained introns as the introns with IR level equal or greater than 0.1, and the candidate non-retained introns as the introns with IR level less than 0.05. As previously reported, we found that retained introns are shorter, and this observation is more pronounced for introns that are flanked by non-coding exons (data not shown). Only retained introns that are flanked by non-coding exons have a preference for genes with shorter length compared to non-retained introns (Figure 5.2).

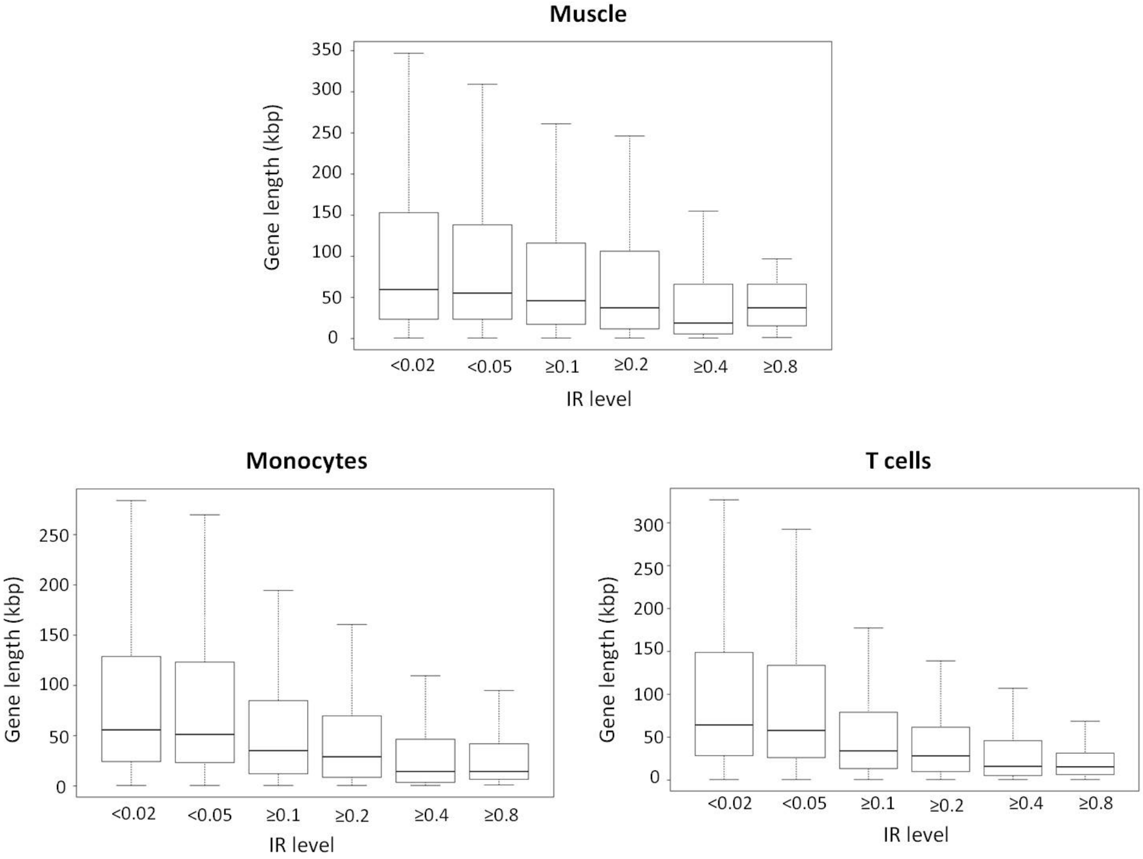


Figure 5.2: Distribution of lengths of host-genes in function of IR level.

The *x* axis represents the IR level ranges and the *y* axis the host gene length in kbp.

IR was detected to variable extents between T cells and muscle, with a significantly higher fraction of mapped intronic reads in T cells compared to muscle ($P = 3.61 \times 10^{-4}$, Mann-Whitney U test). However, we didn't detect differential amount of mapped intronic reads between T cells and monocytes ($P = 0.81$, Mann-Whitney U test), and between monocytes and muscle (*P* = 0.40, Mann-Whitney U test). At the gene level, T cells had the highest proportion of IR with 82% of genes having at least one retained intron (IR level ≥ 0.1), followed by monocytes (74%) and muscle (66%) (Figure 5.3). These differences in IR frequency are unlikely due to differences in RNA-seq data quality (such as DNA contamination) since all samples have comparable rate of intragenic reads (Figure 5.3 and Figure 5.4).

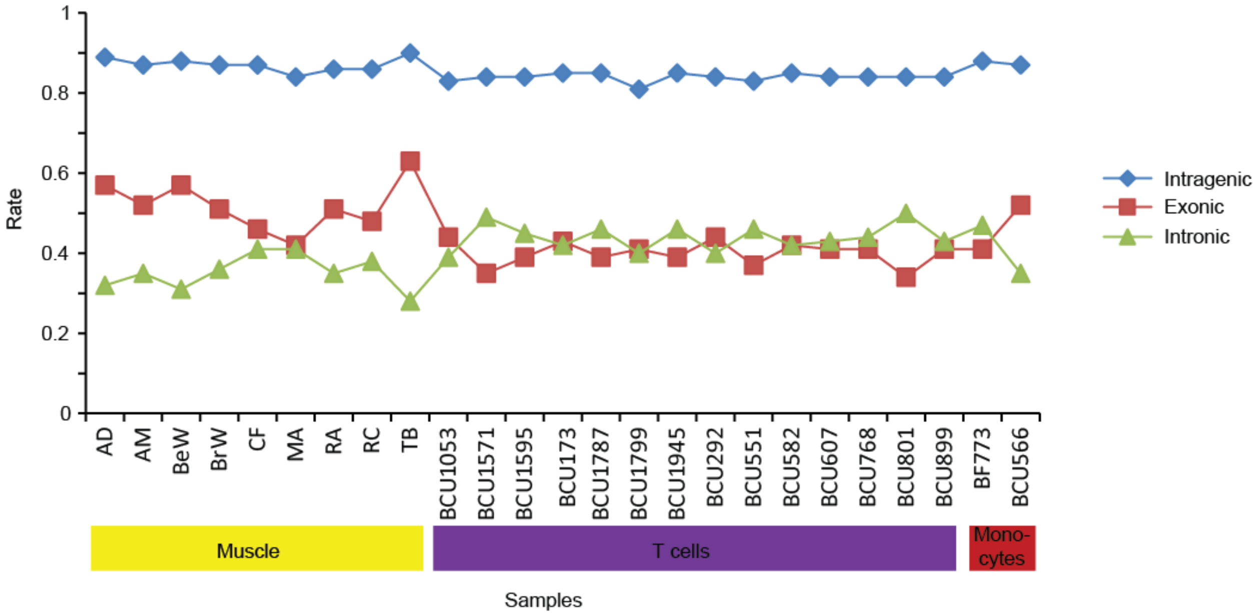


Figure 5.3: Fraction of reads that map within genomic features.

The *x* axis is the sample labels with the muscle underlined in yellow, the T cells in purple and the monocytes in dark red. The *y* axis represents the fraction of mapping reads.

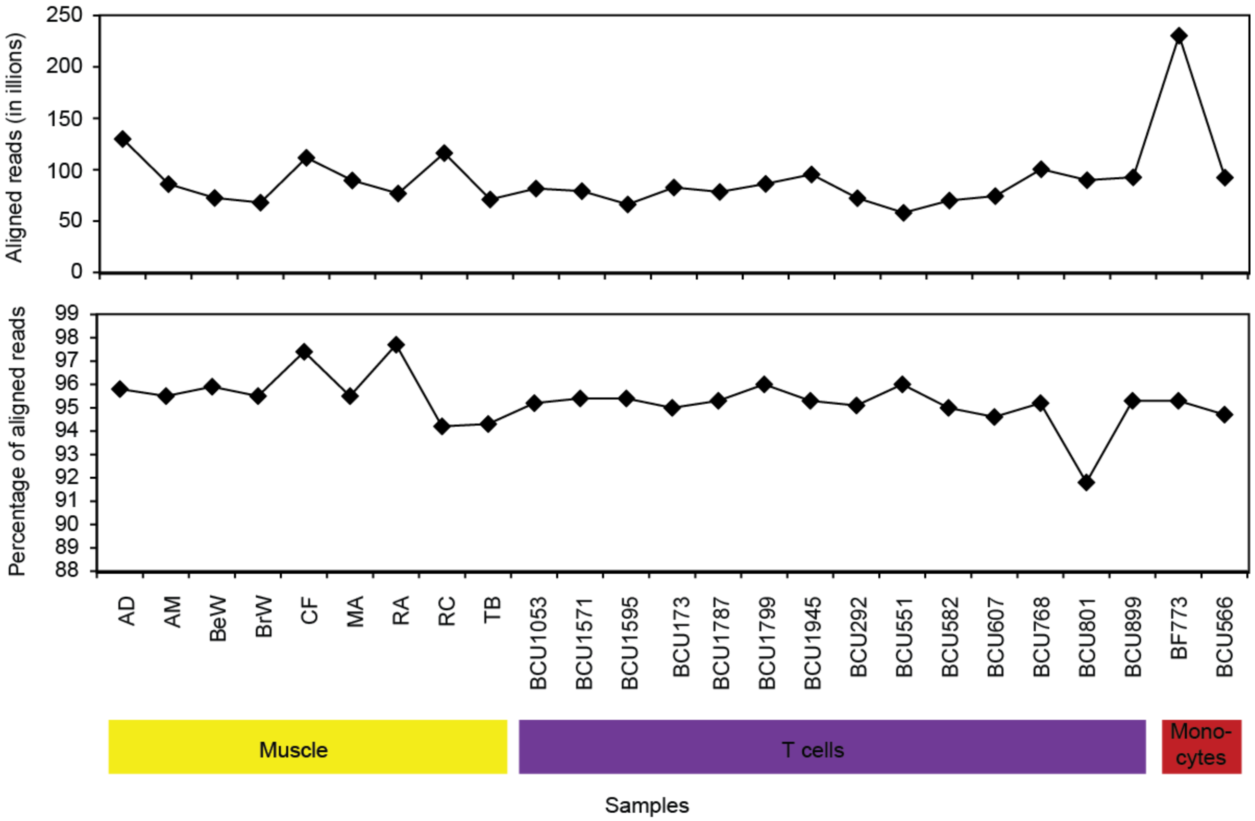


Figure 5.4: Aligned reads in samples.

The *x* axis is the sample labels with the muscle underlined in yellow, the T cells in purple and the monocytes in dark red. In the top panel, the *y* axis is the raw number of aligned reads in millions, and in the bottom panel the *y* axis is the percentage of read aligned over the total number of reads.

5.3.2 Differential intron retention affects mostly genes involved in mRNA processing and translation

First we performed a hierarchical clustering to merge samples with the most similar IR levels. We found that samples from the same cell type cluster together, indicating that patterns of IR are cell type-specific (Figure 5.5). We next investigated differences in intron retention between monocytes, T cells and Muscle (see methods). For the group of introns flanked by coding exons (88802 introns which belong to 9108 genes), 4249(5%) introns which belong to 717(8%) genes were differentially retained between monocytes and muscle; 2136(2%) introns which belong to 277(3%) genes were differentially retained between monocytes and T cells; and 13610(15%) introns which belong to 1656(18%) genes were differentially retained between T cells and muscle. For the group of introns flanked by non-coding exons (11695 introns which belong to 5318 genes), 990(8%) introns which belong to 359(7%) genes were differentially retained between monocytes and muscle; 436(4%) introns which belong to 142(3%) genes were differentially retained between monocytes and T cells; and 2280(19%) introns which belong to 623(12%) genes were differentially retained between T cells and muscle.

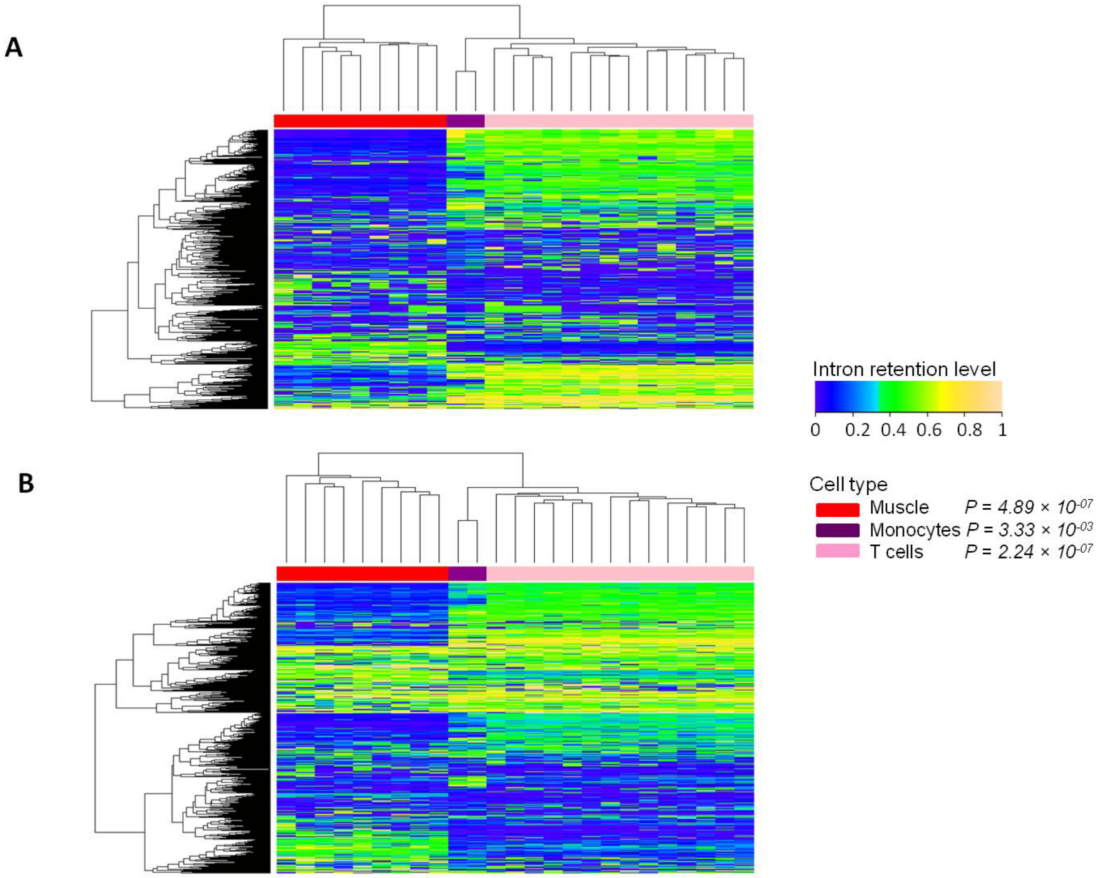


Figure 5.5: Heatmap showing hierarchical clustering based on IR levels of introns.

Heatmap showing hierarchical clustering based on IR levels (the top 1000 most variable introns) of (A) introns flanked by coding exons and (B) introns flanked by non-coding exons. Samples segregate by cell type (muscle, red; monocytes, purple; T cells, pink). The significance (p-value) of the Fisher test testing that samples from the same cell type cluster together is indicated.

Next, we retrieved the host genes of differentially retained introns flanked by coding exons, and we tested whether they are preferentially involved in particular biological processes by performing a Gene Ontology analysis (see methods). The gene ontology analysis revealed that genes associated with differential IR level across cell-types are highly enriched in RNA processing and translation categories, with mRNA splicing as the most enriched term (Figure 5.6-A). This set of genes encodes mainly transcription factor proteins, mRNA splicing factor proteins, and translation initiation/elongation factors. Differentially IR targeting preferentially genes involved in RNA processing and translation has been reported recently by two independent studies [163, 166]. Interestingly, similar analysis at the transcript expression levels revealed that most of the differentially expressed genes(DEseq analysis[150]) are mainly involved in immune system – as would be expected from studies involving white blood cells (Figure 5.6-B). This difference in GO terms indicates that the observed differences in IR level are not driven by the differential expression of the host genes. Altogether, these data suggest that although the main regulation might be via gene expression variation, there are some classes of genes that are targeted by IR-mediated regulation.

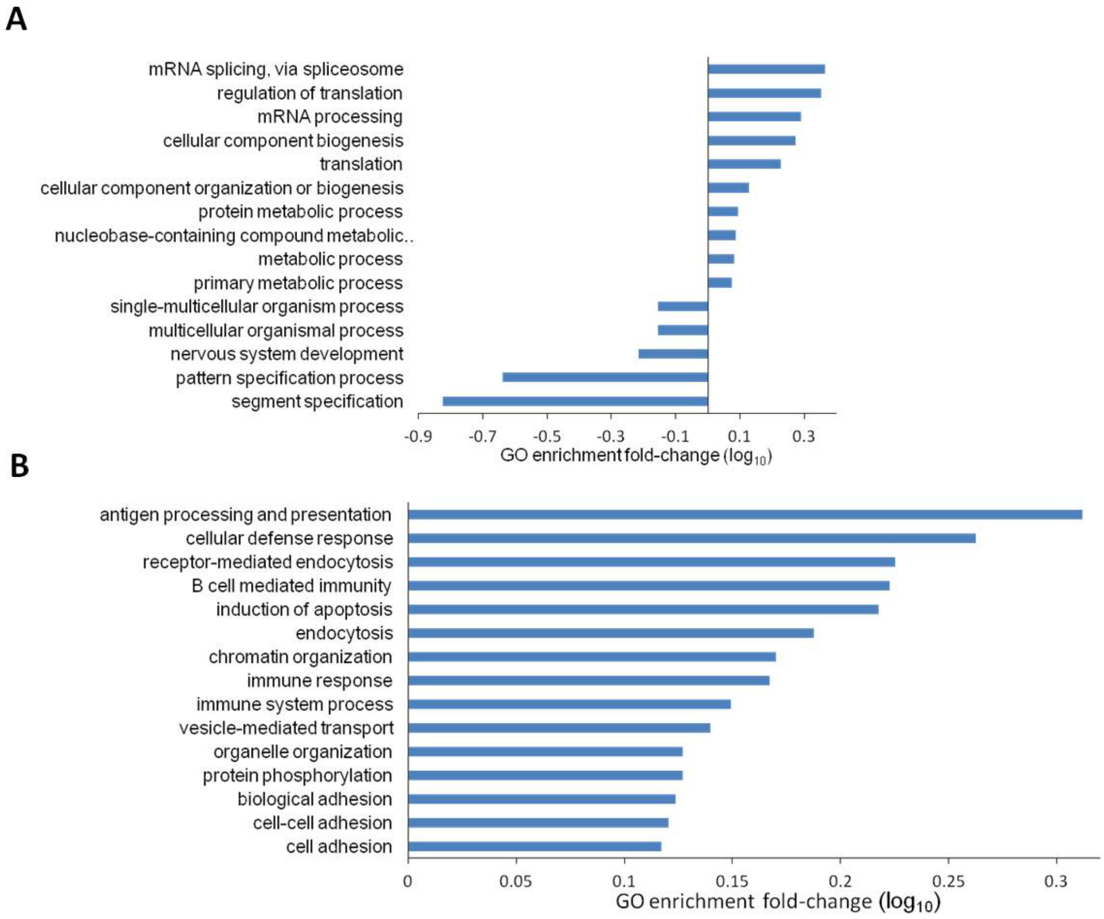


Figure 5.6: Different gene ontology (GO) terms between the differentially expressed genes and the host-genes of differentially retained introns.

Representative gene ontology (GO) terms significantly enriched among (A) the host genes of differentially retained introns and (B) the differentially expressed genes. The *x* axis indicates the enrichment fold change ($\log_{10}$ scale) of the biological process (*y* axis).

5.3.3 CpG content and DNA methylation are not associated with intron retention

Prior works reported that retained introns display higher levels of GC content and DNA methylation compared to non-retained introns [167, 189]. However, after controlling for the influence of gene expression and the positional effect of introns, we did not observe any difference in CpG-density between retained and non-retained introns (Figure 5.7-A). Similarly to the analysis of alternative exons, we did not observe differences in DNA methylation levels (mCpG/CpG) between retained introns and non-retained introns (Figure 5.7-B). Thus, after correcting for a number of confounders (CpG abundance, gene expression levels and the positional effect of introns), our results indicate that the association between IR, CpG and DNA methylation no longer holds.

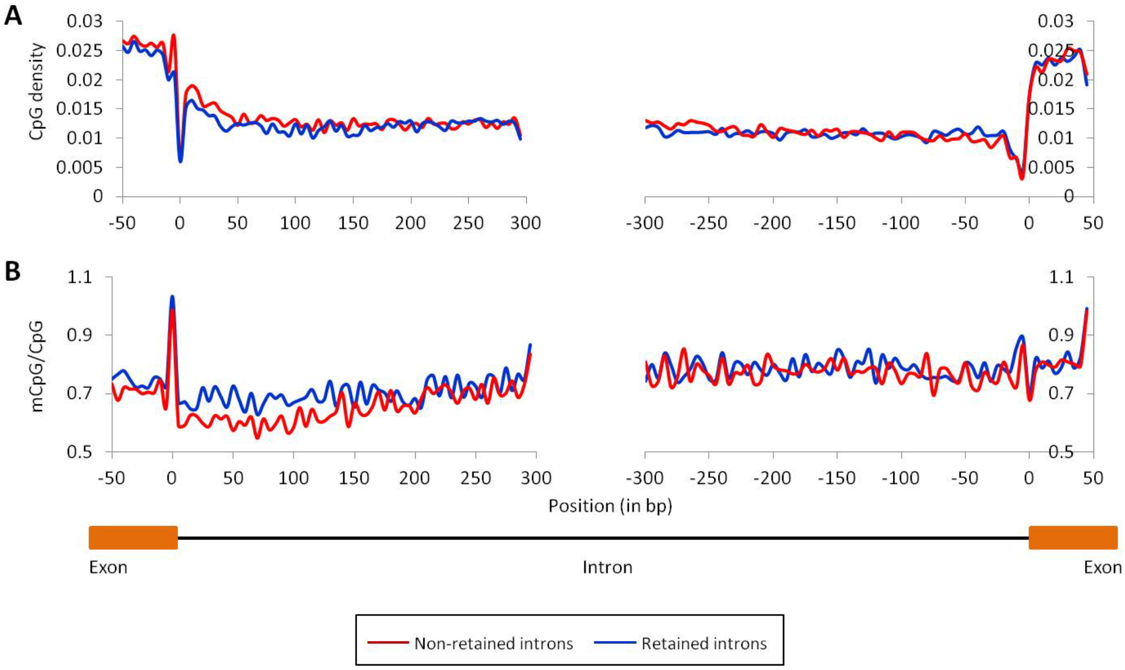


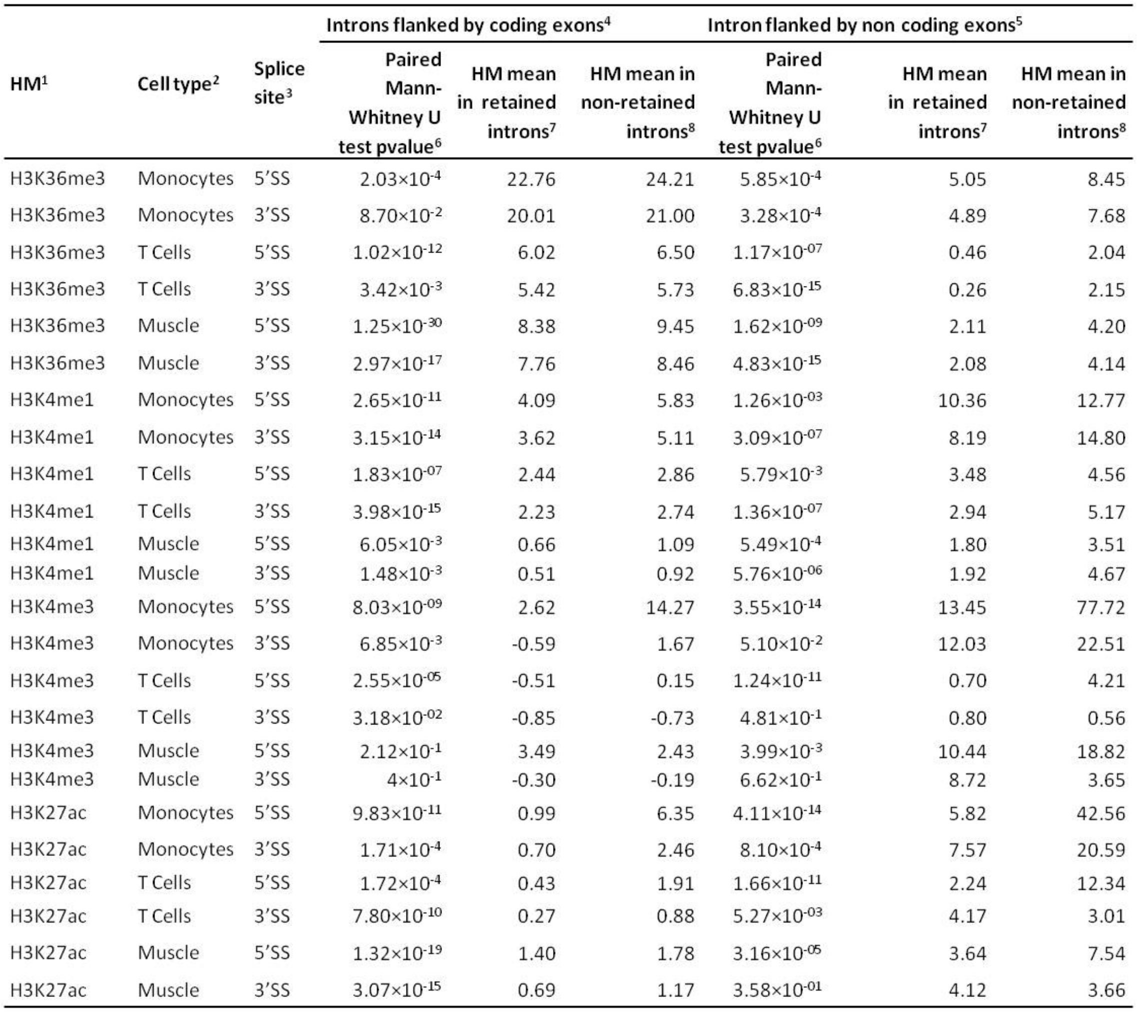
Figure 5.7: CpG and methylation profiles across retained and non-retained introns in muscle.

