UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Depicting epigenetic mechanisms involved in the regulation of pseudogene expression

Ana Margarida Esteves Ferreira

Mestrado em Biologia Molecular e Genética
Dissertação

Dissertação orientada por:
Professora Doutora Ana Rita Grosso
Professora Doutora Mónica Vieira da Cunha

2017

Acknowledgements

Começo com um agradecimento à Doutora Ana Rita Grosso pela confiança que em mim depositou para trabalhar neste projecto. Agradeço-lhe profundamente por ter sido uma constante fonte de ânimo, apoio e disponibilidade em todas as etapas deste trabalho.

Devoto também a minha apreciação à Doutora Mónica Cunha por ter aceitado o convite para co-orientar este projecto e por ter mostrado sempre interesse em ajudar-me na execução desta tarefa.

Agradeço ao Doutor Sérgio de Almeida por me ter recebido no seu laboratório no Instituto de Medicina Molecular e pelos seus contributos científicos neste projecto. Menciono também os meus colegas Alexandra, Ana, Cláudio, João, Mafalda, Ram, Robert e Sílvia por terem facilitado a minha integração no laboratório e por todo o apoio que me deram.

Ao agrupamento 1242 de Ramada que pertence ao movimento CNE que pertence à Associação Mundial de Escutismo. É profundo o orgulho que tenho em pertencer a este movimento e da forma como me ensinou, conforme as palavras do seu fundador, a tentar deixar este mundo um pouco melhor do que o encontrei. Um obrigado especial à Nídia, à Marta, à Catarina, à Inês, ao Pedro, ao Barrué e ao Ricardo, amigos com quem tive a oportunidade de me entregar, construir e trilhar o caminho em direcção ao Homem Novo, tendo como farol os ensinamentos de Karol Wojtyla e o exemplo de conversão de São Paulo.

Agradeço à minha amiga de longa data, Sofia, que sempre assinala com compromisso todos os momentos importantes da minha vida. Ao menino e meninas que alegram a minha existência taciturna com a amizade que surge da sua simples presença. São eles a Tatázinha, a Fontainhas, a Salsa e o Dani. Um agradecimento muito especial e intensamente sentido à Inês por ser a coisa mais parecida que tenho com uma irmã mais velha e, na verdade, por tudo.

Para a maior bênção que possuo: família. Não tenho em mim capacidade para colocar em palavras esta graça que é, não só ter uma família grande, mas também ter uma grande família, tanto a que vem do sangue como a que vem de vincados laços de amizade. Se da minha vida algo valedor construir, a vós o devo. É obrigatório destacar os meus avós pelo constante contributo na minha educação, a minha tia por estar sempre presente e os meus primos-irmãos pela parvoíce saudável.

E por fim, vêm os primeiros. Mãe, Pai, lembram-se quando eu era pequena e vos disse que fiquei muito triste quando descobri que afinal não eram super-heróis? Que tola! Se olhava para vocês com admiração antes, relembro-vos agora que, apesar de serem humanos de carne e osso, creio vivamente que a vossa arma para salvar o mundo é a mais forte de todas. Obrigado por partilharem esse pequeno grande mistério comigo todos os dias. Obrigado por me ensinarem a transportar esse tesouro num vaso de barro.

"Por isso, já não sou eu que vivo; é Cristo que vive em mim. E a minha vida presente vivo-a por meio da fé no Filho de Deus que me amou e deu a sua vida por mim."

Gl 2:20

Resumo Alargado

A epigenética dedica-se ao estudo de modificações que ocorrem, principalmente, sobre a dupla cadeia de DNA, sem que exista a edição da sequência nela contida (Waddington 1942b, 1942a). Graças às descobertas feitas nesta área nos últimos anos, existem vários tipos de modificações epigenéticas já descritas, entre as quais se destaca as modificações de histonas e a metilação do DNA (Li et al. 2007). Assim, avaliando a presença ou ausência destas modificações, poderemos inferir relativamente à activação ou silenciamento de uma determinada região do genoma. Vários estudos têm sido realizados para caracterizar a forma como estas modificações afectam a transcrição de genes codificadores de proteína (Kouzarides 2007), no entanto, pouco se sabe como estas modificações podem condicionar outras classes de genes, nomeadamente, os pseudogenes. Neste sentido, o objectivo deste trabalho consiste na determinação de modificações epigenéticas que possam estar envolvidas na expressão dos pseudogenes, potencialmente exercendo um papel crucial na sua regulação.