Average (5-bp window) of (A) CpG density and (B) CpG-normalized methylation in the first and last 300-bp of introns flanked by coding exons.

5.3.4 Intron retention has a distinct relationship with HMs

Since HMs have been implicated in differential exon inclusion, we wondered whether they may also be involved in differential intron retention. H3K4me3, H3K4me1 and H3K27ac have been reported to be enriched over retained introns compared to non-retained introns [124, 189], whereas H3K36me3 deficiency is associated with global inefficiency of intron removal [82]. We analyzed the distribution profiles of H3K36me3, H3K4me1, H3K4me3 and H3K27ac in the first and last 500bp of introns, after controlling for gene expression and the uneven distribution of HMs. For all three cell types, we observed a significant enrichment of H3K36me3, H3K34me3, H3K34me1 and H3K27ac signals in non-retained introns compared to retained introns for both introns interrupting coding exons and introns interrupting non-coding exons (Table 5.1, and Figure 5.8 - Figure 5.12). In opposition to previous studies, our results suggest that IR is negatively associated with H3K34me3, H3K34me1, and H3K27ac marks; moreover, this association is independent of the relationship between these HMs and gene expression.

Table 5-1: Differential HM signals between retained and non-retained introns.



HM[1]	Cell type[2]	Splice site[3]	Introns flanked by coding exons[4]			Intron flanked by non coding exons[5]		
			Paired Mann-Whitney U test pvalue[6]	HM mean in retained introns[7]	HM mean in non-retained introns[8]	Paired Mann-Whitney U test pvalue[6]	HM mean in retained introns[7]	HM mean in non-retained introns[8]
H3K36me3	Monocytes	5'SS	2.03×10^{-4}	22.76	24.21	5.85×10^{-4}	5.05	8.45
H3K36me3	Monocytes	3'SS	8.70×10^{-2}	20.01	21.00	3.28×10^{-4}	4.89	7.68
H3K36me3	T Cells	5'SS	1.02×10^{-12}	6.02	6.50	1.17×10^{-07}	0.46	2.04
H3K36me3	T Cells	3'SS	3.42×10^{-3}	5.42	5.73	6.83×10^{-15}	0.26	2.15
H3K36me3	Muscle	5'SS	1.25×10^{-30}	8.38	9.45	1.62×10^{-09}	2.11	4.20
H3K36me3	Muscle	3'SS	2.97×10^{-17}	7.76	8.46	4.83×10^{-15}	2.08	4.14
H3K4me1	Monocytes	5'SS	2.65×10^{-11}	4.09	5.83	1.26×10^{-03}	10.36	12.77
H3K4me1	Monocytes	3'SS	3.15×10^{-14}	3.62	5.11	3.09×10^{-07}	8.19	14.80
H3K4me1	T Cells	5'SS	1.83×10^{-07}	2.44	2.86	5.79×10^{-3}	3.48	4.56
H3K4me1	T Cells	3'SS	3.98×10^{-15}	2.23	2.74	1.36×10^{-07}	2.94	5.17
H3K4me1	Muscle	5'SS	6.05×10^{-3}	0.66	1.09	5.49×10^{-4}	1.80	3.51
H3K4me1	Muscle	3'SS	1.48×10^{-3}	0.51	0.92	5.76×10^{-06}	1.92	4.67
H3K4me3	Monocytes	5'SS	8.03×10^{-09}	2.62	14.27	3.55×10^{-14}	13.45	77.72
H3K4me3	Monocytes	3'SS	6.85×10^{-3}	-0.59	1.67	5.10×10^{-2}	12.03	22.51
H3K4me3	T Cells	5'SS	2.55×10^{-05}	-0.51	0.15	1.24×10^{-11}	0.70	4.21
H3K4me3	T Cells	3'SS	3.18×10^{-02}	-0.85	-0.73	4.81×10^{-1}	0.80	0.56
H3K4me3	Muscle	5'SS	2.12×10^{-1}	3.49	2.43	3.99×10^{-3}	10.44	18.82
H3K4me3	Muscle	3'SS	4×10^{-1}	-0.30	-0.19	6.62×10^{-1}	8.72	3.65
H3K27ac	Monocytes	5'SS	9.83×10^{-11}	0.99	6.35	4.11×10^{-14}	5.82	42.56
H3K27ac	Monocytes	3'SS	1.71×10^{-4}	0.70	2.46	8.10×10^{-4}	7.57	20.59
H3K27ac	T Cells	5'SS	1.72×10^{-4}	0.43	1.91	1.66×10^{-11}	2.24	12.34
H3K27ac	T Cells	3'SS	7.80×10^{-10}	0.27	0.88	5.27×10^{-03}	4.17	3.01
H3K27ac	Muscle	5'SS	1.32×10^{-19}	1.40	1.78	3.16×10^{-05}	3.64	7.54
H3K27ac	Muscle	3'SS	3.07×10^{-15}	0.69	1.17	3.58×10^{-01}	4.12	3.66

For each HM[1], cell type[2] and splice site[3], average HM levels (ChIP-Seq read coverage minus its corresponding input-DNA coverage) are indicated for introns flanked by coding exons[4] and introns flanked by non-coding exons[5]. Differences in HM levels between retained[7] and non-retained[8] introns were assessed with the paired Mann-Whitney U test[6].

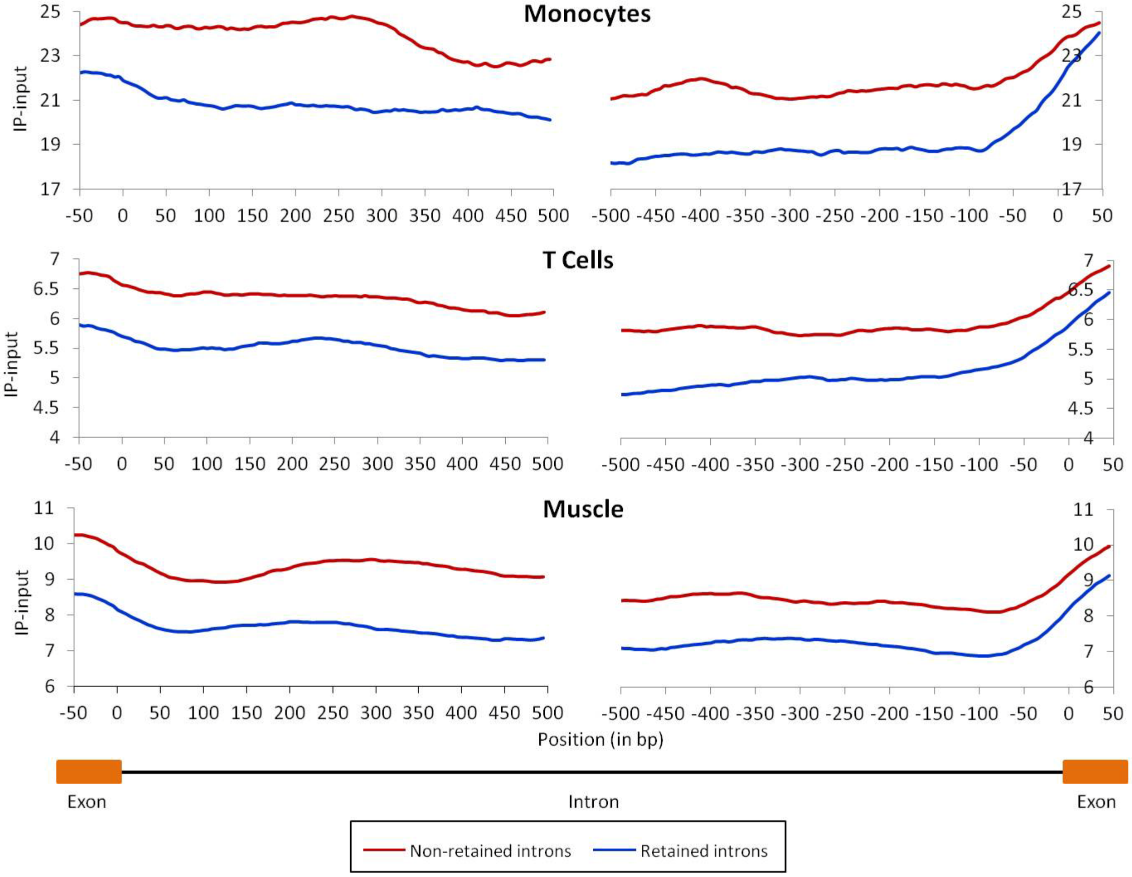


Figure 5.8: Pattern of H3K36me3 in retained and non-retained introns flanked by coding exons.

Average signals of H3K36me3 in the first and last 500-bp of retained (IR level ≥ 0.1; blue line) and non-retained (IR level < 0.05; red line) introns flanked by coding exons. The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows.

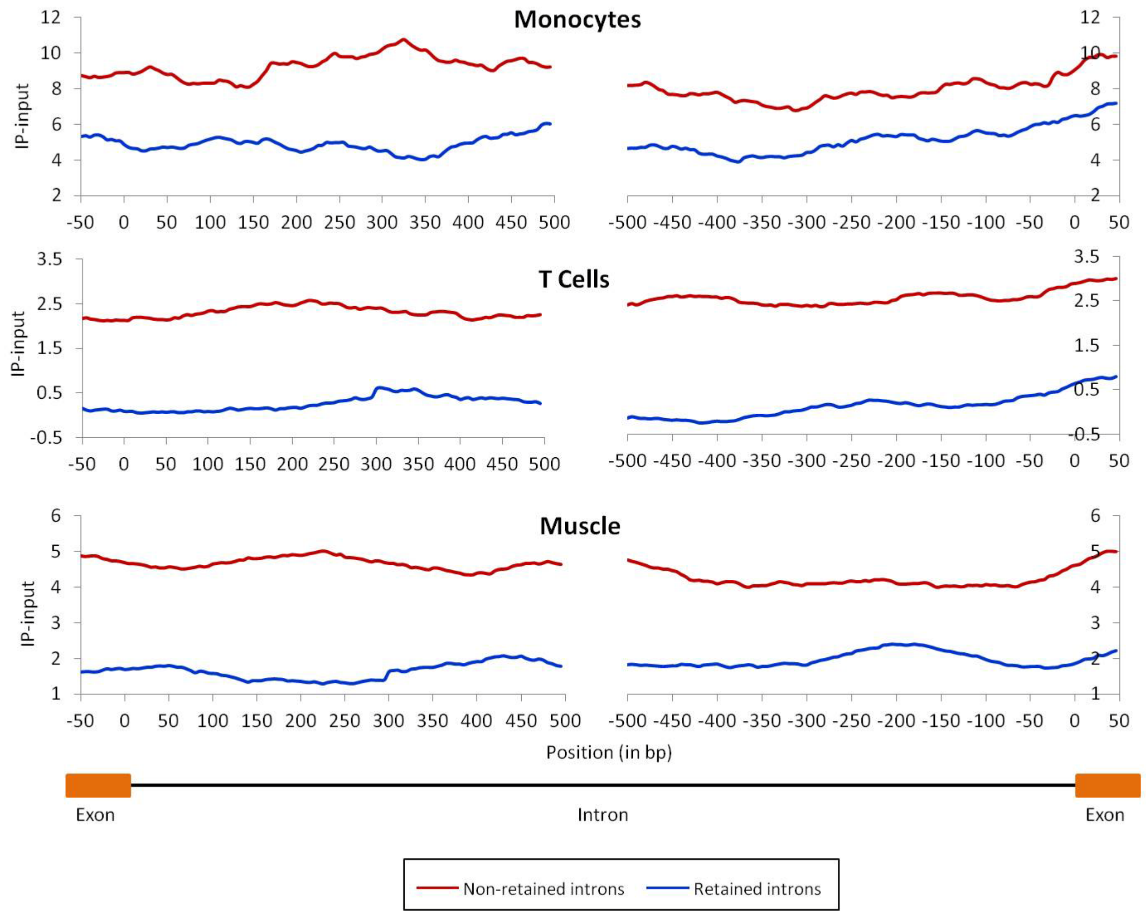


Figure 5.9: Pattern of H3K36me3 in retained and non-retained introns flanked by non-coding exons.

Average signals of H3K36me3 in the first and last 500bp of retained (IR level ≥ 0.1; blue line) and non-retained (IR level < 0.05; red line) introns flanked by non-coding exons. The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows.

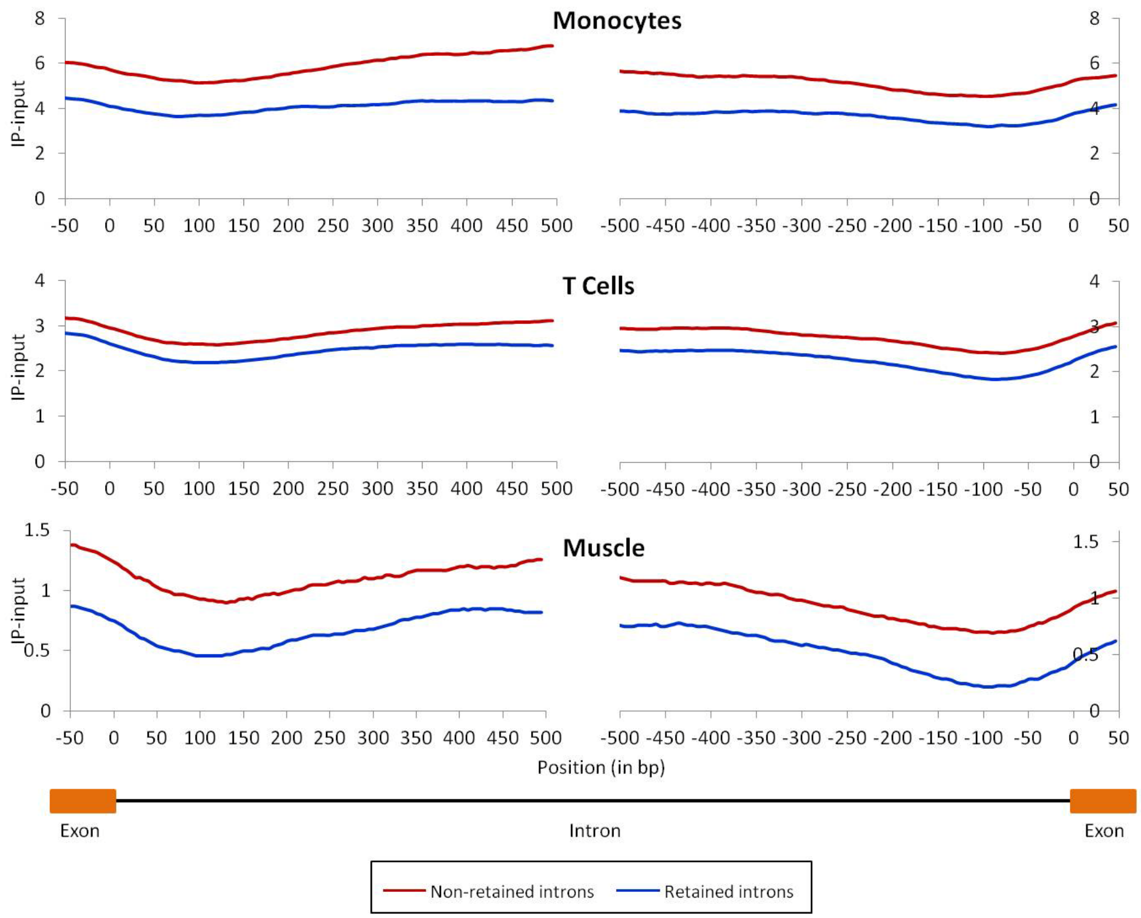


Figure 5.10: Pattern of H3K4me1 in retained and non-retained introns flanked by coding exons.

Average signals of H3K4me1 in the first and last 500bp of retained (IR level ≥ 0.1; blue line) and non-retained (IR level < 0.05; red line) introns flanked by coding exons. The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows.

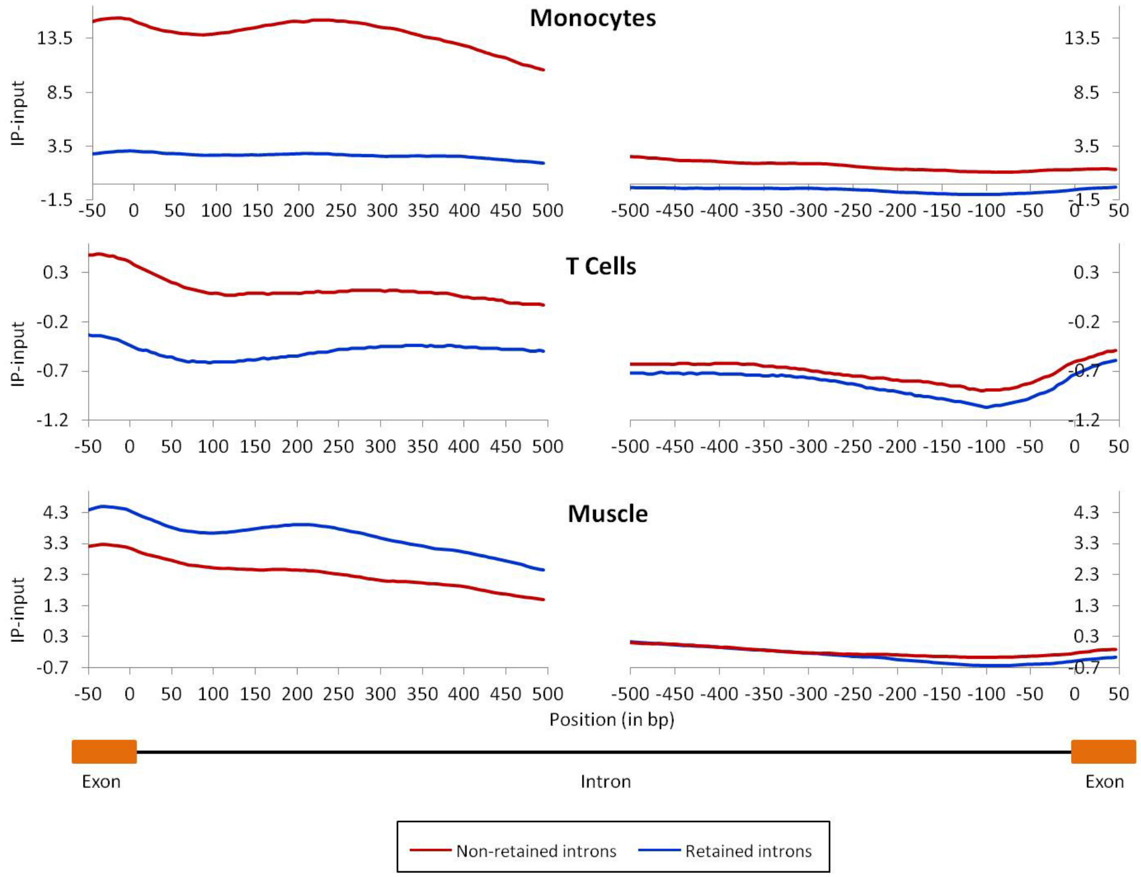


Figure 5.11: Pattern of H3K4me3 in retained and non-retained introns flanked by coding exons.

Average signals of H3K4me3 in the first and last 500bp of retained (IR level ≥ 0.1; blue line) and non-retained (IR level < 0.05; red line) introns flanked by coding exons. The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows.

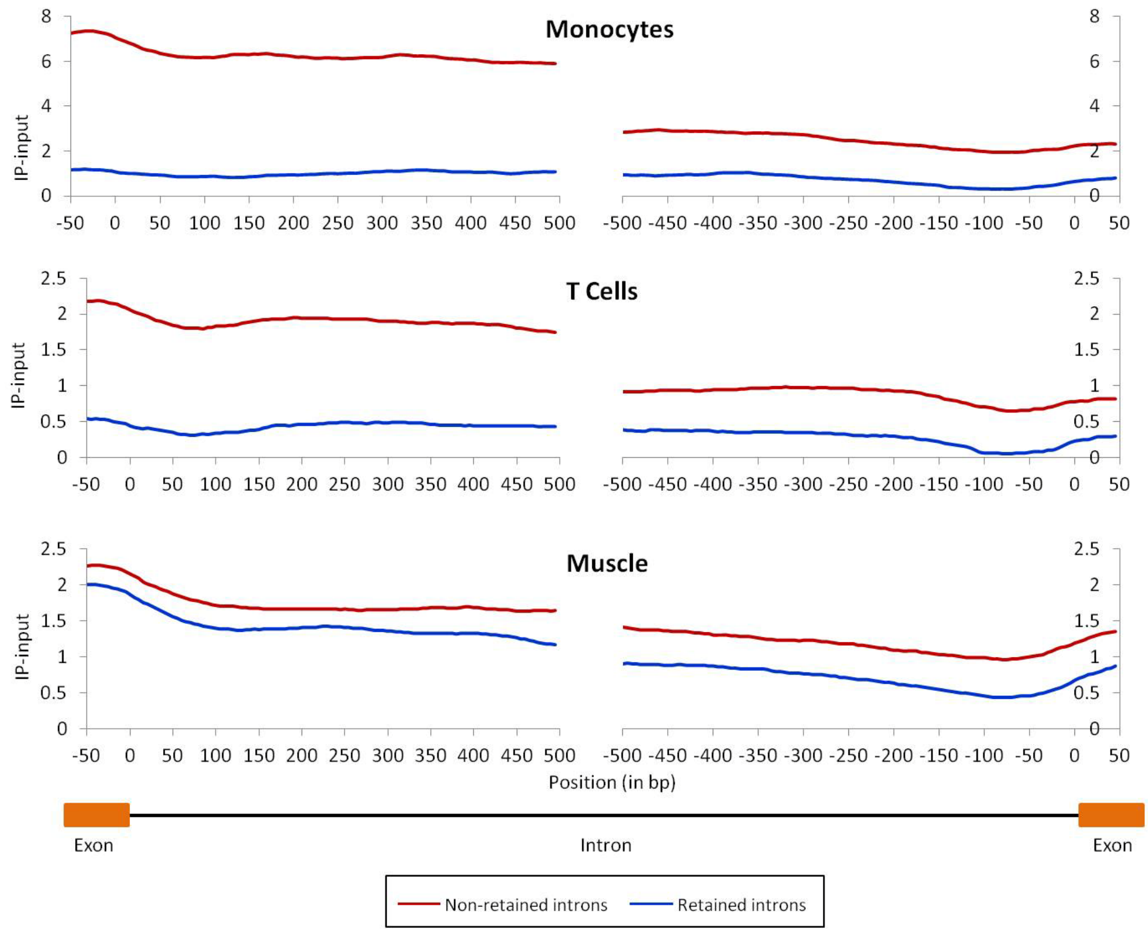


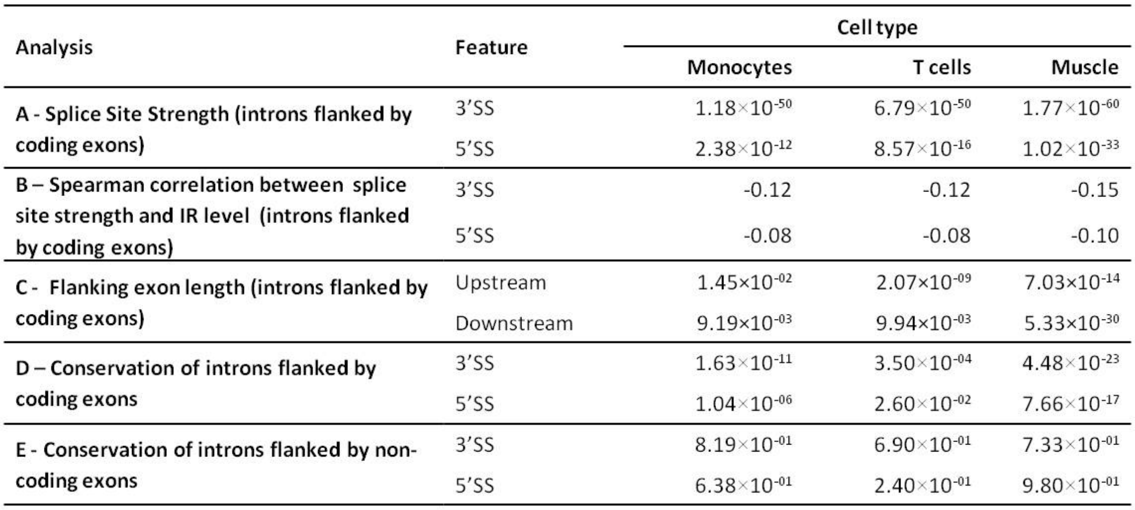
Figure 5.12: Pattern of H3K27ac in retained and non-retained introns flanked by coding exons.

Average ChIP-Seq signals of H3K27ac in the first and last 500bp of retained (IR level ≥ 0.1; blue line) and non-retained (IR level <0.05; red line) introns flanked by coding exons. The *x* axis is the position in bp relative to the splice site, and *y* axis indicates the input-subtracted average ChIP-Seq fragment density, 5-bp windows.

5.3.5 Our corrective approach validates a number of previously reported features specific to IR

Previous genome-wide studies have reported cis-acting sequence features specific to retained introns, including shorter lengths, weaker splice sites, weaker conservation of splice sites, shorter flanking exons, and higher GC content [124, 160]. To compare features between retained introns and non-retained introns, we analyzed matching pairs of retained and non-retained introns within the same gene (see methods). To test the contribution of splice site strength in intron retention, we estimated the strength of 5’SS and 3’SS flanking introns using the maximum entropy approach (see methods). As expected, retained introns are flanked by significantly weaker splice sites compared to non-retained introns (Table 5.2.A, paired one sided Mann-Whitney U test). Also, we found a significant modest anti-correlation (Spearman correlation $P < 2.2\times10^{-16}$) between IR level and splice site strength, with the correlation stronger for 3’ SS (Table 5.2.B). Additionally, we also noticed that introns flanked by non-coding exons have weaker 3’SS and 5’SS compared to introns flanked by coding exons.

Table 5-2: Differential sequence features between retained and non-retained introns.



Analysis	Feature	Cell type		
		Monocytes	T cells	Muscle
A - Splice Site Strength (introns flanked by coding exons)	3’SS	1.18×10^{-50}	6.79×10^{-50}	1.77×10^{-60}
	5’SS	2.38×10^{-12}	8.57×10^{-16}	1.02×10^{-33}
B – Spearman correlation between splice site strength and IR level (introns flanked by coding exons)	3’SS	-0.12	-0.12	-0.15
	5’SS	-0.08	-0.08	-0.10
C - Flanking exon length (introns flanked by coding exons)	Upstream	1.45×10^{-02}	2.07×10^{-09}	7.03×10^{-14}
	Downstream	9.19×10^{-03}	9.94×10^{-03}	5.33×10^{-30}
D – Conservation of introns flanked by coding exons	3’SS	1.63×10^{-11}	3.50×10^{-04}	4.48×10^{-23}
	5’SS	1.04×10^{-06}	2.60×10^{-02}	7.66×10^{-17}
E - Conservation of introns flanked by non-coding exons	3’SS	8.19×10^{-01}	6.90×10^{-01}	7.33×10^{-01}
	5’SS	6.38×10^{-01}	2.40×10^{-01}	9.80×10^{-01}

(A) Paired Mann-Whitney U test comparing the splice site strength between retained and non-retained introns flanked by coding exons. The significance of the test (p-value) is indicated for each cell type and feature. (B) Spearman correlation between splice site strength and IR level; the correlation coefficient is indicated for each cell type and feature. (C) The length of flanking exons and (D, E) the evolutionary conservation score (GERP) were compared between retained and non-retained introns. The significance of the comparisons (p-value), assessed by the paired Mann-Whitney U test, is given for each cell type and feature.

The comparison of exon lengths flanking introns showed that retained introns are flanked by significantly shorter upstream and downstream exons (Table 5.2.C, paired one-sided Mann-Whitney U test). We next evaluated whether evolutionary constraint is involved in intron retention (see methods). In the set of introns flanked by coding exons, the comparison of means of GERP scores between retained and non-retained introns showed that non-retained introns are significantly more conserved than retained introns around splice sites (Table 5.2.D, one sided paired Mann-Whitney U test) (Figure

5.13). The higher conservation of non-retained introns in the vicinity of splice sites could be explained by the presence of functional signals essential for proper splicing processing. However, difference in conservation near splice sites was not observed for introns flanked by non-coding exons (Table 5.2.E, Paired Mann-Whitney U test) (Figure 5.14) indicating that introns interrupting non-coding exons may harbor less or weaker splicing-regulatory signals around splice sites; another explanation is that the conservation of splicing signals might be not rigorous for non-coding exons since they do not code for protein product. We also noticed that introns flanked by non-coding exons are significantly less conserved than introns flanked by coding exons in general (data not shown). Altogether, these analyses show that even after controlling for the influence of gene expression and the position of introns in the gene, the previously reported association between intron retention and specific sequence features hold true. Our results clarify the relationship between IR and specific sequence features by eliminating a number of confounders.

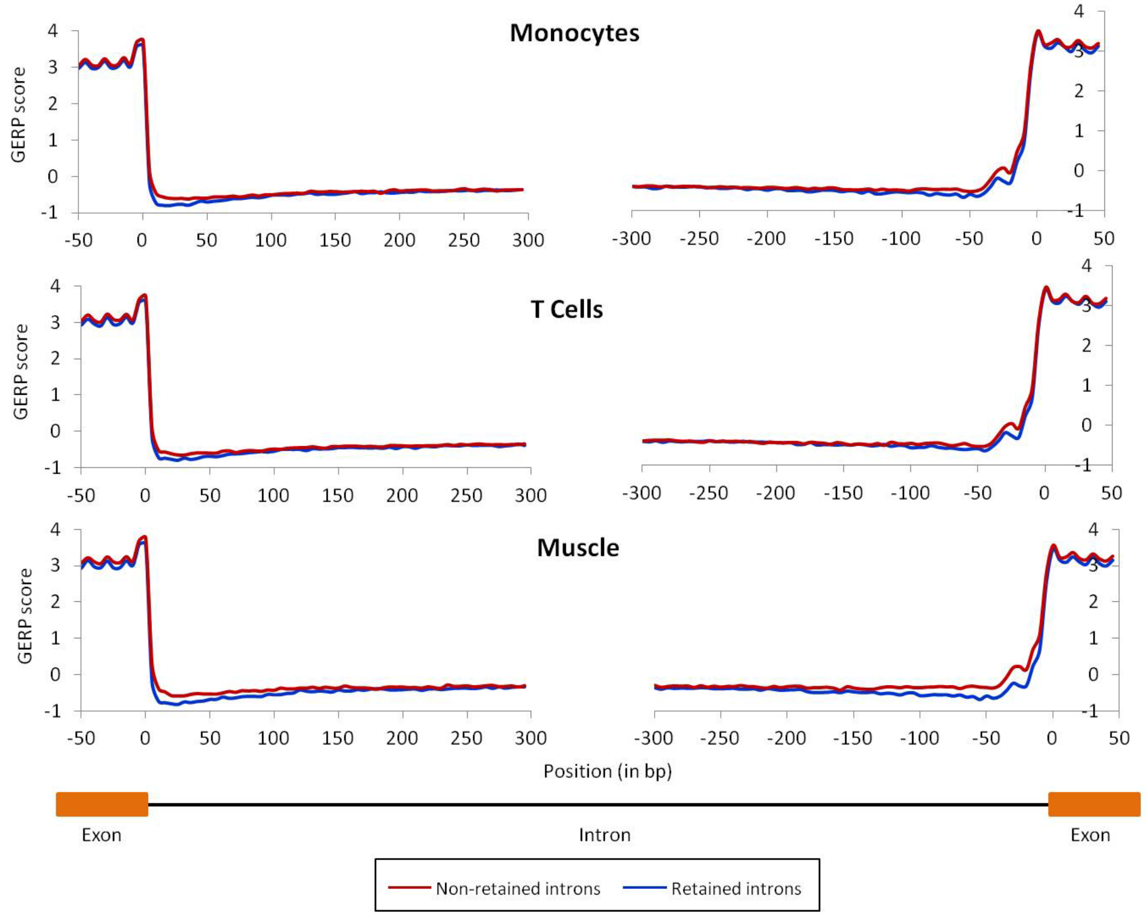


Figure 5.13: Pattern of conservation of retained and non-retained introns flanked by coding exons.

Average GERP conservation scores (5-bp window) around 5’SS and 3’SS of retained introns and non-retained introns flanked by coding exons. The *x* axis is the position in bp relative to the splice site, while the *y* axis is the GERP score.

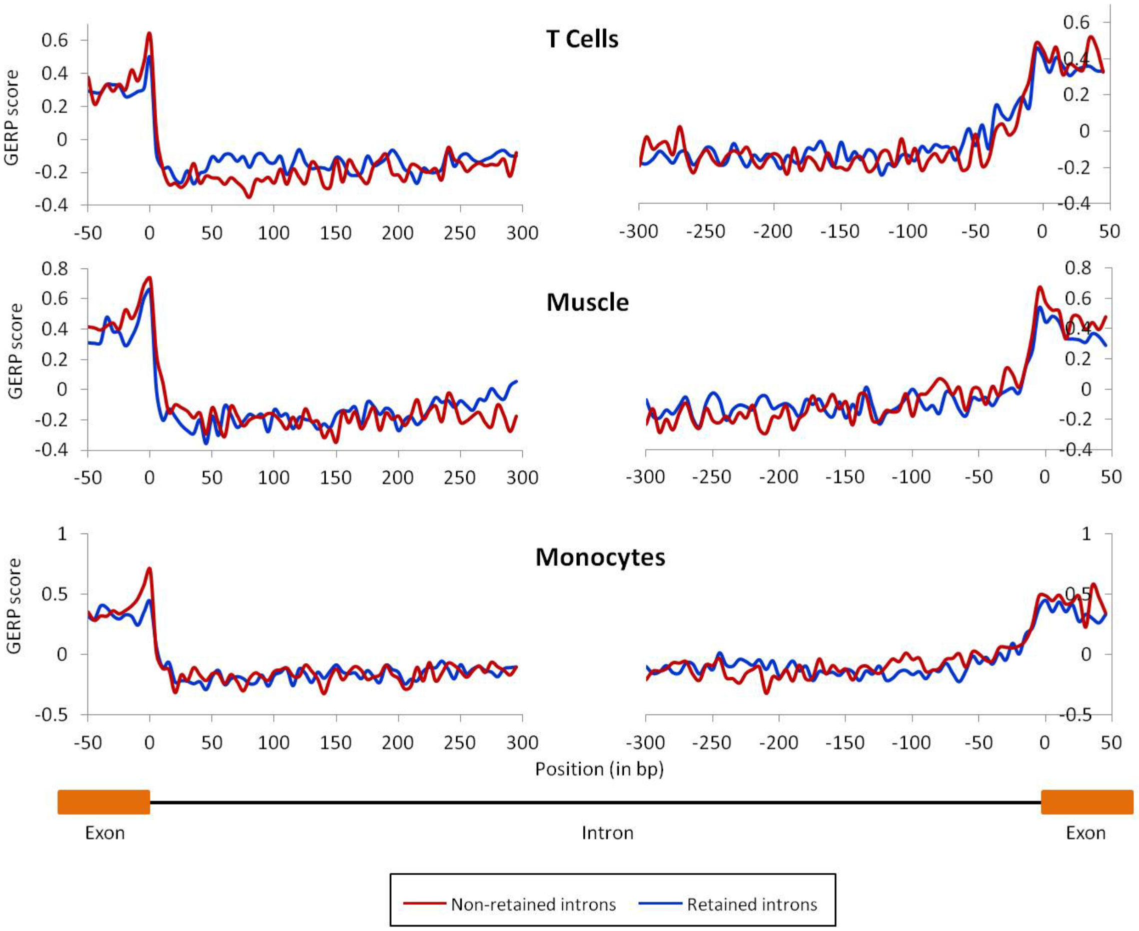


Figure 5.14: Distribution of GERP score in retained and non-retained introns flanked by non-coding exons.

Average GERP conservation scores (5-bp window) around splice sites of retained and non-retained introns flanked by non-coding exons. The *x* axis is the position in bp relative to the splice site, while the *y* axis is the GERP score.

Chapter 6: General discussion and conclusion

6.1 Discussion

Understanding alternative splicing and its regulation is a key component to understanding transcriptome diversity. One way of getting at such an understanding is to find out which variables are associated with alternative splicing, such as epigenetic marks and sequence features. Naively, such associations can be identified using a statistical test to compare the variables of interest between all alternative splicing events and all constitutive events. However, numerous confounders lie hidden in the data, and if not properly addressed can result in both missed associations and spurious associations. Sequencing has revolutionized the analysis of splicing because of its high throughput precision at base pair resolution. Now, with next-generation sequencing techniques (including RNA-seq, ChIP-seq, and whole-genome bisulfite sequencing), it is faster and easier to analyze splicing at multiple levels (e.g at the DNA, RNA or epigenome level) in the entire genome. However, such an integrated analyses are confounded by multiple levels of interdependence in the data. Early studies associating splicing and epigenetic marks have either been limited to individual genes or have not accounted for the potentially confounding influence of gene expression levels and epigenetic mark distribution across gene bodies. This necessitated a more careful investigation of the role of epigenetic marks in splicing at the genome scale. Given the high degree of inter-correlation between exon inclusion, IR, gene expression and epigenetic marks, the possible effects of transcription need to be removed to target the independent association between splicing and epigenetic marks. We also showed that the non-uniform distribution of epigenetic marks across gene bodies is a potential confounder when analyzing the association between splicing and epigenetic marks. In this study, we present an approach that jointly corrects for the confounding effects of the gene expression level and the position of splicing event in the gene body. We performed a comprehensive analysis integrating epigenetic, transcriptomic and sequence features in exon inclusion regulation and IR regulation and how the two latter are related. While it is possible to postulate statistical models that correct for confounding variables, it is not generally possible to fully account for complex or non-linear dependencies. Our simple solution to this problem is to use a “case-control” approach, where each case object is matched to a control object experiencing (in general in the cell population) the same transcriptional conditions. By comparing distinct splicing events within the same gene and with overall identical spatial distribution in the gene, we eliminated bias from gene expression as well as the bias from the non-uniform distribution of epigenetic marks across gene bodies, which are issues that previous studies have not fully resolved. After we corrected for these confounding influences, a range of findings were no longer in agreement with previous studies, namely the differential 5’SS strength and methylation between alternative and constitutive exons, and differential CpG content and methylation between retained and non-retained introns. However, some interesting correlations remained significant.

At the sequence level, we found that alternative exons are accompanied by a higher sequence conservation of the flanking introns, as previously reported [174, 190, 191]. This apparently reflects purifying selection on regulatory motifs according to previous studies. One hypothesis is that those regulatory elements interfere with the splicing machinery to influence exon inclusion level. Gladman *et al* reported such a regulatory mechanism of splicing of the survival motor neuron (*SMN)* gene which is associated with the proximal spinal muscular atrophy. Using mutagenesis, the authors found two conserved regions in intron 7 of the human SMN gene that affect exon 7 splicing [192]. Prior studies reported that alternative exons are generally associated with weaker 3' SS and 5' SS. We found that constitutive exons have statistically stronger 3'SS compared to alternative exons but not 5'SS. The spliceosome, performing the splicing process, is placed around splice sites in each intron. A consensus sequence of each splice site drives its recognition by spliceosomal components. Strong splice sites are those that are more similar to the consensus sequence. They are more efficiently recognized and used in comparison to weak or suboptimal splice sites. In nascent pre-mRNA, the proximity of competing strong and weak splice sites results in alternative splicing [193]. Since alternative and constitutive exons differ in 3'SS strength but not 5'SS strength, we hypothesize that the 3'SS is the most causal splice site and thus needs to be recognized prior to the 5'SS during the splicing process. A plausible explanation supported by previous studies is that the weaker strengths are a consequence of purifying selection to keep the AS sites weaker, may be because strong splice sites are regulated with difficulty. Several experimental studies have shown that strengthening of weak alternative splice sites leads to loss of effective regulation by splicing enhancers and silencers [194-196]. As example, a G → A splice-donor-site mutation at position +3 in the Tau gene has been associated with optimization of a weak alternative site. This naturally occurring mutation, responsible for frontotemporal dementia, increases exon inclusion by reinforcing pairing with U1 RNA or by disrupting the RNA secondary structure [197]. Genome-wide analyses of substitution patterns within orthologous human–mouse–rat splice sites reveal that alternative splice sites are under selection to be weak [198]. Therefore, purifying selection of sequences flanking alternative exons may act to suppress or avoid mutations that strengthen splice site. Our results, along with previous studies, suggest that regulated alternative splicing requires relatively weak splice sites.

At the RNA level, alternative exons display a higher retention level of flanking introns. This is in line with the idea that cassette exons are spliced at a slower rate than constitutive exons [14, 74, 137]. Moreover, it is possible that a coupling between localized IR and RNA pol II, in the vicinity of exon, could contribute to exon inclusion level. Such a mechanism has been proposed to influence gene expression levels, stipulating that increased IR can contribute to decreased expression through a bidirectional cross-regulation mechanism between localized RNA pol II accumulation and impaired recruitment of splicing factors [124]. In the proposed mechanism, low level of expression leads to the reduction of splicing factor recruitment to nascent transcripts. The absence of recruitment of such factors in turn promotes increased IR and localized pausing of RNA pol II over retained introns; reduced Pol II elongation may further increase IR by promoting splicing repressive factors and ultimately transcript turnover. We speculate that at the exon level, a localized accumulation of RNA pol II in introns flanking alternative exons could alter the recruitment of splicing factors and promote the recruitment of splicing repressive factors, and this results in the skipping of the exon. Altogether, these data provide evidence that exon inclusion and IR level of surrounding introns have a unique association that is independent of gene expression influence.

We showed that total RNA-seq can be used for studying nascent RNAs undergoing transcription. We found a large fraction of intronic reads (38%). This high rate of nascent transcripts would not be detectable using poly(A) RNA-seq, indicating that total RNA provides additional insight into transcriptional activity and the global RNA profile of a cell. Braunschweig *et al* previously compared IR level across cell types and found that neural and immune cells contain a higher proportion of IR compared to embryonic stem and muscle cells [124]. Consistently in our study, IR level varies between cell types with the highest levels in T cells and the lowest levels in muscle. This raises several valuable questions: *(i)* Is a high rate of transcription coupled to high turnover of mature transcripts in T cells? *(ii)* Is it possible that a smaller fraction of immature RNA was processed to mRNA in T cells than in muscle or monocyte? *(iii)* Is pre-mRNA more stable in T cells than in muscle or monocyte? However, we cannot assess these possibilities using only total RNA-seq data. Comparing RNA-Seq data from total RNA vs. poly(A)-RNA data from the same sample would provide a quantification of the amount of immature transcript vs. mature transcript. Another question of interest is whether the regulation of gene expression at the pre-mRNA processing level, such as IR, has a specific functional role. We were able to successfully cluster sample types based on IR levels and we found that differentially expressed genes and differentially retained introns differ in enriched-GO biological processes. As previously reported, differential IR was preferentially enriched for genes encoding RNA processing and translation factors [163, 166], whereas differentially expressed genes across white blood and muscle cells were preferentially involved in immune-related processes, as expected. These findings indicate that *(i)* our approach has successfully removed the influence of gene expression in IR analysis, and *(ii)* IR variation has a distinct regulatory signature, in term of the type of genes, compared to transcriptional regulation. A higher IR rate has been previously identified in neural cells [124]; additionally, our results in line with previous findings support a higher level of IR in T cells [124]. Given *(i)* the extensive RNA processing and the transcriptome complexity of neuronal cells, and *(ii)* the coordinated program of gene expression regulation for T cell activation in immune response, we speculate that IR is associated with the regulation of transcriptome complexity. However, the regulation of the transcriptome remains a complex mechanism in which several processes act in concert, such as IR, splicing and epigenetic marks.

At the epigenome level, we found that HMs are not distributed evenly across gene bodies, in accordance with their distinct functions in transcription, as shown before [199, 200]. In particular, the mark of actively transcribed regions H3K36me3 is low in the first exon and sharply increases from the second exon [152, 153], whereas the transcription activation marks H3K4me1, H3K4me3, H3K27ac are higher in the 5’end compared to the rest of the gene. Indeed, we showed that the position of splicing events, in the gene, is a significant confounder of the relationship between splicing and epigenetic marks, due to the non-uniform distribution of epigenetic marks across gene bodies. Introns flanked by non-coding exons, which are preferentially located close to the 5’ end of the gene, have lower levels of H3K36me3 and higher levels of H3K4me1, H3K4me3, and H3K27ac compared to introns flanked by coding exons. However, many early studies didn’t account for this biased distribution of HMs which is a significant confounder [96, 189, 201, 202]. Previous studies, which did not account for these confounders, indicate that H3K27ac, H3K4me1 and H3K4me3 are significantly enriched over retained introns compared to non-retained introns [124, 189]. Conversely, our corrective approach indicates that H3K27ac, H3K4me1 and H3K4me3 are significantly enriched over non-retained introns. By correcting for the position of splicing events, we removed the influence of HM distribution in the assessment of the association between HMs and pre-mRNA processing. In the exon inclusion analysis, after correcting for the effects of transcription and exon position, we found that H3K36me3 is the only HM that is associated with exon inclusion, with an increased occupancy of H3K36me3 over constitutive exons.

Having identified the signatures of HMs (particularly H3K36me3) in pre-mRNA processing, the major challenge remains to understand their biological function. Is the enrichment of certain HMs in exons vs. introns and constitutive vs. alternative exons just an effect of nucleosome density since *(i)* exons display higher nucleosome occupancy compared with their flanking introns [84, 87] and *(ii)* exon inclusion level positively correlates with nucleosome enrichment [203, 204]? If nucleosomes are preferentially positioned at exons, and particularly constitutive exons, is it expected that those exons display more HMs simply because there are more histones to be modified? Tilgner *et al* in their analyses showed that the normalization of H3K36me3 enrichment against nucleosome occupancy leads to the vanishing of H3K36me3 peak within exons [87]. However in that study, nucleosome density upstream of the acceptor sites of strong exons is not correlated with the depletion of H3K36me3. The authors have also observed that some HMs (H4K20me1, for example) show a characteristic exonic pattern even after normalizing for nucleosome density. Therefore, we consider a model that links chromatin and splicing, as previously proposed by several studies. This model argues that exons can be recognized by the splicing machinery through well-positioned nucleosomes carrying H3K36me3, suggesting that HMs may indeed influence splicing.

One possible influence of H3K36me3 on splicing could be the modulation of the recruitment of splicing regulators (splicing factors and/or spliceosome components) promoting the exon recognition. Indeed, such a mechanism has been demonstrated for the H3K36me3-binding protein, Psip1, which recruits members of the splicing regulator family SRSF1 and SRSF3. SRSF1 and SRSF3 bind to cis-elements in the exons and promote exon recognition and splicing by recruiting U1 and U2 snRNPs [205]. An alternative and not mutually exclusive possibility is that splicing enhances the marking of H3K36me3 [152, 153]. Almeida *et al* have shown that the formation of H3K36me3 is directly influenced by splicing [153]. In that study, the authors found that intron containing genes contain considerably higher H3K36me3 marking than intronless genes, irrespective of expression levels and gene size. Most importantly, the authors experimentally inhibited splicing in human cells, and this impairs the recruitment of H3K36 methyltransferase HYPB/Setd2 and decreases the formation of H3K36me3, whereas forcing the inclusion of alternative exons results in the opposite effect of increasing HYPB/Setd2 recruitment and H3K36me3 formation. This supports a model in which co-transcriptional pre-mRNA splicing enhances H3K36me3 by promoting the recruitment of HYPB/Setd2. Based on our findings, along with previous studies, we propose the existence of a bidirectional talk between H3K36me3 and the splicing machinery.

At the methylation level, we observed that the position of splicing event in the gene and the CpG density variation are potential confounders. Introns flanked by non-coding exons are CpG-rich and devoid of methylation because they are generally near promoter regions. Although a number of studies in several species found an association between splicing and DNA methylation, with higher methylation over exons compared to the flanking introns [95, 206, 207], it is unclear whether DNA methylation is involved in splicing regulation, or *(i)* is this a coincidence because coding constraints within exons maintain higher GC density, or *(ii)* is this effect driven by the positional effect of exons in the gene? Singer *et al* in their recent study addressed the issue of CpG variation between gene features. In that study, the authors showed that the non-uniform distribution of CpG associated with varying conservation levels between region types may result in the false detection of differential methylation intensities across these region types [110]. This conservation bias in mCpG calculation has been corrected by simply removing rows with missing CpG average at each location or by applying a matrix completion (estimate the probability of methylation at each site, taking into account the spatial correlation between nearby sites). The two correction methods eliminated previously reported sharp transitions of methylation at exon–intron boundaries. Consistently, our CpG-adjustment method does not support methylation differences between exons and their flanking introns. The implication of methylation in alternative splicing has also been suggested. It has been reported that retained introns have higher CpG and mCpG levels compared to non-retained introns [167], and that alternative exons have lower levels of methylation compared to constitutive ones [104, 167]. However, detecting the methylation effect on pre-mRNA processing can encounter difficulties, particularly when the methylation level can be influenced by many variables such as the position of the methylation site in the gene and the CpG content. For instance, Gelfman *et al* in their genome-wide study compared the methylation profiles of alternative and constitutive exons [104]. In that study, the authors applied the mCpG/CpG ratio to estimate methylation level correcting for GC-content; however the authors did not address the potential positional effect of exons. The authors found that alternative exons had lower DNA methylation levels than constitutive exons. In our study, we investigated the association between mCpG and AS, while controlling for a larger range of confounders including CpG content, gene expression and splicing event position. Contrary to previous studies that did not simultaneously account for these biases, we did not observe a difference in methylation between: *(i)* constitutive and alternative exons and *(ii)* retained and non-retained introns. Therefore our results do not support CpG methylation being primarily associated with alternative splicing.

The goal of this study was to apply an optimal analytical approach to analyze epigenetic and sequences features involved in exon inclusion and IR with putative confounding variables. By correcting for these confounding variables (gene expression levels and splicing event position) we uniquely targeted the relationship between splicing, IR level, epigenetic marks and sequence features. Our approach increases the reliability and robustness of the measured effects. Our approach can be generalized and applied to other studies with similar data structure, for instance to integrated analysis of transcriptome and epigenome in AS in cancer. Now that we have established the requirements to correct for gene expression and AS event position, a next step would be *(i)* to experimentally validate our findings, *(ii)* to correct for more potential confounders, and *(iii)* to investigate more features of interest that might be associated with pre-mRNA processing, such as the RNA pol II occupancy, nucleosome density, and genetic variants affecting splicing or HM formation.