Os pseudogenes são cópias ancestrais de sequências codificantes que, possivelmente devido à perda de pressão selectiva, degeneraram em novas unidades genéticas (Jacq et al. 1977). Actualmente, os pseudogenes são classificados em três grandes grupos que são definidos com base no seu processo de formação: processados, a classe de pseudogenes mais representada e cuja formação envolve um processo de transcrição reversa e integração de um RNA mensageiro novamente no DNA, num processo conhecido por retrotransposição; não processados, no caso do processo de formação do pseudogene acontecer através da duplicação de um gene completo; e unitários, quando a própria estrutura física do gene sofre modificações que levam à perda da capacidade de codificar uma proteína (Pink et al. 2011). O processo de formação dos pseudogenes que resulta na incapacidade do novo pseudogene codificar uma proteína denomina-se "pseudogenização" (Gregório 2016).

Graças ao recente desenvolvimento de plataformas de sequenciação em larga escala, revelou-se que os pseudogenes são transcritos e que a sua transcrição pode estar envolvida na condução de importantes processos celulares nos quais os pseudogenes podem desempenhar funções celulares específicas. Presentemente, sabe-se que os pseudogenes conseguem também actuar através de diferentes mecanismos para modular a regulação dos seus genes parentais, nomeadamente através da competição para esponjas de microRNAs (Thomson and Dinger 2016), transcritos *antisense* ou lncRNAs com a capacidade de conduzir complexos proteicos remodeladores de cromatina (Groen et al. 2014). Para além desta actuação mediada por RNA através dos potenciais transcritos dos pseudogenes, pensa-se também que os pseudogenes podem ter mecanismos de acção ao nível do DNA que podem condicionar a actividade do gene parental, por exemplo através de um evento de recombinação homóloga entre o pseudogene e o gene parental que pode resultar na deleção do gene parental (Poliseno 2012). Dada esta possível contribuição em vários processos celulares, os pseudogenes definem um novo paradigma de como o genoma não codificante pode ter importantes contribuições em diversas funções biológicas, nomeadamente no desenvolvimento e no cancro. Um exemplo destas contribuições é o pseudogene *Oct4p4,* que tem a capacidade de regular a transcrição do seu gene parental, o regulador de pluripotência *Oct4*. Quando expresso, este pseudogene conduz a célula a iniciar o processo de diferenciação neural, através da imposição da modificação repressiva da histona H3 (H3K9me3) na região promotora do gene *Oct4* (Liedtke et al. 2007). Um outro exemplo de um pseudogene com uma função importante, neste caso em cancro, é o *PTENP1,* um pseudogene do gene supressor tumoral *PTEN*. O *PTENP1* é o exemplo de um pseudogene com diversificados mecanismos de acção através de um único pseudogene conseguindo actuar como uma esponja de microRNAs, um catalisador do recrutamento de remodeladores da cromatina para o promotor do gene *PTEN* e um transcrito *antisense* que consegue regular a estabilidade e a função de esponja de microRNAs do próprio transcrito sense do *PTENP1* (Johnsson et al. 2013).

Contudo, os mecanismos pelos quais a expressão dos pseudogenes é regulada e qual o seu papel biológico estão ainda por explorar. Grande porção dos pseudogenes não aparentam ter sequências regulatórias a montante do corpo do pseudogene, o que pode sugerir que outros mecanismos poderão estar envolvidos neste processo, em resultado da observação de modificações nas histonas de pseudogenes que são transcritos e que não são características nos seus genes parentais ou nos restantes genes codificadores de proteínas (Pei et al. 2012). Um destes exemplos é a presença de H3K9me3 na região do promotor de pseudogenes expressos (Guo et al. 2014).