6.2 Future Directions: Towards Experimental Validation

Our accurate transcriptomic and epigenomic screening of AS has identified a distinct relationship between pre-mRNA processing and HMs, demonstrating the power of our approach. An ultimate step would be the experimental confirmation of our findings to help bridge the gap between statistical significance and biological relevance. Validating our findings using experimental assays will provide valuable insight into the mechanics of co-transcriptional splicing. To establish causation link between exon inclusion, IR level and HM, targeted experiments on the sites of interest need to be performed. In the past few years, the first fast, efficient and economical techniques of genetic targeting have emerged. They are based on zinc finger nuclease (ZFN) [208, 209] and transcription activator-like effectors (TALE) [194, 210]. These techniques enable to edit targeted genomic locus via the usage of engineered nucleases that are fused to sequence-specific DNA binding domains. The specificity of the target is provided by the DNA binding domain, and the nuclease cleaves DNA by directing the formation of DNA double-strand breaks (DSBs) at the genomic locus of interest. These DSBs are repaired via homologous recombination (HR) or nonhomologous end-joining (NHEJ) pathway to achieve gene knock-out [211], addition of DNA stretch into a genomic loci [212], gene correction [213], and targeted chromosomal rearrangements such as translocation [214], deletion [215] and duplication [216]. ZFN and TALE techniques can also be used to achieve targeted epigenome editing such as DNA methylation, histone methylation, demethylation and deacethylation [217-220]. However these earlier techniques present some limitations: *(i)* the target specificity relies on protein/DNA recognition which is costly and not readily; *(ii)* for each new target, a large DNA segments (500-1500bp) is required; *(iii)* mutations cannot be introduced in multiple genes at the same time; *(iv)* the long and laborious homologous recombination/embryonic stem cell approach that is required to create targeted mutant.

ZNF and TALE techniques have been outshined, in terms of target design simplicity, multiplexed mutations and efficiency by the latest exciting advance in genome editing technology, known as the CRISPR/Cas9 system. This technique relies on the RNA-guided endonuclease Cas9 (CRISPR-associated protein 9), a component of the type II CRISPR (clustered, regularly interspaced, short palindromic repeats) system of bacterial host defense [221-223]. The cas9 nuclease can be directed to specific region of interest in the genome using sequence complementarity between an engineered guide RNA (gRNA) and the target site [224-226]. Similar to the ZNF and TALE systems, the CRISPR/Cas9 acts via homologous recombination (HR) or nonhomologous end-joining (NHEJ). While CRISPR-Cas9 has been widely used to produce gene editing [227-230], this approach can also be used to indirectly modulate gene expression without editing the genome directly. This relies on a mutated form of Cas9 that lacks nuclease activity (dCas9) [219, 231, 232] ; sgRNA–dCas9, instead of being used to cut the DNA, is used as a scaffold for the recruitment of other modifying enzymes to the targeted site to modulate its function. dCas9 can be fused to various effectors to allow the immunoprecipitation of the bound chromatin, the localization of specific DNA sequences or the repression or activation of transcription. The dCas9 system can also be applied to epigenome editing by modulating the formation of a specific HM at a genomic locus. Recently, Hilton *et al* used the dCas9 system to demonstrate that the recruitment of an acetyltransferase by dCas9 to promoter and enhancer regions, directly modulates the formation of H3K27ac at these regions and activates target-gene expression [233]. With recent developments in the technology, Cas9 system now has the potential to edit RNA [234]. Cas9 can work with short DNA sequences known as “PAM,” for protospacer adjacent motif, to target specific site of single-stranded RNA (ssRNA). In the presence of designed PAMmers (PAM-presenting oligonucleotides) in *trans* as a separate DNA oligonucleotide, Cas9 can be specifically directed to bind or cleaves ssRNA targets that match the sequence of the Cas9-associated gRNA, while avoiding corresponding DNA sequences. RNA-targeting Cas9 (RCas9) has the potential to revolutionize the study of RNA function; for example the modulation of the splicing of a specific exon can be achieved with the nuclease-null RCas9 (dRCas9) by fusing to dRCas9 a splicing repressor or activator domain. To assess the causation links between IR level, exon inclusion and HMs, we propose validation methods based on CRISPR-Cas9 systems.

6.2.1 Functionality of the conserved intronic sequences that flank alternative exons

Based on early studies and our findings, we hypothesize that the conserved intronic regions around alternative exons harbor regulatory elements essential for alternative splicing (low exon inclusion level). To verify the involvement of these conserved intronic sequences in the pathway leading to alternative splicing, we can perform targeted genome editing experiment using the CRISPR/Cas9 system. We can focus on the 50bp intronic regions flanking alternative exons. We can use the CRISPR/Cas9 system, to mutate multiple sites, independently or simultaneously, while avoiding the regions of donor/acceptor splice site and branch point, which are known to be essential for splicing: *(i)* Transfect the cells (monocytes, muscle and T cells) with CRISPR-cas9 plasmids designed to target and mutate the flanking introns of the alternative exon of interest. *(ii)* Isolate total RNA and DNA, and verify the identity of mutations in sites of interest by DNA sequencing, and splicing outcome by qRT-PCR on cDNA. DNA sequencing should reveal that the clones carry the mutations. We expect that these mutations do not affect nascent transcript levels close to the exon/intron boundary, but remarkably increase the exon inclusion level compared to the wild type clones.

6.2.2 Association between exon inclusion and IR level

Our data indicate that exon inclusion is negatively associated with the retention level of flanking introns, and that retained introns are indicative of alternative exon. The major challenge remains to understand the mechanistic of this association. By using the RCas9 technique, we can assess, for any candidate exon, whether IR level in the vicinity affects exon inclusion level. Fusing a splicing repressor or enhancer domain to dRCas9 can allow to control (repress or activate) the inclusion (splicing) of an intron or exon. We can isolate total RNA from cells transfected with the plasmid, and verify the splicing outcome by performing RNA sequencing or qRT-PCR on cDNA. If IR affects exon inclusion level, the dRCas9-based removal of flanking introns should lead to an increased level of the exon, whereas the dRCas9-based retention of flanking introns should reduce the exon inclusion level.

6.2.3 Validation of the association of exon inclusion and IR with HMs

We found that exon inclusion is positively associated with H3K36me3, and that IR level is negatively associated with H3K36me3, H3K27ac, H3K4me1 and H3K4me3. At the DNA level, since CRISPR-Cas9 system could interrogate the causation link between some HMs and transcription [233], it can be also applied to interrogate the relationship between HMs, in particular H3K36me3, and co-transcriptional splicing. Setd2 is known as a histone-H3K36-specific methyltransferase [235], and it has been shown that mutations in setd2 decrease H3K36me3 formation [125, 236]. To assess whether localized H3K36me3 in the vicinity of exon is sufficient to activate exon splicing, we can use CRISPR-dCas9 plasmids fused with setd2 domain to control setd2-dependent formation of H3K36me3 in the vicinity of the exon of interest, and then observe the resulting effect on splicing. To quantify targeted H3K36me3 formation, we can perform chromatin immunoprecipitation with an anti-H3K36me3 antibody followed by quantitative PCR (ChIP-qPCR) in the transfected cells (monocyte, muscle or T cells). We can isolate total RNA from cells transfected with the plasmid, and verify the splicing outcome by performing RNA sequencing or qRT-PCR on cDNA. If H3K36me3 induces splicing, we hypothesize that *(i)* the recruitment of methyltransferase setd2 by dCas9 in the vicinity of the exon will directly modulate the formation of H3K36me3 and further the activation of the exon splicing, and *(ii)* the dCas9-mediated deletion of setd2 in the vicinity of exon will impair H3K36me3 formation and reduce the exon inclusion level. By using the same approaches, we could assess whether localized H3K36me3 has a causative role in IR regulation. Similarly, dCas9 peptide fusion can be used to assess whether localized H3K27ac influences IR level. This can be performed by using a programmable CRISPR-dCas9-based acetyltransferase which consists of the catalytic core of the human acetyltransferase p300 protein fused to dCas9, as previously described [233]. The fusion protein will catalyze the formation of H3K27ac at its target sites. A CRISPR-dCas9-based method for targeted control of H3K4me3 or H3K4me1, has not been documented yet in the literature. However, we speculate that by fusing to dCas9 the appropriate histone methyltransferase domain of the enzymes that catalyze H3K4me3 and H3K4me1, we expect that we can assess whether H3K4me3 or H3K4me1 is causative in intron retention.

At the RNA level, dRCas9 fused to a splicing factor domain can be programmed to modulate the recruitment of splicing factors on the targeted exon to modulate its inclusion level [234]. To test whether cotranscriptional splicing can promote the formation of H3K36me3, the fusion of splicing enhancer domain to dRCas9, that will be directed to the regions adjacent to or inside the exon of interest in the nascent pre-mRNA, should result in an increased formation of localized H3K36me3 in the vicinity in the corresponding DNA sequence. This will imply that splicing activator factor on the nascent pre-mRNA contributes to the formation of H3K36me3, indicating that the splicing machinery can indeed enhance the recruitment of H3K36me3. Conversely, the fusion of a splicing repressor domain to dRCas9 should result in the depletion of H3K36me3 in the vicinity of the exon at the DNA level. By using the same approaches, we could verify whether IR can influence the formation of localized H3K36me3, H3K27ac, H3K4me1 or H3K4me3.

6.3 Conclusion

Collectively, our results show that pre-mRNA splicing and HMs have a distinct relationship beyond the influence of gene expression. To date, AS is postulated to be regulated on at least three different levels: *(i)* At the RNA level, where splicing factors bind to pre-mRNA and modulate the recruitment of the basal splicing machinery onto the splice sites, *(ii)* by the transcription machinery where changes in the elongation rate of RNA pol II influence exon inclusion level, and *(iii)* at the chromatin level, i.e., nucleosome occupancy and HMs. Considering the high inter-correlation between HMs, CpG, sequence features, IR, exon inclusion and gene expression, it is difficult to control for all the confounding factors at the same time. In this study, after controlling for gene expression levels and positional biases, we find that some previously reported variables – such as DNA methylation – have no primary association with pre-mRNA processing. However a number of epigenetic, transcriptomic and sequence features remain significantly associated with levels of exon inclusion and IR. Our analysis is an important step in elucidating those features and will help guide future further work aimed at determining the molecular mechanisms underlying those correlative relationships.

Reference list

1. Beadle, G.W. and E.L. Tatum, *Genetic Control of Biochemical Reactions in Neurospora.* Proc Natl Acad Sci U S A, 1941. **27**(11): p. 499-506.
2. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
3. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
4. Clark, A.G., et al., *Evolution of genes and genomes on the Drosophila phylogeny.* Nature, 2007. **450**(7167): p. 203-18.
5. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
6. *Genome sequence of the nematode C. elegans: a platform for investigating biology.* Science, 1998. **282**(5396): p. 2012-8.
7. Barash, Y., et al., *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-9.
8. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
9. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.
10. Biamonti, G., et al., *The alternative splicing side of cancer.* Semin Cell Dev Biol, 2014. **32**: p. 30-6.
11. David, C.J. and J.L. Manley, *Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged.* Genes Dev, 2010. **24**(21): p. 2343-64.
12. Oltean, S. and D.O. Bates, *Hallmarks of alternative splicing in cancer.* Oncogene, 2014. **33**(46): p. 5311-8.
13. Zhang, J. and J.L. Manley, *Misregulation of pre-mRNA alternative splicing in cancer.* Cancer Discov, 2013. **3**(11): p. 1228-37.
14. Pandya-Jones, A. and D.L. Black, *Co-transcriptional splicing of constitutive and alternative exons.* RNA, 2009. **15**(10): p. 1896-908.
15. Bauren, G. and L. Wieslander, *Splicing of Balbiani ring 1 gene pre-mRNA occurs simultaneously with transcription.* Cell, 1994. **76**(1): p. 183-92.
16. Goldstrohm, A.C., A.L. Greenleaf, and M.A. Garcia-Blanco, *Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing.* Gene, 2001. **277**(1-2): p. 31-47.
17. Shukla, S. and S. Oberdoerffer, *Co-transcriptional regulation of alternative pre-mRNA splicing.* Biochim Biophys Acta, 2012. **1819**(7): p. 673-83.
18. Karin, M., *Too many transcription factors: positive and negative interactions.* New Biol, 1990. **2**(2): p. 126-31.
19. Latchman, D.S., *Transcription factors: an overview.* Int J Biochem Cell Biol, 1997. **29**(12): p. 1305-12.
20. Blanchette, M., et al., *Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.* Genome Res, 2006. **16**(5): p. 656-68.
21. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.

22. Calvo, O. and J.L. Manley, *Strange bedfellows: polyadenylation factors at the promoter.* Genes Dev, 2003. **17**(11): p. 1321-7.
23. McCracken, S., et al., *5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II.* Genes Dev, 1997. **11**(24): p. 3306-18.
24. McCracken, S., et al., *The C-terminal domain of RNA polymerase II couples mRNA processing to transcription.* Nature, 1997. **385**(6614): p. 357-61.
25. Reed, R., *Coupling transcription, splicing and mRNA export.* Curr Opin Cell Biol, 2003. **15**(3): p. 326-31.
26. Neugebauer, K.M., *On the importance of being co-transcriptional.* J Cell Sci, 2002. **115**(Pt 20): p. 3865-71.
27. Shatkin, A.J. and J.L. Manley, *The ends of the affair: capping and polyadenylation.* Nat Struct Biol, 2000. **7**(10): p. 838-42.
28. Lewis, J.D. and E. Izaurralde, *The role of the cap structure in RNA processing and nuclear export.* Eur J Biochem, 1997. **247**(2): p. 461-9.
29. Visa, N., et al., *A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export.* J Cell Biol, 1996. **133**(1): p. 5-14.
30. Gao, M., et al., *Interaction between a poly(A)-specific ribonuclease and the 5' cap influences mRNA deadenylation rates in vitro.* Mol Cell, 2000. **5**(3): p. 479-88.
31. Burkard, K.T. and J.S. Butler, *A nuclear 3'-5' exonuclease involved in mRNA degradation interacts with Poly(A) polymerase and the hnRNA protein Npl3p.* Mol Cell Biol, 2000. **20**(2): p. 604-16.
32. Evdokimova, V., et al., *The major mRNA-associated protein YB-1 is a potent 5' cap-dependent mRNA stabilizer.* EMBO J, 2001. **20**(19): p. 5491-502.
33. Konarska, M.M., R.A. Padgett, and P.A. Sharp, *Recognition of cap structure in splicing in vitro of mRNA precursors.* Cell, 1984. **38**(3): p. 731-6.
34. Shatkin, A.J., *Capping of eucaryotic mRNAs.* Cell, 1976. **9**(4 PT 2): p. 645-53.
35. Sonenberg, N. and A.C. Gingras, *The mRNA 5' cap-binding protein eIF4E and control of cell growth.* Curr Opin Cell Biol, 1998. **10**(2): p. 268-75.
36. Banerjee, A.K., *5'-terminal cap structure in eucaryotic messenger ribonucleic acids.* Microbiol Rev, 1980. **44**(2): p. 175-205.
37. Proudfoot, N., *Connecting transcription to messenger RNA processing.* Trends Biochem Sci, 2000. **25**(6): p. 290-3.
38. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.
39. Blaustein, M., F. Pelisch, and A. Srebrow, *Signals, pathways and splicing regulation.* Int J Biochem Cell Biol, 2007. **39**(11): p. 2031-48.
40. Matera, A.G. and Z. Wang, *A day in the life of the spliceosome.* Nat Rev Mol Cell Biol, 2014. **15**(2): p. 108-21.
41. Taggart, A.J., et al., *Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo.* Nat Struct Mol Biol, 2012. **19**(7): p. 719-21.
42. Corvelo, A., et al., *Genome-wide association between branch point properties and alternative splicing.* PLoS Comput Biol, 2010. **6**(11): p. e1001016.
43. Cheng, Z. and T.M. Menees, *RNA splicing and debranching viewed through analysis of RNA lariats.* Mol Genet Genomics, 2011. **286**(5-6): p. 395-410.
44. Hesselberth, J.R., *Lives that introns lead after splicing.* Wiley Interdiscip Rev RNA, 2013. **4**(6): p. 677-91.

45. Dietrich, R.C., R. Incorvaia, and R.A. Padgett, *Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns.* Mol Cell, 1997. **1**(1): p. 151-60.
46. Alioto, T.S., *U12DB: a database of orthologous U12-type spliceosomal introns.* Nucleic Acids Res, 2007. **35**(Database issue): p. D110-5.
47. Sheth, N., et al., *Comprehensive splice-site analysis using comparative genomics.* Nucleic Acids Res, 2006. **34**(14): p. 3955-67.
48. Tarn, W.Y. and J.A. Steitz, *Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns.* Science, 1996. **273**(5283): p. 1824-32.
49. Beaudoing, E., et al., *Patterns of variant polyadenylation signal usage in human genes.* Genome Res, 2000. **10**(7): p. 1001-10.
50. Tian, B., et al., *A large-scale analysis of mRNA polyadenylation of human and mouse genes.* Nucleic Acids Res, 2005. **33**(1): p. 201-12.
51. Bienroth, S., W. Keller, and E. Wahle, *Assembly of a processive messenger RNA polyadenylation complex.* EMBO J, 1993. **12**(2): p. 585-94.
52. Mandel, C.R., Y. Bai, and L. Tong, *Protein factors in pre-mRNA 3'-end processing.* Cell Mol Life Sci, 2008. **65**(7-8): p. 1099-122.
53. Guhaniyogi, J. and G. Brewer, *Regulation of mRNA stability in mammalian cells.* Gene, 2001. **265**(1-2): p. 11-23.
54. F, C., *The genetic code*, in *What mad pursuit: a personal view of scientific discovery*1988, Basic Books: New York. p. 89-101.
55. Cooper, G.M., *The Cell: A Molecular Approach. 2nd edition*, ed. Sunderland(MA)2000: Sinauer Associates.
56. Scheper, G.C., M.S. van der Knaap, and C.G. Proud, *Translation matters: protein synthesis defects in inherited disease.* Nat Rev Genet, 2007. **8**(9): p. 711-23.
57. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.
58. Curradi, M., et al., *Molecular mechanisms of gene silencing mediated by DNA methylation.* Mol Cell Biol, 2002. **22**(9): p. 3157-73.
59. Baylin, S.B., *DNA methylation and gene silencing in cancer.* Nat Clin Pract Oncol, 2005. **2 Suppl 1**: p. S4-11.
60. Keren, H., G. Lev-Maor, and G. Ast, *Alternative splicing and evolution: diversification, exon definition and function.* Nat Rev Genet, 2010. **11**(5): p. 345-55.
61. Sonenberg, N. and A.G. Hinnebusch, *Regulation of translation initiation in eukaryotes: mechanisms and biological targets.* Cell, 2009. **136**(4): p. 731-45.
62. Buchan, J.R. and R. Parker, *Eukaryotic stress granules: the ins and outs of translation.* Mol Cell, 2009. **36**(6): p. 932-41.
63. Harding, H.P., et al., *Regulated translation initiation controls stress-induced gene expression in mammalian cells.* Mol Cell, 2000. **6**(5): p. 1099-108.
64. Jensen, O.N., *Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry.* Curr Opin Chem Biol, 2004. **8**(1): p. 33-41.
65. Munoz, M.J., M. de la Mata, and A.R. Kornblihtt, *The carboxy terminal domain of RNA polymerase II and alternative splicing.* Trends Biochem Sci, 2010. **35**(9): p. 497-504.
66. Phatnani, H.P. and A.L. Greenleaf, *Phosphorylation and functions of the RNA polymerase II CTD.* Genes Dev, 2006. **20**(21): p. 2922-36.
67. Kim, E., et al., *Splicing factors associate with hyperphosphorylated RNA polymerase II in the absence of pre-mRNA.* J Cell Biol, 1997. **136**(1): p. 19-28.
68. de la Mata, M. and A.R. Kornblihtt, *RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20.* Nat Struct Mol Biol, 2006. **13**(11): p. 973-80.

69. Monsalve, M., et al., *Direct coupling of transcription and mRNA processing through the thermogenic coactivator PGC-1.* Mol Cell, 2000. **6**(2): p. 307-16.
70. Nogues, G., et al., *Transcriptional activators differ in their abilities to control alternative splicing.* J Biol Chem, 2002. **277**(45): p. 43110-4.
71. Schor, I.E., et al., *Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing.* Proc Natl Acad Sci U S A, 2009. **106**(11): p. 4325-30.
72. Kadener, S., et al., *Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing.* EMBO J, 2001. **20**(20): p. 5759-68.
73. de la Mata, M., et al., *A slow RNA polymerase II affects alternative splicing in vivo.* Mol Cell, 2003. **12**(2): p. 525-32.
74. de la Mata, M., C. Lafaille, and A.R. Kornblihtt, *First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal.* RNA, 2010. **16**(5): p. 904-12.
75. Roberts, G.C., et al., *Co-transcriptional commitment to alternative splice site selection.* Nucleic Acids Res, 1998. **26**(24): p. 5568-72.
76. Nogues, G., M.J. Munoz, and A.R. Kornblihtt, *Influence of polymerase II processivity on alternative splicing depends on splice site strength.* J Biol Chem, 2003. **278**(52): p. 52166-71.
77. Ip, J.Y., et al., *Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation.* Genome Res, 2011. **21**(3): p. 390-401.
78. Kadener, S., et al., *Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation.* Proc Natl Acad Sci U S A, 2002. **99**(12): p. 8185-90.
79. Shukla, S., et al., *CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.* Nature, 2011. **479**(7371): p. 74-9.
80. Oberdoerffer, S., *A conserved role for intragenic DNA methylation in alternative pre-mRNA splicing.* Transcription, 2012. **3**(3): p. 106-9.
81. *Bam to wiggle*. Available from: https://github.com/chapmanb/bcbb/blob/master/nextgen/scripts/bam_to_wiggle.py.
82. Simon, J.M., et al., *Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects.* Genome Res, 2014. **24**(2): p. 241-50.
83. Schwartz, S., E. Meshorer, and G. Ast, *Chromatin organization marks exon-intron structure.* Nat Struct Mol Biol, 2009. **16**(9): p. 990-5.
84. Andersson, R., et al., *Nucleosomes are well positioned in exons and carry characteristic histone modifications.* Genome Res, 2009. **19**(10): p. 1732-41.
85. Spies, N., et al., *Biased chromatin signatures around polyadenylation sites and exons.* Mol Cell, 2009. **36**(2): p. 245-54.
86. Berget, S.M., *Exon recognition in vertebrate splicing.* J Biol Chem, 1995. **270**(6): p. 2411-4.
87. Tilgner, H., et al., *Nucleosome positioning as a determinant of exon recognition.* Nat Struct Mol Biol, 2009. **16**(9): p. 996-1001.
88. Petesch, S.J. and J.T. Lis, *Overcoming the nucleosome barrier during transcript elongation.* Trends Genet, 2012. **28**(6): p. 285-94.
89. Schwartz, S. and G. Ast, *Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing.* EMBO J, 2010. **29**(10): p. 1629-36.

90. Izban, M.G. and D.S. Luse, *Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing.* Genes Dev, 1991. **5**(4): p. 683-96.

91. Karlic, R., et al., *Histone modification levels are predictive for gene expression.* Proc Natl Acad Sci U S A, 2010. **107**(7): p. 2926-31.

92. Dhami, P., et al., *Complex exon-intron marking by histone modifications is not determined solely by nucleosome distribution.* PLoS One, 2010. **5**(8): p. e12339.

93. Hon, G.C., R.D. Hawkins, and B. Ren, *Predictive chromatin signatures in the mammalian genome.* Hum Mol Genet, 2009. **18**(R2): p. R195-201.

94. Hodges, E., et al., *High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing.* Genome Res, 2009. **19**(9): p. 1593-605.

95. Chodavarapu, R.K., et al., *Relationship between nucleosome positioning and DNA methylation.* Nature, 2010. **466**(7304): p. 388-92.

96. Luco, R.F., et al., *Regulation of alternative splicing by histone modifications.* Science, 2010. **327**(5968): p. 996-1000.

97. Kornblihtt, A.R., et al., *Alternative splicing: a pivotal step between eukaryotic transcription and translation.* Nat Rev Mol Cell Biol, 2013. **14**(3): p. 153-65.

98. Kolasinska-Zwierz, P., et al., *Differential chromatin marking of introns and expressed exons by H3K36me3.* Nat Genet, 2009. **41**(3): p. 376-81.

99. Fazi, F., et al., *A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis.* Cell, 2005. **123**(5): p. 819-31.

100. Xu, J.F., et al., *Association of GPR126 gene polymorphism with adolescent idiopathic scoliosis in Chinese populations.* Genomics, 2015. **105**(2): p. 101-7.

101. Sarraf, S.A. and I. Stancheva, *Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly.* Mol Cell, 2004. **15**(4): p. 595-605.

102. Klose, R.J. and A.P. Bird, *Genomic DNA methylation: the mark and its mediators.* Trends Biochem Sci, 2006. **31**(2): p. 89-97.

103. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.

104. Gelfman, S., et al., *DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure.* Genome Res, 2013. **23**(5): p. 789-99.

105. Choi, J.K., *Contrasting chromatin organization of CpG islands and exons in the human genome.* Genome biology, 2010. **11**(7): p. R70.

106. Feng, S., et al., *Conservation and divergence of methylation patterning in plants and animals.* Proc Natl Acad Sci U S A, 2010. **107**(19): p. 8689-94.

107. Gao, F., et al., *Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis.* Genome biology, 2012. **13**(10): p. R100.

108. Flores, K., et al., *Genome-wide association between DNA methylation and alternative splicing in an invertebrate.* BMC Genomics, 2012. **13**: p. 480.

109. Choi, J.K., *Contrasting chromatin organization of CpG islands and exons in the human genome.* Genome Biol, 2010. **11**(7): p. R70.

110. Singer, M. and L. Pachter, *Controlling for conservation in genome-wide DNA methylation studies.* BMC Genomics, 2015. **16**: p. 420.

111. Hare, M.P. and S.R. Palumbi, *High intron sequence conservation across three mammalian orders suggests functional constraints.* Mol Biol Evol, 2003. **20**(6): p. 969-78.

112. Beaulieu, E., et al., *Identification of a novel cell type-specific intronic enhancer of macrophage migration inhibitory factor (MIF) and its regulation by mithramycin.* Clin Exp Immunol, 2011. **163**(2): p. 178-88.

113. Bianchi, M., et al., *A potent enhancer element in the 5'-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene.* Gene, 2009. **448**(1): p. 88-101.

114. Scohy, S., et al., *Identification of an enhancer and an alternative promoter in the first intron of the alpha-fetoprotein gene.* Nucleic Acids Res, 2000. **28**(19): p. 3743-51.

115. Tourmente, S., et al., *Enhancer and silencer elements within the first intron mediate the transcriptional regulation of the beta 3 tubulin gene by 20-hydroxyecdysone in Drosophila Kc cells.* Insect Biochem Mol Biol, 1993. **23**(1): p. 137-43.

116. Gaunitz, F., et al., *An intronic silencer element is responsible for specific zonal expression of glutamine synthetase in the rat liver.* Hepatology, 2005. **41**(6): p. 1225-32.

117. Gaunitz, F., K. Heise, and R. Gebhardt, *A silencer element in the first intron of the glutamine synthetase gene represses induction by glucocorticoids.* Mol Endocrinol, 2004. **18**(1): p. 63-9.

118. Zhang, G.R., et al., *The vesicular glutamate transporter-1 upstream promoter and first intron each support glutamatergic-specific expression in rat postrhinal cortex.* Brain Res, 2011. **1377**: p. 1-12.

119. Bornstein, P., et al., *Interactions between the promoter and first intron are involved in transcriptional control of alpha 1(I) collagen gene expression.* Mol Cell Biol, 1988. **8**(11): p. 4851-7.

120. Bradnam, K.R. and I. Korf, *Longer first introns are a general property of eukaryotic gene structure.* PLoS One, 2008. **3**(8): p. e3093.

121. Majewski, J. and J. Ott, *Distribution and characterization of regulatory elements in the human genome.* Genome Res, 2002. **12**(12): p. 1827-36.

122. Rigo, F. and H.G. Martinson, *Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage.* Mol Cell Biol, 2008. **28**(2): p. 849-62.

123. Wong, J.J., et al., *Orchestrated intron retention regulates normal granulocyte differentiation.* Cell, 2013. **154**(3): p. 583-95.

124. Braunschweig, U., et al., *Widespread intron retention in mammals functionally tunes transcriptomes.* Genome Res, 2014. **24**(11): p. 1774-86.

125. Edmunds, J.W., L.C. Mahadevan, and A.L. Clayton, *Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation.* The EMBO journal, 2008. **27**(2): p. 406-20.

126. Koch, C.M., et al., *The landscape of histone modifications across 1% of the human genome in five human cell lines.* Genome Res, 2007. **17**(6): p. 691-707.

127. Krivtsov, A.V., et al., *H3K79 methylation profiles define murine and human MLL-AF4 leukemias.* Cancer Cell, 2008. **14**(5): p. 355-68.

128. Roh, T.Y., S. Cuddapah, and K. Zhao, *Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping.* Genes Dev, 2005. **19**(5): p. 542-52.

129. Guenther, M.G., et al., *A chromatin landmark and transcription initiation at most promoters in human cells.* Cell, 2007. **130**(1): p. 77-88.

130. *McGill Epigenomics Mapping Centre Portal*. Available from: http://epigenomesportal.ca.

131. Bonasio, R., S. Tu, and D. Reinberg, *Molecular signals of epigenetic states.* Science, 2010. **330**(6004): p. 612-6.

132. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.
133. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-12.
134. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-83.
135. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.
136. Peters, A.H., et al., *Partitioning and plasticity of repressive histone methylation states in mammalian chromatin.* Mol Cell, 2003. **12**(6): p. 1577-89.
137. Ameur, A., et al., *Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain.* Nat Struct Mol Biol, 2011. **18**(12): p. 1435-40.
138. Lohse, M., et al., *RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W622-7.
139. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
140. Feng, J., et al., *Identifying ChIP-seq enrichment using MACS.* Nat Protoc, 2012. **7**(9): p. 1728-40.
141. Johnson, M.D., et al., *Single nucleotide analysis of cytosine methylation by whole-genome shotgun bisulfite sequencing.* Curr Protoc Mol Biol, 2012. **Chapter 21**: p. Unit21 23.
142. *nxtgen-utils*. Available from: http://code.google.com/p/nxtgen-utils/downloads/list.
143. *Picard*. Available from: http://picard.sourceforge.net
144. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
145. Liu, Y., et al., *Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data.* Genome biology, 2012. **13**(7): p. R61.
146. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
147. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology, 2009. **10**(3): p. R25.
148. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
149. Dreszer, T.R., et al., *The UCSC Genome Browser database: extensions and updates 2011.* Nucleic Acids Res, 2012. **40**(Database issue): p. D918-23.
150. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome biology, 2010. **11**(10): p. R106.
151. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621-8.
152. Huff, J.T., et al., *Reciprocal intronic and exonic histone modification regions in humans.* Nat Struct Mol Biol, 2010. **17**(12): p. 1495-9.
153. de Almeida, S.F., et al., *Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36.* Nat Struct Mol Biol, 2011. **18**(9): p. 977-83.
154. Galante, P.A., et al., *Detection and evaluation of intron retention events in the human transcriptome.* RNA, 2004. **10**(5): p. 757-65.
155. Seong, M.W., et al., *Deleterious c-Cbl Exon Skipping Contributes to Human Glioma.* Neoplasia, 2015. **17**(6): p. 518-24.

156. Kalsotra, A. and T.A. Cooper, *Functional consequences of developmentally regulated alternative splicing.* Nat Rev Genet, 2011. **12**(10): p. 715-29.

157. Irimia, M. and B.J. Blencowe, *Alternative splicing: decoding an expansive regulatory layer.* Curr Opin Cell Biol, 2012. **24**(3): p. 323-32.

158. Loh, T.J., et al., *CD44 alternative splicing and hnRNP A1 expression are associated with the metastasis of breast cancer.* Oncol Rep, 2015.

159. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

160. Sakabe, N.J. and S.J. de Souza, *Sequence features responsible for intron retention in human.* BMC Genomics, 2007. **8**: p. 59.

161. Pheasant, M. and J.S. Mattick, *Raising the estimate of functional human sequences.* Genome Res, 2007. **17**(9): p. 1245-53.

162. Cho, V., et al., *The RNA-binding protein hnRNPLL induces a T cell alternative splicing program delineated by differential intron retention in polyadenylated RNA.* Genome biology, 2014. **15**(1): p. R26.

163. Boutz, P.L., A. Bhutkar, and P.A. Sharp, *Detained introns are a novel, widespread class of post-transcriptionally spliced introns.* Genes Dev, 2015. **29**(1): p. 63-80.

164. Shalgi, R., et al., *Widespread regulation of translation by elongation pausing in heat shock.* Mol Cell, 2013. **49**(3): p. 439-52.

165. Yap, K., et al., *Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention.* Genes Dev, 2012. **26**(11): p. 1209-23.

166. Dvinge, H. and R.K. Bradley, *Widespread intron retention diversifies most cancer transcriptomes.* Genome Med, 2015. **7**(1): p. 45.

167. Gascard, P., et al., *Epigenetic and transcriptional determinants of the human breast.* Nat Commun, 2015. **6**: p. 6351.

168. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.

169. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.

170. Kuhn, R.M., D. Haussler, and W.J. Kent, *The UCSC genome browser and associated tools.* Brief Bioinform, 2013. **14**(2): p. 144-61.

171. Meyer, L.R., et al., *The UCSC Genome Browser database: extensions and updates 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D64-9.

172. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol, 2004. **11**(2-3): p. 377-94.

173. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

174. Sorek, R. and G. Ast, *Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.* Genome Res, 2003. **13**(7): p. 1631-7.

175. Sugnet, C.W., et al., *Unusual intron conservation near tissue-regulated exons found by splicing microarrays.* PLoS Comput Biol, 2006. **2**(1): p. e4.

176. Itoh, H., T. Washio, and M. Tomita, *Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes.* RNA, 2004. **10**(7): p. 1005-18.

177. Khodor, Y.L., et al., *Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila.* Genes Dev, 2011. **25**(23): p. 2502-12.

178. Vargas, D.Y., et al., *Single-molecule imaging of transcriptionally coupled and uncoupled splicing.* Cell, 2011. **147**(5): p. 1054-65.

179. Pandya-Jones, A., et al., *Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression.* RNA, 2013. **19**(6): p. 811-27.

180. Kurio, H., et al., *Intron retention generates a novel isoform of CEACAM6 that may act as an adhesion molecule in the ectoplasmic specialization structures between spermatids and sertoli cells in rat testis.* Biol Reprod, 2008. **79**(6): p. 1062-73.

181. Forrest, S.T., et al., *Intron retention generates a novel Id3 isoform that inhibits vascular lesion formation.* J Biol Chem, 2004. **279**(31): p. 32897-903.

182. Bell, T.J., et al., *Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations.* Proc Natl Acad Sci U S A, 2010. **107**(49): p. 21152-7.

183. Xie, J. and D.P. McCobb, *Control of alternative splicing of potassium channels by stress hormones.* Science, 1998. **280**(5362): p. 443-6.

184. Tian, L., et al., *Distinct stoichiometry of BKCa channel tetramer phosphorylation specifies channel activation and inhibition by cAMP-dependent protein kinase.* Proc Natl Acad Sci U S A, 2004. **101**(32): p. 11897-902.

185. Sowalsky, A.G., et al., *Whole transcriptome sequencing reveals extensive unspliced mRNA in metastatic castration-resistant prostate cancer.* Mol Cancer Res, 2015. **13**(1): p. 98-106.

186. Mi, H., et al., *Large-scale gene function analysis with the PANTHER classification system.* Nat Protoc, 2013. **8**(8): p. 1551-66.

187. Mi, H., A. Muruganujan, and P.D. Thomas, *PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees.* Nucleic Acids Res, 2013. **41**(Database issue): p. D377-86.

188. Zhao, W., et al., *Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling.* BMC Genomics, 2014. **15**: p. 419.

189. Zhou, Y., Y. Lu, and W. Tian, *Epigenetic features are significantly associated with alternative splicing.* BMC Genomics, 2012. **13**: p. 123.

190. Sugnet, C.W., et al., *Transcriptome and genome conservation of alternative splicing events in humans and mice.* Pac Symp Biocomput, 2004: p. 66-77.

191. Yeo, G.W., et al., *Identification and analysis of alternative splicing events conserved in human and mouse.* Proc Natl Acad Sci U S A, 2005. **102**(8): p. 2850-5.

192. Gladman, J.T. and D.S. Chandler, *Intron 7 conserved sequence elements regulate the splicing of the SMN genes.* Human genetics, 2009. **126**(6): p. 833-41.

193. Ast, G., *How did alternative splicing evolve?* Nat Rev Genet, 2004. **5**(10): p. 773-82.

194. Muro, A.F., A. Iaconcig, and F.E. Baralle, *Regulation of the fibronectin EDA exon alternative splicing. Cooperative role of the exonic enhancer element and the 5' splicing site.* FEBS Lett, 1998. **437**(1-2): p. 137-41.

195. Zheng, Z.M., et al., *Optimization of a weak 3' splice site counteracts the function of a bovine papillomavirus type 1 exonic splicing suppressor in vitro and in vivo.* J Virol, 2000. **74**(13): p. 5902-10.

196. Dirksen, W.P., Q. Sun, and F.M. Rottman, *Multiple splicing signals control alternative intron retention of bovine growth hormone pre-mRNA.* J Biol Chem, 1995. **270**(10): p. 5346-52.

197. Neumann, M., et al., *A new family with frontotemporal dementia with intronic 10+3 splice site mutation in the tau gene: neuropathology and molecular effects.* Neuropathol Appl Neurobiol, 2005. **31**(4): p. 362-73.

198. Garg, K. and P. Green, *Differing patterns of selection in alternative and constitutive splice sites.* Genome Res, 2007. **17**(7): p. 1015-22.

199. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

200. Maunakea, A.K., I. Chepelev, and K. Zhao, *Epigenome mapping in normal and disease States.* Circ Res, 2010. **107**(3): p. 327-39.

201. Shindo, Y., et al., *Computational analysis of associations between alternative splicing and histone modifications.* FEBS Lett, 2013. **587**(5): p. 516-21.

202. Liu, H., et al., *Histone modifications involved in cassette exon inclusions: a quantitative and interpretable analysis.* BMC Genomics, 2014. **15**: p. 1148.

203. Iannone, C., et al., *Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells.* RNA, 2015. **21**(3): p. 360-74.

204. Keren-Shaul, H., G. Lev-Maor, and G. Ast, *Pre-mRNA splicing is a determinant of nucleosome organization.* PLoS One, 2013. **8**(1): p. e53506.

205. Pradeepa, M.M., et al., *Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing.* PLoS Genet, 2012. **8**(5): p. e1002717.

206. Hodges, C., et al., *Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II.* Science, 2009. **325**(5940): p. 626-8.

207. Lyko, F., et al., *The honey bee epigenomes: differential methylation of brain DNA in queens and workers.* PLoS Biol, 2010. **8**(11): p. e1000506.

208. Gersbach, C.A., T. Gaj, and C.F. Barbas, 3rd, *Synthetic zinc finger proteins: the advent of targeted gene regulation and genome modification technologies.* Acc Chem Res, 2014. **47**(8): p. 2309-18.

209. Klug, A., *The discovery of zinc fingers and their applications in gene regulation and genome manipulation.* Annu Rev Biochem, 2010. **79**: p. 213-31.

210. Boch, J., et al., *Breaking the code of DNA binding specificity of TAL-type III effectors.* Science, 2009. **326**(5959): p. 1509-12.

211. Santiago, Y., et al., *Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases.* Proc Natl Acad Sci U S A, 2008. **105**(15): p. 5809-14.

212. Moehle, E.A., et al., *Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3055-60.

213. Urnov, F.D., et al., *Highly efficient endogenous human gene correction using designed zinc-finger nucleases.* Nature, 2005. **435**(7042): p. 646-51.

214. Brunet, E., et al., *Chromosomal translocations induced at specified loci in human stem cells.* Proc Natl Acad Sci U S A, 2009. **106**(26): p. 10620-5.

215. Lee, H.J., E. Kim, and J.S. Kim, *Targeted chromosomal deletions in human cells using zinc finger nucleases.* Genome Res, 2010. **20**(1): p. 81-9.

216. Lee, H.J., et al., *Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases.* Genome Res, 2012. **22**(3): p. 539-48.

217. Rivenbark, A.G., et al., *Epigenetic reprogramming of cancer cells via targeted DNA methylation.* Epigenetics, 2012. **7**(4): p. 350-60.

218. Snowden, A.W., et al., *Gene-specific targeting of H3K9 methylation is sufficient for initiating repression in vivo.* Curr Biol, 2002. **12**(24): p. 2159-66.

219. Maeder, M.L., et al., *Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins.* Nat Biotechnol, 2013. **31**(12): p. 1137-42.

220. Mendenhall, E.M., et al., *Locus-specific editing of histone modifications at endogenous enhancers.* Nat Biotechnol, 2013. **31**(12): p. 1133-6.
221. Hsu, P.D., E.S. Lander, and F. Zhang, *Development and applications of CRISPR-Cas9 for genome engineering.* Cell, 2014. **157**(6): p. 1262-78.
222. Bortesi, L. and R. Fischer, *The CRISPR/Cas9 system for plant genome editing and beyond.* Biotechnol Adv, 2015. **33**(1): p. 41-52.
223. Maeder, M.L. and C.A. Gersbach, *Genome Editing Technologies for Gene and Cell Therapy.* Mol Ther, 2016.
224. Cong, L., et al., *Multiplex genome engineering using CRISPR/Cas systems.* Science, 2013. **339**(6121): p. 819-23.
225. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science, 2012. **337**(6096): p. 816-21.
226. Mali, P., et al., *RNA-guided human genome engineering via Cas9.* Science, 2013. **339**(6121): p. 823-6.
227. Zhou, Y., et al., *High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells.* Nature, 2014. **509**(7501): p. 487-91.
228. Koike-Yusa, H., et al., *Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library.* Nat Biotechnol, 2014. **32**(3): p. 267-73.
229. Shalem, O., et al., *Genome-scale CRISPR-Cas9 knockout screening in human cells.* Science, 2014. **343**(6166): p. 84-7.
230. Wang, T., et al., *Genetic screens in human cells using the CRISPR-Cas9 system.* Science, 2014. **343**(6166): p. 80-4.
231. Cheng, A.W., et al., *Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system.* Cell Res, 2013. **23**(10): p. 1163-71.
232. Gilbert, L.A., et al., *CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes.* Cell, 2013. **154**(2): p. 442-51.
233. Hilton, I.B., et al., *Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers.* Nat Biotechnol, 2015. **33**(5): p. 510-7.
234. O'Connell, M.R., et al., *Programmable RNA recognition and cleavage by CRISPR/Cas9.* Nature, 2014. **516**(7530): p. 263-6.
235. Sun, X.J., et al., *Identification and characterization of a novel human histone H3 lysine 36-specific methyltransferase.* J Biol Chem, 2005. **280**(42): p. 35261-71.
236. Zhu, X., et al., *Identification of functional cooperative mutations of SETD2 in human acute leukemia.* Nat Genet, 2014. **46**(3): p. 287-93.

Appendix

Supplementary figures

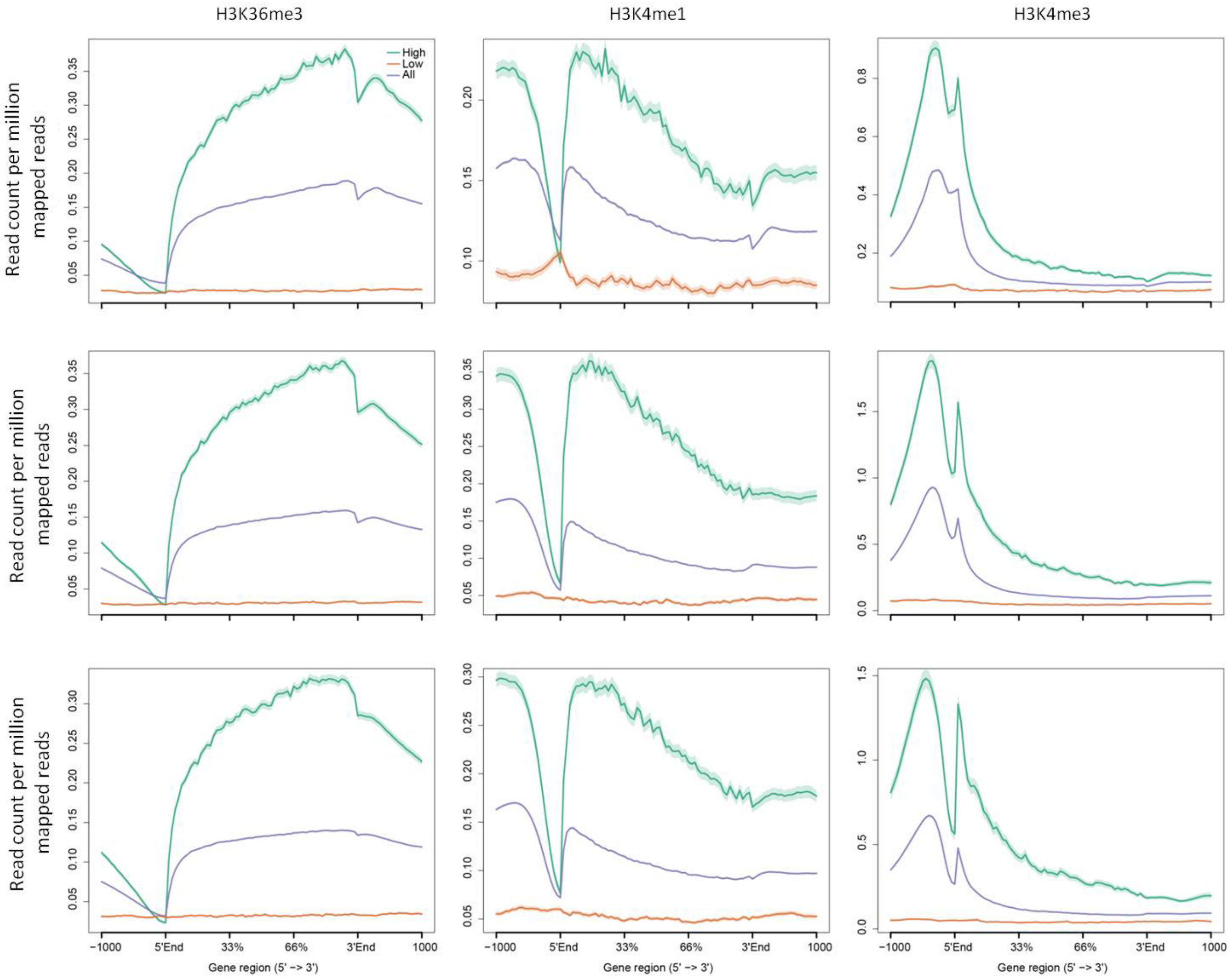


Figure 0.1: Composite plots of patterns of H3K36me3, H3K4me1 and H3K4me3 ChIP-seq signals across genic regions, in muscle.

H3K36me3, H3K4me1 and H3K4me3 are plotted across genic regions (*x* axis), of either the 1000 highly expressed genes (green), the 1000 lowly expressed genes (orange) or all genes (purple), in three muscle samples. Data represented as read count per million mapped reads (*y* axis).

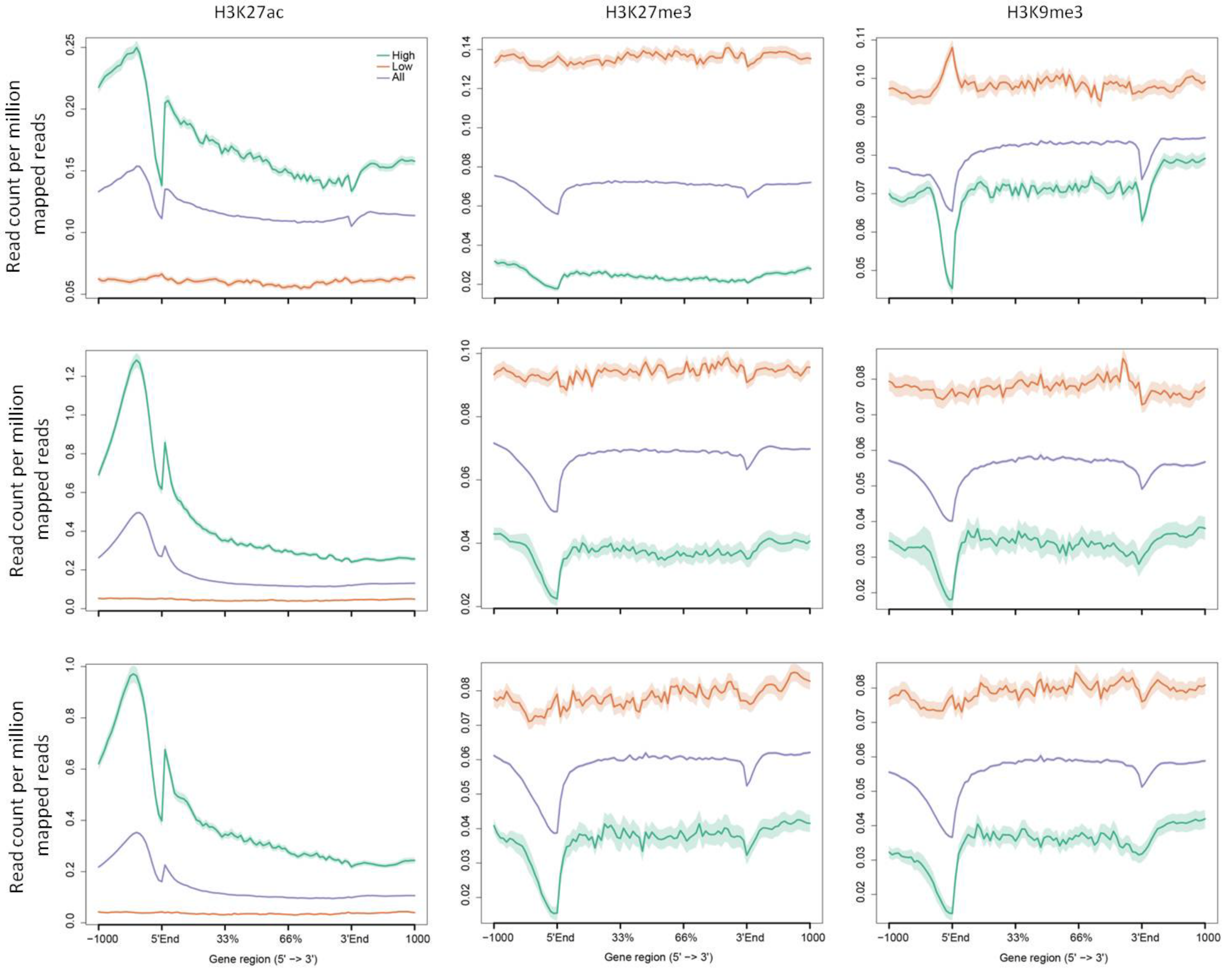


Figure 0.2: Composite plots of patterns of H3K27ac, H3K27me3 and H3K9me3 ChIP-seq signals across genic regions, in muscle.

H3K27ac, H3K27me3 and H3K9me3 are plotted across genic regions (*x* axis), of either the 1000 highly expressed genes (green), the 1000 lowly expressed genes (orange) or all genes (purple), in three muscle samples. Data represented as read count per million mapped reads (*y* axis).