Tendo em consideração estas observações, propomos a hipótese que os pseudogenes possuem mecanismos epigenéticos próprios a regular a sua transcrição. Para testar esta hipótese, estudámos o transcriptoma e epigenoma dos pseudogenes durante a diferenciação neural de células estaminais embrionárias, através da combinação de análises de dados em larga de escala do transcriptoma (RNA-seq e GRO-seq), metilação de DNA (BS-seq), regiões de cromatina aberta (hipersensibilidade à DNase) e modificações de histona (ChIP-seq). Os dados usados foram obtidos através da plataforma NIH Roadmap Epigenomics Consortium (Bernstein et al. 2010), consistindo em 72 amostras e um total de 194 replicados. Devido à elevada expressão de pseudogenes no cérebro (Pei et al. 2012), este projecto incidiu essencialmente na diferenciação neural, durante a qual células estaminais embrionárias (H1) foram diferenciadas *in vitro* em células progenitoras neuronais (H1N).

As nossas análises referentes ao transcriptoma revelaram um número mais elevado de pseudogenes a serem expressos durante a diferenciação neural quando comparado com a diferenciação mesenquimal. No entanto, observámos que a detecção da transcrição dos pseudogenes pode ser incorrectamente determinada usando dados de RNA-seq, pois os perfis obtidos por esta tecnologia são influenciados pela estabilidade dos transcritos. Em concordância, os resultados obtidos usando dados de GRO-seq suportam esta hipótese, dado que permitem identificar um maior número de pseudogenes a serem transcritos. Após a identificação dos pseudogenes transcritos e silenciados, analisámos o seu enriquecimento em modificações de histonas. De todas as alterações observadas, destacamos três importantes observações associadas com a transcrição de pseudogenes, nomeadamente a presença de: H3K36me3 no corpo do pseudogenes transcritos, associada a episódios de continuação da transcrição do gene na região a montante ("read-through"); H3K9me3, uma marca epigenética usualmente associada a regiões não transcritas; e, por fim, domínios bivalentes (H3K4me3 e H3K27me3) na região promotora de alguns pseudogenes. Estas observações parecem sustentar a hipótese que sugere que a transcrição dos pseudogenes é regulada. Estudos mais profundos são necessários para perceber a extensão destas modificações na expressão dos pseudogenes, apesar da presença de H3K36me3 e H3K9me3 terem sido já observadas previamente em pseudogenes transcritos (Pei et al. 2012; Guo et al. 2014).

No entanto, são ainda muitas as limitações associadas ao estudo dos pseudogenes e que precisam de um melhoramento no futuro. Primeiramente, a semelhança existente entre pseudogenes e os genes parentais dificulta o mapeamento destas regiões usando dados de sequenciação de transcriptoma. Adicionalmente, a expressão de pseudogenes por "read-through" do gene a montante pode sugerir a existência de erros na anotação de bases de dados e pressiona para a crescente necessidade de melhoramento na caracterização de genomas.

Concluindo, os resultados aqui observados e discutidos confirmam que os pseudogenes são transcritos e que a sua transcrição parece ser regulada, sugerindo que o seu papel não será assim tão "pseudo" como previamente se pensava. Contudo, mais esforços são necessários para caracterizar a extensão destas alterações, bem como para aferir a contribuição da metilação do DNA na regulação da expressão dos pseudogenes.

Palavras-chave: Pseudogenes, transcrição, epigenética, biologia computacional

Abstract

Pseudogenes are genetic elements that derive from normal protein-coding genes which, through the accumulation of deteriorating mutations, have lost coding potential in a process which is known as "pseudogenization". However, recent high throughput sequencing technology has shown that pseudogenes are transcribed and that their transcription is tissue-specific, which suggests that pseudogenes might have an important role in biological processes. Many pseudogenes have been described to regulate important processes in development or cancer. Yet, not much is known about how pseudogene expression is regulated. Most pseudogenes seem to have lost their upstream regulatory sequences, indicating that trans-acting mechanisms might be responsible for this regulation. Studies evidence that pseudogenes have different histone modifications compared to their parental genes, suggesting that they might have specific transcriptional mechanisms.

In this project, we aimed at identifying the epigenetic pattern responsible for the regulation of pseudogene transcription through a genome-wide analysis. For this analysis, we used transcriptomic data (RNA-seq and GRO-seq) to detect pseudogene transcription and epigenomic data (ChIP-seq, DNase Hypersensitivity and WGBS-seq) to assess epigenomic changes in silent and expressed pseudogenes. Since pseudogene expression has been shown to be higher in the brain, we choose to address our research questions using *in vitro* neural differentiation of embryonic stem cells (ESCs) as a cell differentiation model system.