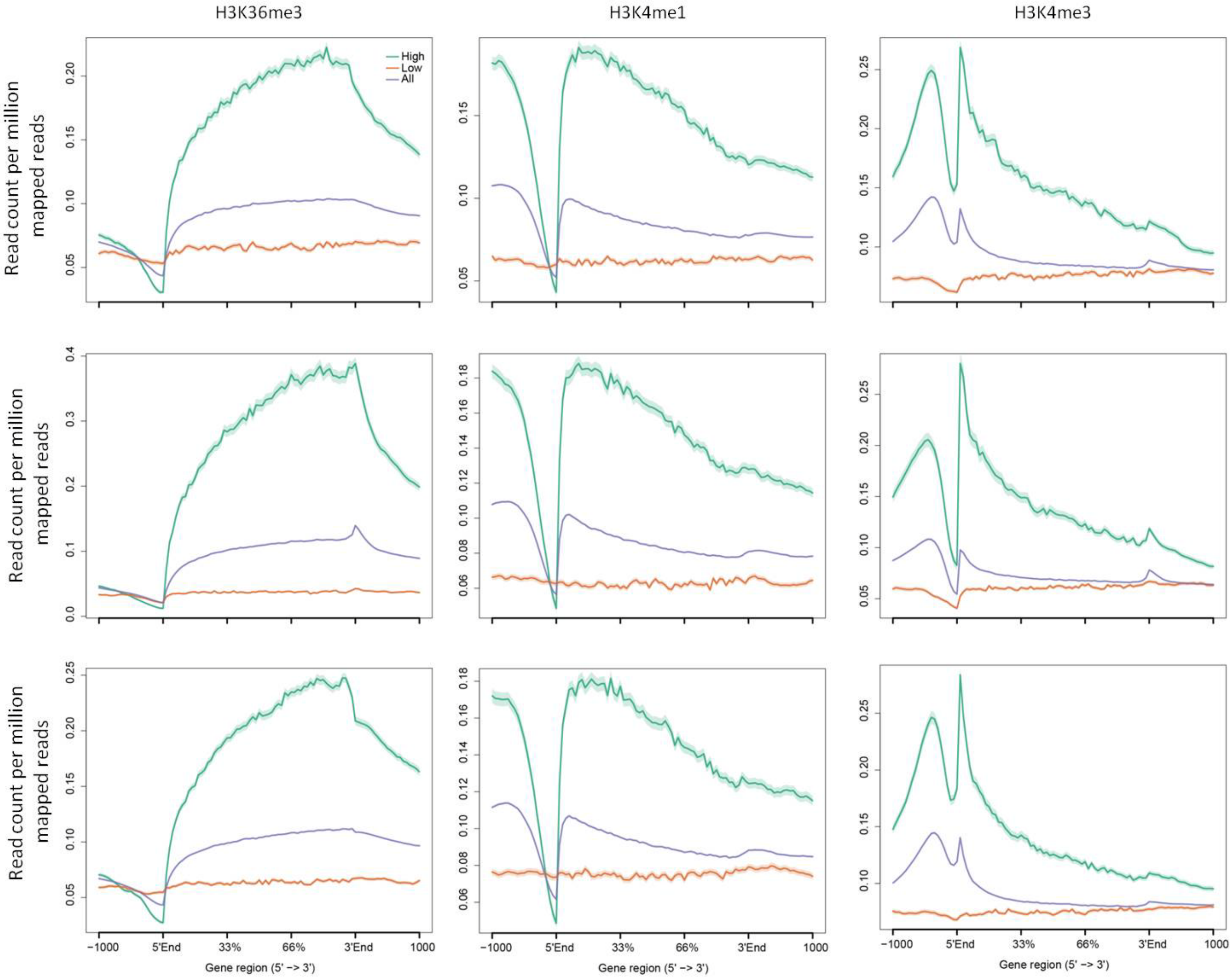


Figure 0.3: Composite plots of patterns of H3K36me3, H3K4me1 and H3K4me3 ChIP-seq signals across genic regions, in T cells.

H3K36me3, H3K4me1 and H3K4me3 are plotted across genic regions (*x* axis), of either the 1000 highly expressed genes (green), the 1000 lowly expressed genes (orange) or all genes (purple), in three T cell samples. Data represented as read count per million mapped reads (*y* axis).

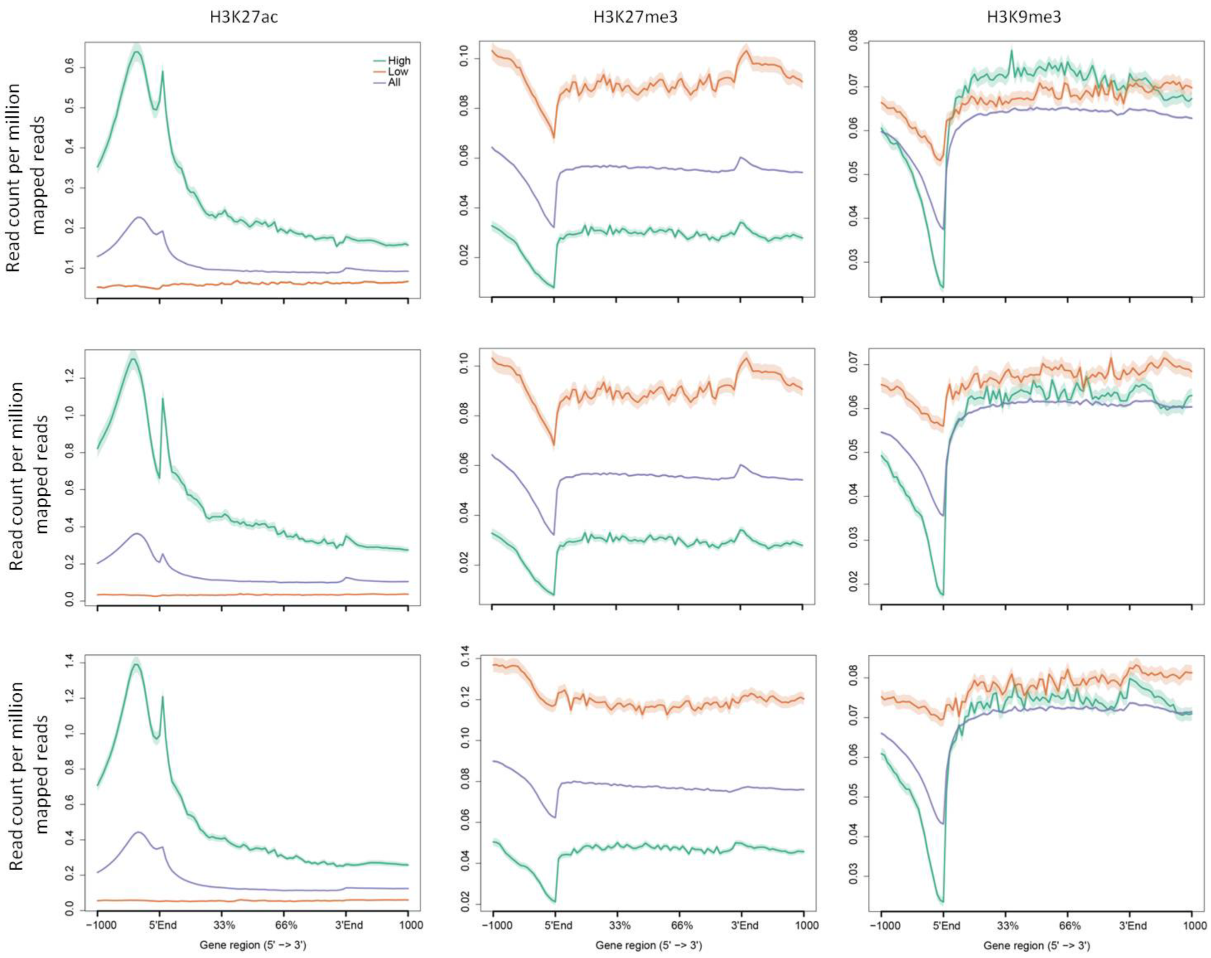


Figure 0.4: Composite plots of patterns of H3K27ac, H3K27me3 and H3K9me3 ChIP-seq signals across genic regions, in T cells.

H3K27ac, H3K27me3 and H3K9me3 are plotted across genic regions (*x* axis), of either the 1000 highly expressed genes (green), the 1000 lowly expressed genes (orange) or all genes (purple), in three T cell samples. Data represented as read count per million mapped reads (*y* axis).

Supplementary tables

Table 0.1: Detailed alignment statistics of RNA high-throughput sequencing samples

cell type[1]	sample_name[2]	raw_reads[3]	filtered_reads[4]	rRNA_reads[5]	aligned_reads[6]	%_aligned_reads[7]	%_duplicates[8]	mitochondrial_reads[9]	nmb_genes_detected (>=5 reads)[10]	intragenic_rate[11]	exonic_rate[12]	intronic_rate[13]	strand_specificity[14]
Monocyte	BF776_Mono_RNASeq_2	774,753,394	743,612,506	667,667,960	637,974,290	95.6	38.9	1,967,457	17,567	0.85	0.46	0.4	0.998
Monocyte	BF776_Mono_RNASeq_1	791,503,468	762,250,892	699,738,042	669,641,195	95.7	35	1,907,999	17,604	0.85	0.47	0.38	0.998
Monocyte	BF775_Mono_RNASeq_1	171,234,104	160,730,776	100,450,952	94,695,973	94.3	24.9	366,742	15,236	0.86	0.51	0.36	0.998
Monocyte	BF773_Mono_RNASeq_2	261,228,092	249,232,848	241,706,382	230,451,394	95.3	20.4	593,905	16,174	0.88	0.41	0.47	0.998
Monocyte	BF773_Mono_RNASeq_1	208,715,154	201,300,720	188,499,740	180,728,268	95.9	23.6	409,738	16,160	0.88	0.51	0.37	0.998
Monocyte	BF772_Mono_RNASeq_2	198,032,124	189,791,952	185,916,120	178,450,157	96	17.6	414,436	16,067	0.89	0.43	0.45	0.998
Monocyte	BF772_Mono_RNASeq_1	200,403,138	194,006,640	190,926,548	183,841,054	96.3	21.1	504,207	16,176	0.86	0.52	0.35	0.998
Monocyte	BF771_Mono_RNASeq_1	195,736,402	180,184,700	60,587,862	55,039,676	90.8	25	369,649	14,676	0.85	0.47	0.38	0.998
Monocyte	BF770_Mono_RNASeq_2	290,420,648	282,279,698	237,046,814	228,599,577	96.4	24.7	1,790,288	15,933	0.88	0.47	0.41	0.998
Monocyte	BF770_Mono_RNASeq_1	226,268,378	219,750,122	210,513,364	202,961,402	96.4	17.8	829,470	16,102	0.88	0.4	0.48	0.998
Monocyte	BF764_Mono_RNASeq_2	304,894,004	293,399,160	276,663,982	265,657,737	96	22.4	1,013,713	16,787	0.87	0.48	0.38	0.997
Monocyte	BF764_Mono_RNASeq_1	233,895,692	222,455,430	201,635,612	191,960,909	95.2	20.8	1,005,148	16,265	0.88	0.44	0.44	0.998
Monocyte	BF761_Mono_RNASeq_1	108,017,850	103,875,062	99,720,708	95,960,865	96.2	22.2	220,379	15,386	0.86	0.5	0.37	0.998
Monocyte	BF758_Mono_RNASeq_1	118,316,194	114,311,426	95,587,170	91,661,870	95.9	27.1	580,948	15,468	0.86	0.49	0.37	0.996
Monocyte	BF757_Mono_RNASeq_2	223,041,314	213,731,352	194,459,780	185,614,688	95.5	25.3	892,681	16,370	0.87	0.5	0.37	0.998
Monocyte	BF757_Mono_RNASeq_1	230,324,518	221,121,430	201,901,618	192,436,612	95.3	28.2	760,504	16,349	0.86	0.53	0.33	0.996

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]number of rRNA reads, [6]number of aligned reads,[7]rate of aligned reads, [8]rate of duplicated reads, [9]number of mitochondrial reads, [10]number of genes with ≥5 read counts, [11]rate of intragenic reads, [12]rate of exonic reads, [13]rate of intronic reads and [14]percentage of strand specificity

cell type[1]	sample_name[2]	raw_reads[3]	filtered_reads[4]	rRNA_reads[5]	aligned_reads[6]	%_aligned_reads[7]	%_duplicates[8]	mitochondrial_reads[9]	nmb_genes_detected (>=5 reads)[10]	intragenic_rate[11]	exonic_rate[12]	intronic_rate[13]	strand_specificity[14]
Monocyte	BCU899_Mono_RNASeq_1	89,664,712	85,423,656	84,355,554	80,526,339	95.5	23.7	318,611	15,349	0.87	0.51	0.36	0.998
Monocyte	BCU801_Mono_RNASeq_1	88,943,378	85,307,912	84,345,446	77,809,378	92.3	25.2	146,322	15,124	0.84	0.49	0.35	0.996
Monocyte	BCU768_Mono_RNASeq_1	108,684,308	102,525,804	100,380,900	95,637,391	95.3	21.2	318,369	15,640	0.86	0.49	0.37	0.998
Monocyte	BCU607_Mono_RNASeq_1	80,676,708	76,118,192	75,185,240	71,619,725	95.3	24.3	172,020	15,181	0.87	0.52	0.34	0.998
Monocyte	BCU582_Mono_RNASeq_1	98,177,890	92,390,130	87,356,546	82,948,593	95	25.2	171,727	15,405	0.84	0.51	0.33	0.998
Monocyte	BCU566_Mono_RNASeq_1	102,583,614	99,010,212	97,598,060	92,397,514	94.7	25.9	276,693	15,426	0.87	0.52	0.35	0.998
Monocyte	BCU551_Mono_RNASeq_1	173,338,252	161,267,500	135,758,214	130,048,221	95.8	24.9	268,512	15,742	0.84	0.48	0.35	0.997
Monocyte	BCU292_Mono_RNASeq_1	78,455,516	74,337,314	72,880,540	69,499,618	95.4	21.8	165,890	15,246	0.85	0.5	0.35	0.998
Monocyte	BCU1945_Mono_RNASeq_1	112,072,642	108,911,308	106,586,954	101,737,506	95.5	23.1	358,988	15,514	0.87	0.51	0.36	0.998
Monocyte	BCU1900_Mono_RNASeq_1	77,554,424	73,093,604	70,426,862	66,861,437	94.9	27.4	111,711	15,081	0.86	0.56	0.3	0.998
Monocyte	BCU1799_Mono_RNASeq_1	82,764,552	78,544,652	77,086,156	73,640,229	95.5	18.7	149,464	14,985	0.86	0.47	0.39	0.996
Monocyte	BCU1787_Mono_RNASeq_1	89,518,120	85,088,856	84,019,966	80,215,632	95.5	22	268,299	15,405	0.87	0.49	0.37	0.998
Monocyte	BCU1744_Mono_RNASeq_1	65,277,718	62,404,974	61,885,068	59,772,676	96.6	20.1	130,150	14,785	0.86	0.51	0.35	0.996
Monocyte	BCU173_Mono_RNASeq_1	81,108,558	77,416,838	76,231,006	72,874,628	95.6	21.7	210,530	15,155	0.87	0.49	0.38	0.998
Monocyte	BCU1731_Mono_RNASeq_1	133,541,040	128,126,856	123,241,294	116,717,031	94.7	20.2	353,295	15,744	0.89	0.5	0.38	0.998
Monocyte	BCU1657_Mono_RNASeq_1	76,562,180	71,373,372	70,206,340	66,650,686	94.9	21.6	115,444	14,912	0.84	0.46	0.37	0.996
Monocyte	BCU1595_Mono_RNASeq_1	80,103,198	77,500,030	76,131,242	73,135,060	96.1	21.1	137,243	15,162	0.83	0.45	0.38	0.996
Monocyte	BCU1571_Mono_RNASeq_1	100,901,532	96,620,706	95,314,272	87,888,086	92.2	24	175,700	15,306	0.85	0.5	0.35	0.996
Monocyte	BCU133_Mono_RNASeq_1	143,108,442	137,572,414	113,521,848	108,867,882	95.9	23.3	376,772	15,593	0.88	0.55	0.33	0.998

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]number of rRNA reads, [6]number of aligned reads,[7]rate of aligned reads, [8]rate of duplicated reads, [9]number of mitochondrial reads, [10]number of genes with ≥5 read counts, [11]rate of intragenic reads, [12]rate of exonic reads, [13]rate of intronic reads and [14]percentage of strand specificity

cell type[1]	sample_name[2]	raw_reads[3]	filtered_reads[4]	rRNA_reads[5]	aligned_reads[6]	%_aligned_reads[7]	%_duplicates[8]	mitochondrial_reads[9]	nmb_genes_detected (>=5 reads)[10]	intragenic_rate[11]	exonic_rate[12]	intronic_rate[13]	strand_specificity[14]
Monocyte	BCU120_Mono_RNASeq_1	77,616,296	73,274,080	72,209,288	68,717,231	95.2	25.2	136,819	15,091	0.84	0.54	0.3	0.998
Monocyte	BCU1053_Mono_RNASeq_1	90,407,318	84,523,310	82,674,442	78,406,440	94.8	23.4	153,062	15,355	0.84	0.5	0.34	0.998
Muscle	TB_Muscle_RNASeq_1	80,556,010	76,058,870	75,360,624	71,036,545	94.3	30.4	217,551	16,359	0.9	0.63	0.28	0.999
Muscle	RC_Muscle_RNASeq_1	132,109,246	124,621,584	123,397,930	116,189,526	94.2	23.1	622,972	17,877	0.86	0.48	0.38	0.998
Muscle	RA_Muscle_RNASeq_1	114,251,246	112,028,278	78,518,474	76,675,946	97.7	26	524,462	17,343	0.86	0.51	0.35	0.998
Muscle	MA_Muscle_RNASeq_1	107,968,514	101,353,606	93,692,796	89,473,909	95.5	17.8	1,279,905	17,744	0.84	0.42	0.41	0.998
Muscle	CF_Muscle_RNASeq_1	125,946,926	122,709,252	114,621,800	111,649,555	97.4	22.4	517,897	17,304	0.87	0.46	0.41	0.998
Muscle	BrW_Muscle_RNASeq_1	86,824,484	82,814,784	71,100,838	67,877,889	95.5	20.5	761,262	16,898	0.87	0.51	0.36	0.997
Muscle	BeW_Muscle_RNASeq_1	82,007,234	76,550,840	75,582,724	72,462,464	95.9	31.3	823,966	17,170	0.88	0.57	0.31	0.998
Muscle	AM_Muscle_RNASeq_1	119,717,310	114,724,372	90,041,404	85,947,570	95.5	24.3	737,653	17,119	0.87	0.52	0.35	0.997
Muscle	AD_Muscle_RNASeq_1	149,658,112	139,989,100	135,578,082	129,839,664	95.8	28	1,072,753	17,954	0.89	0.57	0.32	0.998
T Cells	BF805_TC_RNASeq_1	140,130,756	135,945,730	106,548,980	102,743,341	96.4	22.4	729,205	16,071	0.87	0.49	0.38	0.996
T Cells	BF776_TC_RNASeq_2	887,893,658	824,883,940	744,881,276	705,286,381	94.7	48.8	1,166,332	18,246	0.85	0.41	0.44	0.997
T Cells	BF776_TC_RNASeq_1	917,777,156	855,746,826	738,304,336	699,889,895	94.8	36.3	1,423,223	18,237	0.83	0.43	0.4	0.997
T Cells	BF775_TC_RNASeq_1	158,274,938	151,361,470	126,729,782	121,457,393	95.8	19.4	199,920	16,352	0.86	0.41	0.45	0.997
T Cells	BF773_TC_RNASeq_2	144,102,878	137,660,058	129,983,620	123,905,412	95.3	20.8	181,138	15,918	0.85	0.32	0.54	0.995
T Cells	BF772_TC_RNASeq_2	238,984,416	228,207,622	215,972,030	205,160,755	95	15.7	386,203	16,877	0.86	0.29	0.57	0.996
T Cells	BF772_TC_RNASeq_1	197,485,334	189,555,974	159,902,518	151,995,096	95.1	18.3	341,459	16,745	0.85	0.41	0.44	0.997
T Cells	BF770_TC_RNASeq_1	132,424,054	128,175,266	119,201,286	115,043,230	96.5	22.7	360,396	16,152	0.85	0.44	0.41	0.997

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]number of rRNA reads, [6]number of aligned reads,[7]rate of aligned reads, [8]rate of duplicated reads, [9]number of mitochondrial reads, [10]number of genes with ≥5 read counts, [11]rate of intragenic reads, [12]rate of exonic reads, [13]rate of intronic reads and [14]percentage of strand specificity

cell type[1]	sample_name[2]	raw_reads[3]	filtered_reads[4]	rRNA_reads[5]	aligned_reads[6]	%_aligned_reads[7]	%_duplicates[8]	mitochondrial_reads[9]	nmb_genes_detected (>=5 reads)[10]	intragenic_rate[11]	exonic_rate[12]	intronic_rate[13]	strand_specificity[14]
T Cells	BF764_TC_RNASeq_2	340,015,710	325,233,378	232,435,704	222,437,975	95.7	21.9	872,044	17,274	0.83	0.43	0.39	0.996
T Cells	BF764_TC_RNASeq_1	293,211,416	276,072,028	213,084,976	202,116,243	94.9	25.7	999,323	16,921	0.84	0.44	0.4	0.996
T Cells	BF761_TC_RNASeq_1	108,428,714	103,948,152	94,262,712	90,471,690	96	23.1	165,206	16,017	0.85	0.44	0.4	0.997
T Cells	BF758_TC_RNASeq_2	295,300,776	280,347,324	227,071,152	216,013,648	95.1	26.8	1,206,008	16,921	0.84	0.42	0.42	0.997
T Cells	BF758_TC_RNASeq_1	240,704,102	227,941,634	197,851,824	187,520,980	94.8	22.7	778,499	16,950	0.84	0.42	0.42	0.997
T Cells	BF757_TC_RNASeq_2	233,917,314	221,817,606	203,161,008	192,969,762	95	20.9	666,520	16,926	0.85	0.41	0.43	0.997
T Cells	BF757_TC_RNASeq_1	223,528,978	212,563,252	189,717,634	180,039,848	94.9	22	673,805	17,008	0.85	0.42	0.43	0.997
T Cells	BF705_TC_RNASeq_1	125,590,442	122,349,394	91,657,760	88,471,385	96.5	41.1	694,436	15,488	0.82	0.63	0.19	0.998
T Cells	BCU899_TC_RNASeq_1	105,056,662	99,893,832	97,235,956	92,660,702	95.3	21.5	255,602	16,177	0.84	0.41	0.43	0.997
T Cells	BCU801_TC_RNASeq_1	105,677,062	99,968,466	97,725,886	89,760,206	91.8	18	90,495	15,979	0.84	0.34	0.5	0.994
T Cells	BCU768_TC_RNASeq_1	113,223,594	107,010,626	105,454,818	100,405,536	95.2	20.8	222,301	16,254	0.84	0.41	0.44	0.996
T Cells	BCU607_TC_RNASeq_1	95,596,612	89,044,428	78,422,160	74,186,966	94.6	22.2	135,563	15,932	0.84	0.41	0.43	0.997
T Cells	BCU582_TC_RNASeq_1	79,750,894	75,092,736	73,668,584	70,020,595	95	23.4	92,027	15,793	0.85	0.42	0.42	0.997
T Cells	BCU566_TC_RNASeq_1	113,797,208	110,688,466	106,036,822	101,059,076	95.3	22.4	222,033	16,156	0.86	0.42	0.43	0.997
T Cells	BCU551_TC_RNASeq_1	71,642,128	68,305,082	60,571,036	58,122,650	96	17.8	69,707	15,475	0.83	0.37	0.46	0.994
T Cells	BCU292_TC_RNASeq_1	81,683,444	76,923,366	75,839,290	72,147,146	95.1	22	122,217	15,916	0.84	0.44	0.4	0.997
T Cells	BCU1945_TC_RNASeq_1	103,994,004	101,411,178	100,066,810	95,368,087	95.3	21.8	175,662	16,233	0.85	0.39	0.46	0.996
T Cells	BCU1900_TC_RNASeq_1	80,285,330	76,161,684	74,730,320	71,040,852	95.1	23.7	90,102	15,805	0.85	0.41	0.44	0.997
T Cells	BCU1799_TC_RNASeq_1	95,155,232	91,528,910	89,803,766	86,206,755	96	23.8	124,517	15,955	0.81	0.41	0.4	0.995

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]number of rRNA reads, [6]number of aligned reads,[7]rate of aligned reads, [8]rate of duplicated reads, [9]number of mitochondrial reads, [10]number of genes with ≥5 read counts, [11]rate of intragenic reads, [12]rate of exonic reads, [13]rate of intronic reads and [14]percentage of strand specificity

cell type[1]	sample_name[2]	raw_reads[3]	filtered_reads[4]	rRNA_reads[5]	aligned_reads[6]	%_aligned_reads[7]	%_duplicates[8]	mitochondrial_reads[9]	nmb_genes_detected (>=5 reads)[10]	intragenic_rate[11]	exonic_rate[12]	intronic_rate[13]	strand_specificity[14]
T Cells	BCU1787_TC_RNASeq_1	87,535,592	83,320,994	82,079,812	78,209,913	95.3	19.1	140,100	15,986	0.85	0.39	0.46	0.997
T Cells	BCU1744_TC_RNASeq_1	62,575,846	60,126,444	59,560,210	57,447,941	96.5	18.9	98,159	16,087	0.81	0.39	0.43	0.995
T Cells	BCU173_TC_RNASeq_1	94,414,222	88,555,094	87,154,450	82,763,560	95	21.3	171,802	16,007	0.85	0.43	0.42	0.997
T Cells	BCU1731_TC_RNASeq_1	124,870,162	119,548,024	115,270,620	108,944,016	94.5	18	184,324	16,340	0.86	0.39	0.47	0.996
T Cells	BCU1657_TC_RNASeq_1	73,727,184	71,504,554	70,413,676	67,708,496	96.2	19.6	100,456	15,644	0.82	0.39	0.43	0.995
T Cells	BCU1595_TC_RNASeq_1	73,814,546	69,887,056	69,389,062	66,195,813	95.4	18.2	92,525	15,670	0.84	0.39	0.45	0.995
T Cells	BCU1571_TC_RNASeq_1	88,205,236	84,175,112	82,810,128	79,023,888	95.4	18.1	85,145	15,895	0.84	0.35	0.49	0.994
T Cells	BCU133_TC_RNASeq_1	119,780,920	114,859,534	103,764,932	99,438,195	95.8	26.6	202,414	16,008	0.85	0.45	0.4	0.997
T Cells	BCU120_TC_RNASeq_1	97,255,134	90,765,416	87,047,200	82,319,779	94.6	25.1	113,243	16,091	0.81	0.42	0.4	0.997
T Cells	BCU1053_TC_RNASeq_1	92,702,814	87,361,056	85,688,092	81,567,635	95.2	22.2	126,259	16,074	0.83	0.44	0.39	0.997