Our analysis confirmed that there are more pseudogenes being expressed during neural differentiation when compared to mesenchymal differentiation. Regarding their epigenetic modifications, our results show that some pseudogenes, in which the histone modification H3K36me3 is present, might be transcribed as a consequence of transcription read-through from the upstream gene. Expressed pseudogenes also seem to be enriched with the histone modification H3K9me3, a modification that is known to be associated with inactive transcription. As well as in protein-coding genes and lncRNAs, pseudogenes are enriched with bivalent promoters features, such as the co-localized presence of H3K4me3 and H3K27me3 in both undifferentiated and neural differentiated cell lines.

To conclude, although the regulation of pseudogene transcription still requires further work to truly apprehend the epigenetic mechanisms that contribute to pseudogene expression, our work has confirmed that mainly histone modification such as H3K36me3 and H3K9me3 may indeed play a role, either direct or indirect, that can help modulate the expression of these very particular genes.

Keywords: pseudogenes; transcriptional regulation; epigenome; computational biology

Table of Contents

List of Figures and Tables

Figures

Tables

List of Abbreviations

bp base pair (s)

ChIP-seq Chromatin Immunoprecipitation sequencing

DNA Desoxirribonucleic Acid

ESC Embryonic stem cells

FDR False discovery rate

GRCh38 Genome Reference Consortium human genome (build 38)

GRO-seq Globan Run-On sequencing

H1 H1 human embryonic stem cells

H1M H1 derived mesenchymal stem cell

H1N H1 derived neuronal progenitor cultured cells

HTS High-throughput Sequencing

lincRNA Long intergenic non coding RNAs

lncRNA Long non coding RNAs

mRNA messenger RNA

NGS Next Generation Sequencing

NIH National Institute of Health

PCR Polimerase Chain Reaction

RNA Ribonucleic Acid

RISC RNA-induced silencing complex

RPKMs Reads per kilobase per million

RNA-seq RNA sequencing

TPMs transcripts per million

TSS transcription start site

TTS transcription termination site

UCSC University of California Santa Cruz

UTR untranslated terminal region

WGBS-seq Whole genome bisulfite sequencing

1. Introduction

1.1. The role of the epigenome in the regulation of gene expression

The term "*epigenetics*" arises for the first time in 1942 attributing now a word to the concept of phenotypic change without genotypic change (Waddington 1942b, 1942a). Nowadays, it is known that the DNA template works in collaboration with epigenetic programs to regulate gene expression through several mechanisms (Figure 1.1). The advances in epigenetics have gone from the identification of different levels of chromatin condensation (transcriptionally active regions as "euchromatin" and silent regions as "heterochromatin"), to more detailed insights such as the effects of histone modification and DNA methylation in transcriptional programs (Li et al. 2007). These last two epigenetic features will be discussed in this project.

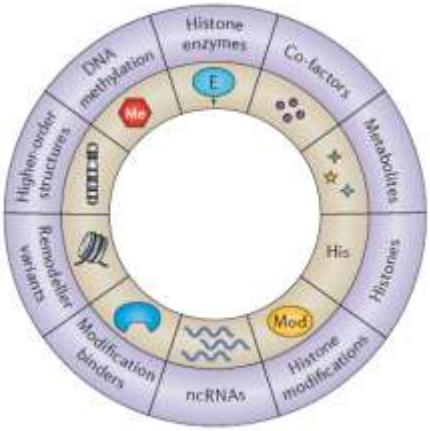

Figure 1.1 – Examples of different molecular mechanisms of epigenetic control (adapted from Allis and Jenuwein 2016**).**

1.1.1. Histone Modifications

The DNA inside the nucleus is compacted as a chromatin fibber which is organized in nucleosomes, the building-block structure of the genome. The nucleosome structure is composed of an octamer of four histones, namely H3, H4, H2A, and H2B, which are encircled by 147 base pairs of DNA (Kornberg and Lorch 1999). It is known that histones have large N-terminal tails that can suffer several modifications, such as methylation, acetylation or phosphorylation, which can affect several DNA-related processes, including transcription (Karlić et al. 2010), splicing (Kornblihtt et al. 2009) or DNA repair (Fillingham et al. 2006) (Figure 1.2, A).

The development of high-throughput sequencing (HTS) technologies has allowed the mapping of all these modifications across the genome. This transformation allowed the association of specific modifications to regulatory processes (Kouzarides 2007), namely transcription (Li et al. 2007; Soboleva et al. 2014). There have been many histone modifications that were described to regulate transcription. Transcription regulation can be performed at several levels, namely through specific modifications in enhancer, promoter and gene body regions (Figure 1.2, B). For instance, trimethylation of lysine 4 in histone H3 (H3K4me3) in promoter region and trimethylation of lysine 36 also in histone H3 (H3K36me3) throughout the gene body correlates positively with active transcription. On the contrary, presence of trimethylation of lysine 9 in histone H3 (H3K9me3) and trimethylation of lysine 27 in histone H3 (H3K27me3) are usually associated with transcriptional

repression. Interestingly, H3K4me3 and H3K27me3 histone modifications are also concomitant with bivalent chromatin associated structures that are characteristic of early development stages in cell differentiation. In these bivalent domains, transcription activation and repression histone modifications are co-existent in a regulated equilibrium (Voigt et al. 2013). Throughout differentiation, there is a tendency for these regions to undergo silencing, which results in the decrease of these bimodal domains. It has been observed that, in ES cells, these bivalent regions are associated to pluripotency factors, such as OCT4, delivering their contribution at maintaining basal activation levels for these factors (Voigt et al. 2013).