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]number of rRNA reads, [6]number of aligned reads,[7]rate of aligned reads, [8]rate of duplicated reads, [9]number of mitochondrial reads, [10]number of genes with ≥5 read counts, [11]rate of intragenic reads, [12]rate of exonic reads, [13]rate of intronic reads and [14]percentage of strand specificity

Table 0.2: Detailed alignment statistics of ChIP-Seq samples

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
T Cells	BF776_TC_ChIP_H3K4me3_2	BF776_TC_ChIP_Input_2	85,997,076	84,697,900	88,401,972	87,447,750	92.6	95.2	35.3	21.8
T Cells	BF776_TC_ChIP_H3K4me1_2	BF776_TC_ChIP_Input_2	88,715,486	87,766,122	88,401,972	87,447,750	96.3	95.2	22.7	21.8
T Cells	BF776_TC_ChIP_H3K27ac_2	BF776_TC_ChIP_Input_2	84,001,186	81,819,688	88,401,972	87,447,750	97.7	95.2	25.6	21.8
Monocyte	BF776_Mono_ChIP_H3K4me3_2	BF776_Mono_ChIP_Input_2	41,291,840	38,421,440	77,888,856	76,931,046	84.3	95.6	51	2.1
Monocyte	BF776_Mono_ChIP_H3K4me1_2	BF776_Mono_ChIP_Input_2	73,605,210	72,629,660	77,888,856	76,931,046	97.6	95.6	2.5	2.1
Monocyte	BF776_Mono_ChIP_H3K27ac_2	BF776_Mono_ChIP_Input_2	73,698,330	72,003,094	77,888,856	76,931,046	98.1	95.6	5.6	2.1
T Cells	BF775_TC_ChIP_H3K4me3_1	BF775_TC_ChIP_Input_1	68,161,028	62,797,378	92,143,016	58,641,470	90.8	94.9	12.8	3.1
T Cells	BF775_TC_ChIP_H3K36me3_1	BF775_TC_ChIP_Input_1	87,627,106	85,409,634	92,143,016	58,641,470	96.9	94.9	4.3	3.1
T Cells	BF775_TC_ChIP_H3K27me3_1	BF775_TC_ChIP_Input_1	67,577,824	64,187,470	92,143,016	58,641,470	97.3	94.9	12.5	3.1
Monocyte	BF775_Mono_ChIP_H3K4me3_1	BF775_Mono_ChIP_Input_1	79,989,924	73,975,952	96,981,564	82,717,804	85.1	96.1	14.9	1.9
Monocyte	BF775_Mono_ChIP_H3K36me3_1	BF775_Mono_ChIP_Input_1	86,840,224	85,079,254	96,981,564	82,717,804	96.5	96.1	4.5	1.9
Monocyte	BF775_Mono_ChIP_H3K27me3_1	BF775_Mono_ChIP_Input_1	59,906,888	55,068,022	96,981,564	82,717,804	93	96.1	50.8	1.9
Monocyte	BF773_Mono_ChIP_H3K9me3_2	BF773_Mono_ChIP_Input_2	107,042,608	104,648,076	102,780,848	100,338,496	89.6	96.5	20	2.3
Monocyte	BF773_Mono_ChIP_H3K4me3_2	BF773_Mono_ChIP_Input_2	84,807,594	82,558,886	102,780,848	100,338,496	96.5	96.5	19.6	2.3
Monocyte	BF773_Mono_ChIP_H3K4me1_2	BF773_Mono_ChIP_Input_2	254,654,416	248,830,602	102,780,848	100,338,496	97.5	96.5	15.3	2.3
Monocyte	BF773_Mono_ChIP_H3K36me3_2	BF773_Mono_ChIP_Input_2	90,636,568	81,503,324	102,780,848	100,338,496	96.1	96.5	40.2	2.3
Monocyte	BF773_Mono_ChIP_H3K27me3_2	BF773_Mono_ChIP_Input_2	77,605,776	73,591,802	102,780,848	100,338,496	97.6	96.5	23.7	2.3
Monocyte	BF773_Mono_ChIP_H3K27ac_2	BF773_Mono_ChIP_Input_2	83,249,222	79,385,956	102,780,848	100,338,496	97.7	96.5	55.5	2.3
Monocyte	BF772_Mono_ChIP_H3K9me3_2	BF772_Mono_ChIP_Input_2	96,446,164	94,268,754	117,734,680	114,272,924	91.5	95.8	7.3	71.4
Monocyte	BF772_Mono_ChIP_H3K4me3_2	BF772_Mono_ChIP_Input_2	102,139,436	98,437,562	117,734,680	114,272,924	96	95.8	11.4	71.4
Monocyte	BF772_Mono_ChIP_H3K4me1_2	BF772_Mono_ChIP_Input_2	83,995,994	80,045,008	117,734,680	114,272,924	97	95.8	92.1	71.4
Monocyte	BF772_Mono_ChIP_H3K36me3_2	BF772_Mono_ChIP_Input_2	104,284,660	100,804,604	117,734,680	114,272,924	97.4	95.8	10.8	71.4
Monocyte	BF772_Mono_ChIP_H3K27me3_2	BF772_Mono_ChIP_Input_2	105,106,116	102,245,692	117,734,680	114,272,924	97	95.8	15.2	71.4

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
T Cells	BF770_TC_ChIP_H3K4me3_1	BF770_TC_ChIP_Input_1	245,913,320	240,082,310	162,240,698	146,968,918	95.3	95.4	11.5	3.2
T Cells	BF770_TC_ChIP_H3K36me3_1	BF770_TC_ChIP_Input_1	229,271,342	224,726,864	162,240,698	146,968,918	96.9	95.4	19.9	3.2
T Cells	BF770_TC_ChIP_H3K27me3_1	BF770_TC_ChIP_Input_1	133,077,268	130,566,820	162,240,698	146,968,918	97.1	95.4	19.7	3.2
T Cells	BF764_TC_ChIP_H3K4me3_2	BF764_TC_ChIP_Input_2	121,976,446	120,368,328	117,975,254	113,874,424	93.8	95.7	9.4	4.6
T Cells	BF764_TC_ChIP_H3K4me3_1	BF764_TC_ChIP_Input_1	82,454,126	80,414,112	78,469,500	71,197,550	93.1	96	7.7	1.9
T Cells	BF764_TC_ChIP_H3K4me1_2	BF764_TC_ChIP_Input_2	131,331,896	128,961,694	117,975,254	113,874,424	96.6	95.7	4	4.6
T Cells	BF764_TC_ChIP_H3K36me3_2	BF764_TC_ChIP_Input_2	106,254,944	103,826,864	117,975,254	113,874,424	96.6	95.7	8.9	4.6
T Cells	BF764_TC_ChIP_H3K36me3_1	BF764_TC_ChIP_Input_1	84,826,152	83,114,818	78,469,500	71,197,550	97.3	96	5.1	1.9
T Cells	BF764_TC_ChIP_H3K27me3_2	BF764_TC_ChIP_Input_2	136,358,384	133,460,250	117,975,254	113,874,424	97.4	95.7	8.2	4.6
T Cells	BF764_TC_ChIP_H3K27me3_1	BF764_TC_ChIP_Input_1	73,238,886	71,854,108	78,469,500	71,197,550	97.6	96	19.5	1.9
Monocyte	BF764_Mono_ChIP_H3K4me3_1	BF764_Mono_ChIP_Input_1	76,771,150	65,038,626	91,343,384	81,087,966	85.3	96.1	20	2
Monocyte	BF764_Mono_ChIP_H3K36me3_1	BF764_Mono_ChIP_Input_1	89,890,768	82,754,924	91,343,384	81,087,966	96.8	96.1	12.3	2
Monocyte	BF764_Mono_ChIP_H3K27me3_1	BF764_Mono_ChIP_Input_1	67,248,950	36,492,808	91,343,384	81,087,966	60.2	96.1	46.2	2
T Cells	BF761_TC_ChIP_H3K4me3_1	BF761_TC_ChIP_Input_1	95,582,672	92,428,020	50,431,816	47,320,668	93.9	95.9	2.6	6.6
T Cells	BF761_TC_ChIP_H3K36me3_1	BF761_TC_ChIP_Input_1	92,436,146	90,423,792	50,431,816	47,320,668	97.8	95.9	1.6	6.6
T Cells	BF761_TC_ChIP_H3K27me3_1	BF761_TC_ChIP_Input_1	94,105,626	92,104,482	50,431,816	47,320,668	98.6	95.9	1.9	6.6
Monocyte	BF761_Mono_ChIP_H3K4me3_1	BF761_Mono_ChIP_Input_1	104,532,072	82,741,298	90,237,136	64,723,970	64.6	95.8	48.1	3.3
Monocyte	BF761_Mono_ChIP_H3K36me3_1	BF761_Mono_ChIP_Input_1	92,691,174	70,466,362	90,237,136	64,723,970	74.8	95.8	55.1	3.3
Monocyte	BF761_Mono_ChIP_H3K27me3_1	BF761_Mono_ChIP_Input_1	87,159,890	69,049,816	90,237,136	64,723,970	95.9	95.8	56.6	3.3
T Cells	BF757_TC_ChIP_H3K4me3_2	BF757_TC_ChIP_Input_2	100,440,768	97,479,744	108,591,830	105,686,412	93.2	95.8	11.4	3.7
T Cells	BF757_TC_ChIP_H3K4me1_2	BF757_TC_ChIP_Input_2	98,795,516	95,951,590	108,591,830	105,686,412	97.1	95.8	2.8	3.7
T Cells	BF757_TC_ChIP_H3K27ac_2	BF757_TC_ChIP_Input_2	99,482,766	97,161,552	108,591,830	105,686,412	98	95.8	3.5	3.7
Monocyte	BF757_Mono_ChIP_H3K4me3_2	BF757_Mono_ChIP_Input_2	27,609,396	22,384,312	95,303,484	93,907,650	82.6	96.2	44.7	2.7
Monocyte	BF757_Mono_ChIP_H3K4me1_2	BF757_Mono_ChIP_Input_2	89,029,866	86,962,070	95,303,484	93,907,650	98	96.2	8.2	2.7

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
Monocyte	BF757_Mono_ChIP_H3K27ac_2	BF757_Mono_ChIP_Input_2	90,276,870	88,053,750	95,303,484	93,907,650	97.6	96.2	14.5	2.7
T Cells	BCU899_TC_ChIP_H3K9me3_1	BCU899_TC_ChIP_Input_1	108,710,322	106,683,126	185,450,298	178,743,740	95	96	7.7	75.6
T Cells	BCU899_TC_ChIP_H3K4me3_1	BCU899_TC_ChIP_Input_1	85,383,972	83,779,152	185,450,298	178,743,740	96.1	96	21.6	75.6
T Cells	BCU899_TC_ChIP_H3K4me1_1	BCU899_TC_ChIP_Input_1	100,096,432	97,829,390	185,450,298	178,743,740	96.6	96	21	75.6
T Cells	BCU899_TC_ChIP_H3K36me3_1	BCU899_TC_ChIP_Input_1	85,302,386	79,765,678	185,450,298	178,743,740	96	96	39.6	75.6
T Cells	BCU899_TC_ChIP_H3K27me3_1	BCU899_TC_ChIP_Input_1	101,433,402	99,652,870	185,450,298	178,743,740	97.1	96	8.5	75.6
T Cells	BCU899_TC_ChIP_H3K27ac_1	BCU899_TC_ChIP_Input_1	102,179,064	93,501,526	185,450,298	178,743,740	95.4	96	79.1	75.6
T Cells	BCU801_TC_ChIP_H3K9me3_1	BCU801_TC_ChIP_Input_1	83,665,870	80,802,662	109,206,430	104,712,880	95	96.9	2.2	1.4
T Cells	BCU801_TC_ChIP_H3K4me3_1	BCU801_TC_ChIP_Input_1	81,383,716	78,553,296	109,206,430	104,712,880	96.8	96.9	2.1	1.4
T Cells	BCU801_TC_ChIP_H3K4me1_1	BCU801_TC_ChIP_Input_1	196,726,626	192,106,606	109,206,430	104,712,880	97.4	96.9	2.2	1.4
T Cells	BCU801_TC_ChIP_H3K36me3_1	BCU801_TC_ChIP_Input_1	74,055,938	70,330,496	109,206,430	104,712,880	97.4	96.9	5.8	1.4
T Cells	BCU801_TC_ChIP_H3K27me3_1	BCU801_TC_ChIP_Input_1	76,011,376	71,246,338	109,206,430	104,712,880	97.3	96.9	7.4	1.4
T Cells	BCU801_TC_ChIP_H3K27ac_1	BCU801_TC_ChIP_Input_1	79,893,792	75,324,832	109,206,430	104,712,880	95.7	96.9	21.8	1.4
T Cells	BCU768_TC_ChIP_H3K9me3_1	BCU768_TC_ChIP_Input_1	82,341,522	77,818,252	126,037,706	119,911,288	94.4	96.2	5.8	1.6
T Cells	BCU768_TC_ChIP_H3K4me3_1	BCU768_TC_ChIP_Input_1	124,963,104	122,561,272	126,037,706	119,911,288	96.1	96.2	9.3	1.6
T Cells	BCU768_TC_ChIP_H3K4me1_1	BCU768_TC_ChIP_Input_1	83,808,014	81,675,186	126,037,706	119,911,288	97	96.2	2	1.6
T Cells	BCU768_TC_ChIP_H3K36me3_1	BCU768_TC_ChIP_Input_1	123,416,984	110,237,228	126,037,706	119,911,288	96.4	96.2	29.6	1.6
T Cells	BCU768_TC_ChIP_H3K27me3_1	BCU768_TC_ChIP_Input_1	54,948,578	53,433,150	126,037,706	119,911,288	97.6	96.2	13.7	1.6
T Cells	BCU768_TC_ChIP_H3K27ac_1	BCU768_TC_ChIP_Input_1	77,938,600	73,249,064	126,037,706	119,911,288	95.8	96.2	21	1.6
T Cells	BCU607_TC_ChIP_H3K9me3_1	BCU607_TC_ChIP_Input_1	96,817,828	93,827,678	100,468,200	95,564,924	95	96.4	2.3	1.7
T Cells	BCU607_TC_ChIP_H3K4me3_1	BCU607_TC_ChIP_Input_1	85,646,856	83,284,128	100,468,200	95,564,924	96.5	96.4	1.9	1.7
T Cells	BCU607_TC_ChIP_H3K4me1_1	BCU607_TC_ChIP_Input_1	104,306,360	100,132,736	100,468,200	95,564,924	96.5	96.4	3.6	1.7
T Cells	BCU607_TC_ChIP_H3K36me3_1	BCU607_TC_ChIP_Input_1	104,980,796	100,268,448	100,468,200	95,564,924	96.6	96.4	2.9	1.7
T Cells	BCU607_TC_ChIP_H3K27me3_1	BCU607_TC_ChIP_Input_1	97,723,686	95,801,402	100,468,200	95,564,924	96.9	96.4	10.5	1.7

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
T Cells	BCU607_TC_ChIP_H3K27ac_1	BCU607_TC_ChIP_Input_1	106,088,890	102,569,978	100,468,200	95,564,924	96.7	96.4	45.1	1.7
T Cells	BCU582_TC_ChIP_H3K9me3_1	BCU582_TC_ChIP_Input_1	104,776,114	94,399,376	106,180,416	95,419,826	92.4	94.1	25.1	1.5
T Cells	BCU582_TC_ChIP_H3K4me3_1	BCU582_TC_ChIP_Input_1	84,162,836	75,027,868	106,180,416	95,419,826	94.7	94.1	10.1	1.5
T Cells	BCU582_TC_ChIP_H3K4me1_1	BCU582_TC_ChIP_Input_1	343,548,556	325,077,014	106,180,416	95,419,826	96.1	94.1	9.7	1.5
T Cells	BCU582_TC_ChIP_H3K36me3_1	BCU582_TC_ChIP_Input_1	95,220,012	87,478,914	106,180,416	95,419,826	96.4	94.1	2.3	1.5
T Cells	BCU582_TC_ChIP_H3K27me3_1	BCU582_TC_ChIP_Input_1	98,386,658	90,354,272	106,180,416	95,419,826	97	94.1	21	1.5
T Cells	BCU582_TC_ChIP_H3K27ac_1	BCU582_TC_ChIP_Input_1	90,191,170	81,538,934	106,180,416	95,419,826	96.3	94.1	15.4	1.5
Monocyte	BCU566_Mono_ChIP_H3K9me3_1	BCU566_Mono_ChIP_Input_1	238,005,740	225,213,748	139,413,336	135,363,466	82.4	96.2	38.4	6.9
Monocyte	BCU566_Mono_ChIP_H3K4me3_1	BCU566_Mono_ChIP_Input_1	170,973,420	160,707,846	139,413,336	135,363,466	95.4	96.2	63	6.9
Monocyte	BCU566_Mono_ChIP_H3K4me1_1	BCU566_Mono_ChIP_Input_1	230,305,188	197,621,006	139,413,336	135,363,466	95	96.2	51	6.9
Monocyte	BCU566_Mono_ChIP_H3K36me3_1	BCU566_Mono_ChIP_Input_1	220,174,640	209,403,864	139,413,336	135,363,466	95.2	96.2	38.2	6.9
Monocyte	BCU566_Mono_ChIP_H3K27me3_1	BCU566_Mono_ChIP_Input_1	165,476,352	157,672,968	139,413,336	135,363,466	98.5	96.2	24	6.9
Monocyte	BCU566_Mono_ChIP_H3K27ac_1	BCU566_Mono_ChIP_Input_1	119,936,570	112,044,772	139,413,336	135,363,466	98.3	96.2	18.5	6.9
T Cells	BCU551_TC_ChIP_H3K9me3_1_Rep	BCU551_TC_ChIP_Input_1	60,650,776	59,478,874	112,182,380	108,265,400	93.8	96.1	2.1	7.5
T Cells	BCU551_TC_ChIP_H3K9me3_1	BCU551_TC_ChIP_Input_1	97,251,240	92,308,854	112,182,380	108,265,400	93.8	96.1	17.3	7.5
T Cells	BCU551_TC_ChIP_H3K4me3_1_Rep	BCU551_TC_ChIP_Input_1	99,953,172	96,803,514	112,182,380	108,265,400	96.1	96.1	3.4	7.5
T Cells	BCU551_TC_ChIP_H3K4me3_1	BCU551_TC_ChIP_Input_1	111,486,788	106,619,656	112,182,380	108,265,400	96.1	96.1	30.9	7.5
T Cells	BCU551_TC_ChIP_H3K4me1_1_Rep	BCU551_TC_ChIP_Input_1	167,864,242	161,613,026	112,182,380	108,265,400	95.9	96.1	96.2	7.5
T Cells	BCU551_TC_ChIP_H3K4me1_1	BCU551_TC_ChIP_Input_1	94,568,946	91,427,962	112,182,380	108,265,400	96.6	96.1	53.4	7.5
T Cells	BCU551_TC_ChIP_H3K36me3_1_Rep	BCU551_TC_ChIP_Input_1	117,322,506	113,549,134	112,182,380	108,265,400	97.2	96.1	3.6	7.5
T Cells	BCU551_TC_ChIP_H3K36me3_1	BCU551_TC_ChIP_Input_1	107,237,216	103,726,636	112,182,380	108,265,400	97.4	96.1	10.2	7.5
T Cells	BCU551_TC_ChIP_H3K27me3_1_Rep	BCU551_TC_ChIP_Input_1	44,203,042	40,480,534	112,182,380	108,265,400	96	96.1	4.9	7.5
T Cells	BCU551_TC_ChIP_H3K27me3_1	BCU551_TC_ChIP_Input_1	91,725,664	88,788,218	112,182,380	108,265,400	97.4	96.1	54.5	7.5
T Cells	BCU551_TC_ChIP_H3K27ac_1_Rep	BCU551_TC_ChIP_Input_1	62,849,344	57,340,740	112,182,380	108,265,400	96.8	96.1	10.9	7.5

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
T Cells	BCU551_TC_ChIP_H3K27ac_1	BCU551_TC_ChIP_Input_1	89,171,734	86,267,776	112,182,380	108,265,400	97.9	96.1	68.8	7.5
T Cells	BCU292_TC_ChIP_H3K9me3_1	BCU292_TC_ChIP_Input_1	111,210,312	99,491,674	103,045,968	95,637,898	93.6	95.8	3.6	4.1
T Cells	BCU292_TC_ChIP_H3K4me3_1	BCU292_TC_ChIP_Input_1	108,582,818	99,419,952	103,045,968	95,637,898	96.1	95.8	7.6	4.1
T Cells	BCU292_TC_ChIP_H3K4me1_1	BCU292_TC_ChIP_Input_1	302,586,222	288,659,310	103,045,968	95,637,898	97.3	95.8	12.5	4.1
T Cells	BCU292_TC_ChIP_H3K36me3_1	BCU292_TC_ChIP_Input_1	100,977,754	91,021,538	103,045,968	95,637,898	96.5	95.8	10.6	4.1
T Cells	BCU292_TC_ChIP_H3K27me3_1	BCU292_TC_ChIP_Input_1	93,708,140	84,004,772	103,045,968	95,637,898	96.7	95.8	2.3	4.1
T Cells	BCU292_TC_ChIP_H3K27ac_1	BCU292_TC_ChIP_Input_1	108,409,628	96,157,052	103,045,968	95,637,898	96.5	95.8	12.6	4.1
T Cells	BCU1945_TC_ChIP_H3K9me3_1	BCU1945_TC_ChIP_Input_1	46,617,628	45,181,956	52,163,906	49,554,386	94	95.9	3.8	3.1
T Cells	BCU1945_TC_ChIP_H3K4me3_1	BCU1945_TC_ChIP_Input_1	63,304,468	61,881,802	52,163,906	49,554,386	96	95.9	3.1	3.1
T Cells	BCU1945_TC_ChIP_H3K4me1_1	BCU1945_TC_ChIP_Input_1	104,813,642	99,773,356	52,163,906	49,554,386	96.3	95.9	10.1	3.1
T Cells	BCU1945_TC_ChIP_H3K36me3_1	BCU1945_TC_ChIP_Input_1	56,937,732	54,494,208	52,163,906	49,554,386	96.9	95.9	3.1	3.1
T Cells	BCU1945_TC_ChIP_H3K27me3_1	BCU1945_TC_ChIP_Input_1	70,092,926	64,594,540	52,163,906	49,554,386	96.2	95.9	6.6	3.1
T Cells	BCU1945_TC_ChIP_H3K27ac_1	BCU1945_TC_ChIP_Input_1	87,549,478	80,679,054	52,163,906	49,554,386	97.1	95.9	14.9	3.1
T Cells	BCU1799_TC_ChIP_H3K9me3_1	BCU1799_TC_ChIP_Input_1	156,717,698	153,568,878	200,663,598	191,370,152	93.1	96.4	2.7	0.9
T Cells	BCU1799_TC_ChIP_H3K4me3_1	BCU1799_TC_ChIP_Input_1	73,932,674	71,756,482	200,663,598	191,370,152	95.2	96.4	2	0.9
T Cells	BCU1799_TC_ChIP_H3K4me1_1	BCU1799_TC_ChIP_Input_1	235,229,500	224,152,354	200,663,598	191,370,152	96.5	96.4	2	0.9
T Cells	BCU1799_TC_ChIP_H3K36me3_1	BCU1799_TC_ChIP_Input_1	181,377,816	169,490,224	200,663,598	191,370,152	96.8	96.4	14.5	0.9
T Cells	BCU1799_TC_ChIP_H3K27me3_1	BCU1799_TC_ChIP_Input_1	186,734,250	176,238,032	200,663,598	191,370,152	97	96.4	3.8	0.9
T Cells	BCU1799_TC_ChIP_H3K27ac_1	BCU1799_TC_ChIP_Input_1	79,132,188	67,903,128	200,663,598	191,370,152	93.9	96.4	6.9	0.9
T Cells	BCU1787_TC_ChIP_H3K9me3_1	BCU1787_TC_ChIP_Input_1	120,340,548	116,122,568	125,060,354	112,970,030	94.4	93.7	4.7	1.9
T Cells	BCU1787_TC_ChIP_H3K4me3_1	BCU1787_TC_ChIP_Input_1	129,563,068	126,004,602	125,060,354	112,970,030	96.2	93.7	4.6	1.9
T Cells	BCU1787_TC_ChIP_H3K4me1_1	BCU1787_TC_ChIP_Input_1	112,065,770	99,442,226	125,060,354	112,970,030	94.3	93.7	28.3	1.9
T Cells	BCU1787_TC_ChIP_H3K36me3_1	BCU1787_TC_ChIP_Input_1	66,681,824	64,941,124	125,060,354	112,970,030	96.8	93.7	2.3	1.9
T Cells	BCU1787_TC_ChIP_H3K27me3_1	BCU1787_TC_ChIP_Input_1	115,925,128	112,558,218	125,060,354	112,970,030	97.9	93.7	15.6	1.9

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
T Cells	BCU1787_TC_ChIP_H3K27ac_1	BCU1787_TC_ChIP_Input_1	48,442,614	44,908,508	125,060,354	112,970,030	96.2	93.7	78	1.9
T Cells	BCU173_TC_ChIP_H3K9me3_1	BCU173_TC_ChIP_Input_1	90,871,374	84,114,176	95,805,330	88,615,808	94.2	94.8	14.7	1.5
T Cells	BCU173_TC_ChIP_H3K4me3_1	BCU173_TC_ChIP_Input_1	100,643,600	92,412,272	95,805,330	88,615,808	95.3	94.8	17.6	1.5
T Cells	BCU173_TC_ChIP_H3K4me1_1	BCU173_TC_ChIP_Input_1	109,464,264	99,658,978	95,805,330	88,615,808	95.6	94.8	32.7	1.5
T Cells	BCU173_TC_ChIP_H3K36me3_1	BCU173_TC_ChIP_Input_1	77,399,016	71,826,548	95,805,330	88,615,808	95.9	94.8	5.9	1.5
T Cells	BCU173_TC_ChIP_H3K27me3_1	BCU173_TC_ChIP_Input_1	99,539,622	91,450,626	95,805,330	88,615,808	96.7	94.8	31.3	1.5
T Cells	BCU173_TC_ChIP_H3K27ac_1	BCU173_TC_ChIP_Input_1	99,592,998	91,240,286	95,805,330	88,615,808	96.5	94.8	28.8	1.5
T Cells	BCU1657_TC_ChIP_H3K9me3_1	BCU1657_TC_ChIP_Input_1	98,883,444	95,996,658	107,823,298	98,375,614	94.6	96	18.5	21.2
T Cells	BCU1657_TC_ChIP_H3K4me3_1	BCU1657_TC_ChIP_Input_1	101,349,458	90,143,112	107,823,298	98,375,614	95.9	96	43.9	21.2
T Cells	BCU1657_TC_ChIP_H3K4me1_1	BCU1657_TC_ChIP_Input_1	97,545,950	88,695,396	107,823,298	98,375,614	96	96	36.9	21.2
T Cells	BCU1657_TC_ChIP_H3K36me3_1	BCU1657_TC_ChIP_Input_1	117,462,434	112,148,062	107,823,298	98,375,614	97.5	96	15.2	21.2
T Cells	BCU1657_TC_ChIP_H3K27me3_1	BCU1657_TC_ChIP_Input_1	93,898,560	88,546,036	107,823,298	98,375,614	97.1	96	47.8	21.2
T Cells	BCU1657_TC_ChIP2_H3K27ac_1	BCU1657_TC_ChIP2_Input_1	87,787,492	43,288,646	104,831,876	92,937,700	86.3	95.7	41.5	13.3
T Cells	BCU1595_TC_ChIP_H3K9me3_1	BCU1595_TC_ChIP_Input_1	103,558,764	101,716,034	82,773,480	80,745,042	94.8	96.1	5.4	9.4
T Cells	BCU1595_TC_ChIP_H3K4me3_1	BCU1595_TC_ChIP_Input_1	74,044,818	72,981,358	82,773,480	80,745,042	96.3	96.1	6.8	9.4
T Cells	BCU1595_TC_ChIP_H3K4me1_1	BCU1595_TC_ChIP_Input_1	214,211,844	202,262,888	82,773,480	80,745,042	94.9	96.1	27.6	9.4
T Cells	BCU1595_TC_ChIP_H3K36me3_1	BCU1595_TC_ChIP_Input_1	52,803,894	48,442,898	82,773,480	80,745,042	95.4	96.1	75.3	9.4
T Cells	BCU1595_TC_ChIP_H3K27me3_1	BCU1595_TC_ChIP_Input_1	103,706,922	100,886,308	82,773,480	80,745,042	97.2	96.1	5.2	9.4
T Cells	BCU1595_TC_ChIP_H3K27ac_1	BCU1595_TC_ChIP_Input_1	63,891,320	56,749,058	82,773,480	80,745,042	92	96.1	53.6	9.4
T Cells	BCU1571_TC_ChIP_H3K9me3_1	BCU1571_TC_ChIP_Input_1	86,863,022	82,607,920	87,587,868	80,953,860	92.8	96.4	4.6	1.2
T Cells	BCU1571_TC_ChIP_H3K4me3_1	BCU1571_TC_ChIP_Input_1	82,430,370	79,257,958	87,587,868	80,953,860	96.4	96.4	2.8	1.2
T Cells	BCU1571_TC_ChIP_H3K4me1_1	BCU1571_TC_ChIP_Input_1	202,163,068	193,694,544	87,587,868	80,953,860	96.6	96.4	7.6	1.2
T Cells	BCU1571_TC_ChIP_H3K36me3_1	BCU1571_TC_ChIP_Input_1	113,770,750	110,975,028	87,587,868	80,953,860	97.7	96.4	3.5	1.2
T Cells	BCU1571_TC_ChIP_H3K27me3_1	BCU1571_TC_ChIP_Input_1	106,077,946	102,570,104	87,587,868	80,953,860	97.9	96.4	23.5	1.2

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
T Cells	BCU1571_TC_ChIP_H3K27ac_1	BCU1571_TC_ChIP_Input_1	65,434,854	62,424,484	87,587,868	80,953,860	96.4	96.4	17.7	1.2
T Cells	BCU1571_TC_ChIP2_H3K9me3_1	BCU1571_TC_ChIP2_Input_1	104,528,632	95,668,804	148,588,780	135,411,316	92.4	95.3	4.5	3.3
T Cells	BCU1571_TC_ChIP2_H3K4me3_1	BCU1571_TC_ChIP2_Input_1	125,199,072	102,276,658	148,588,780	135,411,316	93.4	95.3	4.5	3.3
T Cells	BCU1571_TC_ChIP2_H3K4me1_1	BCU1571_TC_ChIP2_Input_1	81,711,498	69,931,382	148,588,780	135,411,316	92	95.3	6.8	3.3
T Cells	BCU1571_TC_ChIP2_H3K36me3_1	BCU1571_TC_ChIP2_Input_1	105,188,104	94,711,450	148,588,780	135,411,316	92.4	95.3	10.6	3.3
T Cells	BCU1571_TC_ChIP2_H3K27me3_1	BCU1571_TC_ChIP2_Input_1	99,289,422	90,375,762	148,588,780	135,411,316	94.7	95.3	5.8	3.3
T Cells	BCU133_TC_ChIP_H3K4me3_1	BCU133_TC_ChIP_Input_1	120,404,062	118,424,594	122,679,248	119,606,612	92.6	96.4	24.9	2.8
T Cells	BCU133_TC_ChIP_H3K4me1_1	BCU133_TC_ChIP_Input_1	124,080,086	121,860,912	122,679,248	119,606,612	97.3	96.4	5.5	2.8
Monocyte	BCU133_Mono_ChIP_H3K4me3_1	BCU133_Mono_ChIP_Input_1	85,431,646	80,493,644	121,716,226	119,745,890	87.3	97.2	57.3	5.1
Monocyte	BCU133_Mono_ChIP_H3K4me1_1	BCU133_Mono_ChIP_Input_1	120,176,822	116,513,064	121,716,226	119,745,890	98.2	97.2	8.7	5.1
Monocyte	BCU133_Mono_ChIP_H3K27ac_1	BCU133_Mono_ChIP_Input_1	114,883,734	110,395,654	121,716,226	119,745,890	98.3	97.2	19.1	5.1
T Cells	BCU1053_TC_ChIP_H3K9me3_1	BCU1053_TC_ChIP_Input_1	98,413,794	84,519,144	115,682,520	106,983,764	93	95.1	2.1	3.7
T Cells	BCU1053_TC_ChIP_H3K4me3_1	BCU1053_TC_ChIP_Input_1	100,007,790	84,288,228	115,682,520	106,983,764	94.6	95.1	5.5	3.7
T Cells	BCU1053_TC_ChIP_H3K4me1_1	BCU1053_TC_ChIP_Input_1	356,151,640	330,959,256	115,682,520	106,983,764	95.6	95.1	16.3	3.7
T Cells	BCU1053_TC_ChIP_H3K36me3_1	BCU1053_TC_ChIP_Input_1	104,858,348	96,763,564	115,682,520	106,983,764	95.7	95.1	33.5	3.7
T Cells	BCU1053_TC_ChIP_H3K27me3_1	BCU1053_TC_ChIP_Input_1	107,572,674	92,636,848	115,682,520	106,983,764	96.3	95.1	4.6	3.7
T Cells	BCU1053_TC_ChIP_H3K27ac_1	BCU1053_TC_ChIP_Input_1	70,732,760	60,930,908	115,682,520	106,983,764	95.1	95.1	85.2	3.7
Muscle	TB_Muscle_ChIP_H3K4me1_1	TB_Muscle_ChIP_Input_1	67,934,800	58,204,688	118,566,468	103,214,680	94	94.7	26.4	2
Muscle	TB_Muscle_ChIP_H3K36me3_1	TB_Muscle_ChIP_Input_1	102,342,796	77,817,272	118,566,468	103,214,680	93.4	94.7	7.5	2
Muscle	TB_Muscle_ChIP_H3K27me3_1	TB_Muscle_ChIP_Input_1	104,090,356	90,576,776	118,566,468	103,214,680	84.3	94.7	11.2	2
Muscle	TB_Muscle_ChIP2_H3K9me3_1	TB_Muscle_ChIP2_Input_1	151,981,796	127,761,548	64,326,758	56,874,584	91.5	95.8	7.4	1.1
Muscle	TB_Muscle_ChIP2_H3K4me3_1	TB_Muscle_ChIP2_Input_1	61,273,804	55,599,438	64,326,758	56,874,584	98.4	95.8	5.8	1.1
Muscle	RC_Muscle_ChIP_H3K9me3_1	RC_Muscle_ChIP_Input_1	113,190,242	94,865,822	84,831,728	74,443,196	91.5	95.3	19.6	1.6
Muscle	RC_Muscle_ChIP_H3K4me3_1	RC_Muscle_ChIP_Input_1	129,847,806	95,781,686	84,831,728	74,443,196	74.1	95.3	33.9	1.6

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
Muscle	RC_Muscle_ChIP_H3K4me1_1	RC_Muscle_ChIP_Input_1	115,211,096	109,493,096	84,831,728	74,443,196	98.6	95.3	10.8	1.6
Muscle	RC_Muscle_ChIP_H3K36me3_1	RC_Muscle_ChIP_Input_1	130,336,670	114,244,436	84,831,728	74,443,196	97.9	95.3	10.8	1.6
Muscle	RC_Muscle_ChIP_H3K27me3_1	RC_Muscle_ChIP_Input_1	98,931,668	90,576,614	84,831,728	74,443,196	95.6	95.3	12.3	1.6
Muscle	RC_Muscle_ChIP_H3K27ac_1	RC_Muscle_ChIP_Input_1	75,610,904	69,428,630	84,831,728	74,443,196	98.8	95.3	5.1	1.6
Muscle	RA_Muscle_ChIP_H3K9me3_1	RA_Muscle_ChIP_Input_1	63,806,416	59,266,020	141,046,192	121,817,052	91.3	95.5	7.1	7.1
Muscle	RA_Muscle_ChIP_H3K4me3_1	RA_Muscle_ChIP_Input_1	116,062,294	94,638,400	141,046,192	121,817,052	92.8	95.5	44.2	7.1
Muscle	RA_Muscle_ChIP_H3K4me1_1	RA_Muscle_ChIP_Input_1	118,332,380	93,931,506	141,046,192	121,817,052	95.1	95.5	55.6	7.1
Muscle	RA_Muscle_ChIP_H3K36me3_1	RA_Muscle_ChIP_Input_1	95,587,854	79,604,414	141,046,192	121,817,052	97.3	95.5	25.8	7.1
Muscle	RA_Muscle_ChIP_H3K27me3_1	RA_Muscle_ChIP_Input_1	103,780,622	88,456,384	141,046,192	121,817,052	79.2	95.5	80.5	7.1
Muscle	RA_Muscle_ChIP_H3K27ac_1	RA_Muscle_ChIP_Input_1	92,134,798	83,568,418	141,046,192	121,817,052	93.1	95.5	34.9	7.1
Muscle	MA_Muscle_ChIP_H3K9me3_1	MA_Muscle_ChIP_Input_1	72,756,656	63,694,512	104,471,744	98,251,978	89.9	96	21.5	4
Muscle	MA_Muscle_ChIP_H3K4me3_1	MA_Muscle_ChIP_Input_1	104,105,326	96,831,178	104,471,744	98,251,978	97.4	96	22.8	4
Muscle	MA_Muscle_ChIP_H3K4me1_1	MA_Muscle_ChIP_Input_1	142,134,326	130,486,114	104,471,744	98,251,978	98.1	96	2	4
Muscle	MA_Muscle_ChIP_H3K36me3_1	MA_Muscle_ChIP_Input_1	62,850,510	60,532,110	104,471,744	98,251,978	98.3	96	2.6	4
Muscle	MA_Muscle_ChIP_H3K27me3_1	MA_Muscle_ChIP_Input_1	146,490,040	135,053,606	104,471,744	98,251,978	83.5	96	20.8	4
Muscle	MA_Muscle_ChIP_H3K27ac_1	MA_Muscle_ChIP_Input_1	116,165,714	101,373,752	104,471,744	98,251,978	85.1	96	18.1	4
Muscle	CF_Muscle_ChIP_H3K9me3_1	CF_Muscle_ChIP_Input_1	93,276,674	85,361,152	130,117,546	127,549,312	92.7	96.3	25.8	2.2
Muscle	CF_Muscle_ChIP_H3K4me3_1	CF_Muscle_ChIP_Input_1	115,167,458	110,459,466	130,117,546	127,549,312	96	96.3	26.6	2.2
Muscle	CF_Muscle_ChIP_H3K4me1_1	CF_Muscle_ChIP_Input_1	95,667,480	88,948,132	130,117,546	127,549,312	97.6	96.3	12.2	2.2
Muscle	CF_Muscle_ChIP_H3K36me3_1	CF_Muscle_ChIP_Input_1	118,177,814	111,035,660	130,117,546	127,549,312	98.1	96.3	7.2	2.2
Muscle	CF_Muscle_ChIP_H3K27me3_1	CF_Muscle_ChIP_Input_1	125,087,950	119,936,162	130,117,546	127,549,312	87.1	96.3	16.4	2.2
Muscle	CF_Muscle_ChIP2_H3K27ac_1	CF_Muscle_ChIP2_Input_1	66,316,252	62,762,316	73,970,694	65,241,532	98.6	95.9	21.8	1.7
Muscle	BrW_Muscle_ChIP_H3K9me3_1	BrW_Muscle_ChIP_Input_1	129,160,288	120,393,816	61,727,108	58,719,166	93.2	96.7	3.9	1.1
Muscle	BrW_Muscle_ChIP_H3K4me3_1	BrW_Muscle_ChIP_Input_1	69,540,268	62,789,158	61,727,108	58,719,166	97	96.7	13.4	1.1

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Cell type[1]	Treatment name[2]	Control name[3]	Treatment raw_reads[4]	Treatment filtered_reads[5]	Control raw_reads[6]	Control filtered_reads[7]	%Treatment alignment[8]	%Control alignment[9]	%Treatment duplicate[10]	%Control duplicate[11]
Muscle	BrW_Muscle_ChIP_H3K4me1_1	BrW_Muscle_ChIP_Input_1	119,517,354	113,204,550	61,727,108	58,719,166	98.1	96.7	11.8	1.1
Muscle	BrW_Muscle_ChIP_H3K36me3_1	BrW_Muscle_ChIP_Input_1	131,095,368	126,522,822	61,727,108	58,719,166	98.9	96.7	5.3	1.1
Muscle	BrW_Muscle_ChIP_H3K27me3_1	BrW_Muscle_ChIP_Input_1	129,635,198	124,930,362	61,727,108	58,719,166	56.7	96.7	18.2	1.1
Muscle	BrW_Muscle_ChIP_H3K27ac_1	BrW_Muscle_ChIP_Input_1	68,635,108	64,196,436	61,727,108	58,719,166	99	96.7	8	1.1
Muscle	BeW_Muscle_ChIP_H3K9me3_1	BeW_Muscle_ChIP_Input_1	132,194,366	125,473,020	65,898,052	63,440,836	94	96.5	12.5	1.6
Muscle	BeW_Muscle_ChIP_H3K4me3_1	BeW_Muscle_ChIP_Input_1	70,409,804	65,612,736	65,898,052	63,440,836	97.4	96.5	23.2	1.6
Muscle	BeW_Muscle_ChIP_H3K4me1_1	BeW_Muscle_ChIP_Input_1	118,526,084	114,288,660	65,898,052	63,440,836	98.5	96.5	24.3	1.6
Muscle	BeW_Muscle_ChIP_H3K36me3_1	BeW_Muscle_ChIP_Input_1	121,025,784	115,865,268	65,898,052	63,440,836	98.7	96.5	4.5	1.6
Muscle	BeW_Muscle_ChIP_H3K27me3_1	BeW_Muscle_ChIP_Input_1	129,870,628	124,866,766	65,898,052	63,440,836	94.9	96.5	7.1	1.6
Muscle	BeW_Muscle_ChIP_H3K27ac_1	BeW_Muscle_ChIP_Input_1	61,220,034	57,238,454	65,898,052	63,440,836	98.4	96.5	20.1	1.6
Muscle	AM_Muscle_ChIP_H3K9me3_1	AM_Muscle_ChIP_Input_1	146,496,114	143,559,638	70,427,534	68,614,628	93.3	96.8	30	2.2
Muscle	AM_Muscle_ChIP_H3K4me3_1	AM_Muscle_ChIP_Input_1	71,662,862	65,410,990	70,427,534	68,614,628	97.2	96.8	33.2	2.2
Muscle	AM_Muscle_ChIP_H3K36me3_1	AM_Muscle_ChIP_Input_1	140,941,542	137,739,794	70,427,534	68,614,628	98.9	96.8	6.7	2.2
Muscle	AM_Muscle_ChIP_H3K27me3_1	AM_Muscle_ChIP_Input_1	123,942,856	119,870,202	70,427,534	68,614,628	96.5	96.8	28.3	2.2
Muscle	AM_Muscle_ChIP_H3K27ac_1	AM_Muscle_ChIP_Input_1	59,961,376	57,066,648	70,427,534	68,614,628	98.4	96.8	4.5	2.2
Muscle	AD_Muscle_NChIP_H3K4me3_1	AD_Muscle_NChIP_Input_1	67,087,240	64,373,406	77,082,118	74,001,218	97.2	96.8	2.8	2.7
Muscle	AD_Muscle_NChIP_H3K27ac_1	AD_Muscle_NChIP_Input_1	84,358,100	79,675,056	77,082,118	74,001,218	99.2	96.8	1.2	2.7
Muscle	AD_Muscle_ChIP_H3K9me3_1	AD_Muscle_ChIP_Input_1	78,402,412	72,785,404	101,507,100	94,486,088	91.6	96.6	10.5	1.8
Muscle	AD_Muscle_ChIP_H3K4me3_1	AD_Muscle_ChIP_Input_1	102,878,242	87,598,382	101,507,100	94,486,088	95.7	96.6	9.2	1.8
Muscle	AD_Muscle_ChIP_H3K4me1_1	AD_Muscle_ChIP_Input_1	102,337,350	94,307,338	101,507,100	94,486,088	97.9	96.6	17.2	1.8
Muscle	AD_Muscle_ChIP_H3K36me3_1	AD_Muscle_ChIP_Input_1	67,576,398	59,400,090	101,507,100	94,486,088	96.6	96.6	5.5	1.8
Muscle	AD_Muscle_ChIP_H3K27me3_1	AD_Muscle_ChIP_Input_1	88,988,076	85,701,024	101,507,100	94,486,088	86.9	96.6	2.8	1.8
Muscle	AD_Muscle_ChIP2_H3K27ac_1	AD_Muscle_ChIP2_Input_1	71,397,222	63,884,870	81,982,886	74,149,378	98.9	95.9	1.2	1.9

[1]Cell type, [2]HM ChIP-Seq ID, [3]DNA input ID, [4]number of HM reads, [5]number of filtered HM reads, [6]number of DNA input reads, [7]number of filtered DNA input reads, [8]rate of aligned HM reads, [9]rate of aligned DNA input reads, [10]rate of duplicated HM reads, [11]rate of duplicated DNA input reads

Table 0.3: Detailed alignment statistics of WGBS samples

Cell type[1]	Sample[2]	raw_reads[3]	filtered_reads[4]	%_forward_aln[5]	%_reverse_aln[6]	%_total_aln[7]	%_forward_aligned_duplicate[8]	%_reverse_aligned_duplicate[9]	lambda_conversion_rate[10]	cc_conversion_rate[11]	ct_conversion_rate[12]	ca_conversion_rate[13]	mean_genome_coverage[14]
Monocyte	BCU566_Mono_BS_1	339,815,792	313,986,398	33.5	33.7	67.2	4.6	4.6	95.8	99.4	99.3	99.2	6.8
Monocyte	BF773_Mono_BS_2	1,142,654,628	1,078,902,884	17.4	17.5	34.9	51.7	51.5	96.9	99.6	99.5	99.4	12.4
Muscle	AD_Muscle_BS_1	884,712,012	736,716,362	26.7	26.8	53.4	10.1	10.1	99.5	99.6	99.3	98.7	14.5
Muscle	AM_Muscle_BS_1	374,838,210	359,034,042	35.9	36	71.9	2.9	3	99.7	99.8	99.6	99.2	7.6
Muscle	BeW_Muscle_BS_1	1,179,833,910	1,138,560,782	36	36.1	72.1	4.9	5.2	99.4	99.5	99.2	98.6	24.6
Muscle	BrW_Muscle_BS_1	1,226,848,282	1,165,071,590	35	35.1	70.1	5.4	5.5	99.6	99.7	99.5	99	23.8
Muscle	CF_Muscle_BS_1	287,196,476	261,960,912	29.5	29.8	59.3	12	12.1	99.5	99.7	99.5	99	5
Muscle	MA_Muscle_BS_1	1,043,932,700	980,518,236	34.2	34.4	68.6	5.8	5.9	99.6	99.7	99.5	99	21.2
Muscle	RA_Muscle_BS_1	1,044,494,338	1,002,758,084	32.5	32.8	65.2	14.1	14.1	97.4	99.6	99.5	99.1	21
Muscle	RC_Muscle_BS_1	1,162,398,196	1,099,804,692	32.6	32.8	65.3	9.3	9.3	99.6	99.7	99.5	99.1	22.1
Muscle	TB_Muscle_BS_1	1,343,921,156	1,187,931,522	27.4	27.5	54.9	10.7	10.7	99.6	99.7	99.5	99	22.3
T Cells	BCU1053_TC_BS_1	970,130,460	935,957,848	35.8	36	71.8	5.3	5.2	97.5	99.5	99.4	99.3	20.9
T Cells	BCU1571_TC_BS_1	948,135,542	915,594,088	36.8	36.9	73.7	3.8	3.7	97.3	99.3	99.2	98.8	21.1
T Cells	BCU1595_TC_BS_1	1,118,252,158	1,043,536,276	32	32.1	64.1	7.7	7.6	96.7	99.2	99.2	98.8	21.7
T Cells	BCU173_TC_BS_1	1,004,859,052	954,129,620	35.2	35.4	70.7	4.8	5	97.1	99.5	99.4	99.3	21.9
T Cells	BCU1787_TC_BS_1	831,607,796	796,105,736	33.4	33.6	67	7.4	7.2	94.6	99	98.9	98.8	16.4
T Cells	BCU1799_TC_BS_1	1,108,961,994	1,073,944,080	26.9	27	53.9	29	28.9	97.6	99.6	99.6	99.4	18.7
T Cells	BCU1945_TC_BS_1	824,988,074	794,552,634	35.4	35.7	71.2	4.4	4.5	95.8	99.3	99.2	99.1	17.8

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]rate of forward alignment, [6]rate of reverse alignment, [7]total alignment rate, [8]rate of forward aligned duplicated reads, [9]rate of reverse aligned duplicated reads, [10]bisulfite conversion rate(CG context), [11]bisulfite conversion rate(CC context), [12]bisulfite conversion rate(CT context), [13]bisulfite conversion rate(CA context) and [14]genome coverage.

Cell type[1]	Sample[2]	raw_reads[3]	filtered_reads[4]	%_forward_aln[5]	%_reverse_aln[6]	%_total_aln[7]	%_forward_aligned_duplicate[8]	%_reverse_aligned_duplicate[9]	lambda_conversion_rate[10]	cc_conversion_rate[11]	ct_conversion_rate[12]	ca_conversion_rate[13]	mean_genome_coverage[14]
T Cells	BCU292_TC_BS_1	923,804,274	880,022,470	32.6	32.9	65.5	6.8	6.8	95.8	98.7	98.7	98.6	16.9
T Cells	BCU551_TC_BS_1	862,671,794	821,677,360	34.4	34.5	68.8	7.4	7.4	97.1	99.3	99.3	98.9	18.5
T Cells	BCU582_TC_BS_1	934,175,856	893,467,248	34.4	34.6	69	5.4	5.8	97.2	99.3	99.3	99.1	20
T Cells	BCU607_TC_BS_1	918,620,832	868,239,402	35.8	35.9	71.7	4	3.8	97.2	99.5	99.5	99.3	19.8
T Cells	BCU768_TC_BS_1	864,883,540	808,125,624	34.2	34.3	68.5	4.8	4.7	97.3	99.5	99.4	99.2	17.8
T Cells	BCU801_TC_BS_1	899,289,090	846,652,132	33.6	33.7	67.2	7.1	7	97.5	99.3	99.3	99	18.5
T Cells	BCU899_TC_BS_1	600,257,168	574,308,006	35	35.2	70.3	5.4	5.4	96.3	99.5	99.5	99.3	12.9

[1]Cell type, [2]sample ID, [3]number of raw reads, [4]number of filtered reads, [5]rate of forward alignment, [6]rate of reverse alignment, [7]total alignment rate, [8]rate of forward aligned duplicated reads, [9]rate of reverse aligned duplicated reads, [10]bisulfite conversion rate(CG context), [11]bisulfite conversion rate(CC context), [12]bisulfite conversion rate(CT context), [13]bisulfite conversion rate(CA context) and [14]genome coverage.