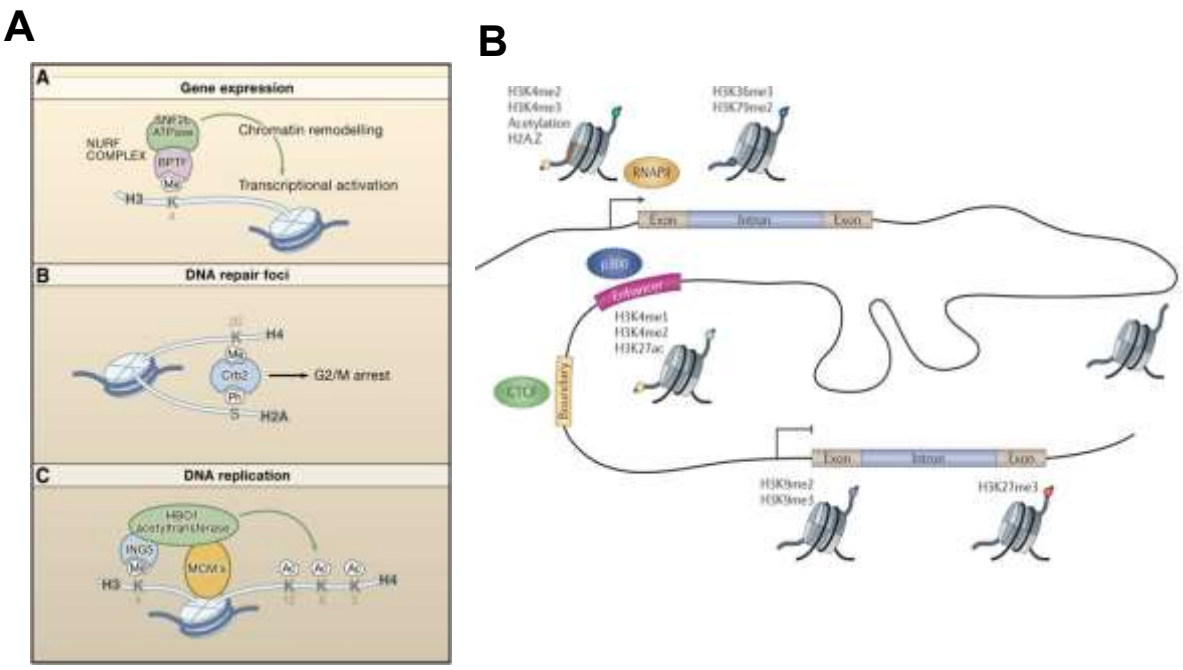

Figure 1.2 - Effects of different histone modifications on the determination of functional activity in the genome. (A) (A.A) NURF complex being transported to H3K4me locations to induce changes in transcription. (A.B) DNA repair response prompted recruitment of Crb2 complex to DNA repair foci. (A.C) Transport of HBO1 acetyltransferase to H3K4me3 in DNA replication sites (adapted from Kouzarides 2007). (B) Functional consequences of histone modifications in transcriptional dynamics**.** According to their expression level, genes and promoter regions are enriched with different histone modifications that help shape the robustness of transcriptional programs**.** Besides genes, other genomic transcriptional regulatory elements, such as enhancers, also seem to be characterized by the presence of specific histone modifications (adapted from Zhou et al. 2011).

1.1.2. DNA methylation

DNA methylation is an epigenetic modification that occurs in the DNA itself, through which, in eukaryotes, a methyl group is added to the fifth position of the cytosine nitrogenous base ring in the cytosine-guanine dinucleotides (CpG) (Holliday and Pugh 1975). This modification is highly conserved in both animal and plants (Law and Jacobsen 2010) and is thought to be present in 60-80% of the estimated 28 million CpG dinucleotides found in somatic cells (Smith and Meissner 2013). CpGs occur in CG-dense regions called CpG islands, predominant in transcription initiation sites. DNA methylation of gene promoter regions is associated with a decrease in gene expression and can lead to gene silencing (Suzuki and Bird 2008). These findings have established the role of DNA methylation in the definition of repressed chromatin states and silent gene activity. In mammals, the addition of this methyl group is completed by the family of methyltransferases DNA methyltransferase 3 (*DNMT3*), for de novo methylation, and DNA methyltransferase 1 (*DNMT1*),

responsible for the maintenance of the methylation pattern during replication (Cheng and Blumenthal 2008). The tight regulation of CpG methylation heritability (Smith and Meissner 2013) suggests that this modification must be of great importance in the maintenance of the stability of several DNA metabolic processes.

1.2. Pseudogene – a mysterious genetic element

The definition of pseudogene appeared for the first time in 1977, when Jacq et al. described the 5S DNA, coding for oocyte type 5S RNA, of *Xenopus laevis*. This 5S DNA is composed of several repeats of a 700 base pairs sequence which included a long spacer, the gene, a linker and a 101 base pairs sequence almost identical to a portion of the 121 base pairs gene to which no function was associated. The finding of this sequence led to the definition of a new genetic element which, up to this day, can still be partially described by the same words as in 1977: *"Further studies showed that this homologous structure was nearly as long as, and almost an exact repeat of, the gene itself; hence the name – pseudogene"* (Jacq et al. 1977). From 1977 to present date, the definition of pseudogene has grown more intricate. While before it was thought that pseudogenes had no coding potential, mainly due to the accumulation of deteriorating mutations, recent evidence has shown that pseudogenes are transcribed (Harrison et al. 2005; Groen et al. 2014; Kandouz et al. 2004) and can impact the expression levels of their parental genes (Liedtke et al. 2007; Poliseno et al. 2010).

1.2.1. Pseudogene Typology

Pseudogenes are very similar to regular protein-coding genes however, through time, they accumulated mutations that were capable to damage their coding potential. According to the mechanism through which they were generated, pseudogenes can be divided in 3 different classes: unprocessed, processed and unitary pseudogenes (Figure 1.3) (Pink et al. 2011).

Unprocessed pseudogenes (Figure 1.3 A) are generated by a process of duplication of an original protein-coding gene and subsequent accumulation of mutations which could have led to the loss of the coding potential and also the transcription initiation signals. (Milligan et al. 2016)

Processed pseudogenes (Figure 1.3 B) are derived through retrotransposition, a process where the transcriptional product from an original protein-coding gene is converted back to DNA and integrated in the genome. Since the mature (spliced) mRNA is the template for the reverse transcription, these pseudogenes are usually intronless (Sakai et al. 2007). Processed pseudogenes are the most represented class in mammalian genomes, possibly due to several bursts of retrotransposition (Ohshima et al. 2003).

Unitary pseudogenes (Figure 1.3 C) are the only type of pseudogenes that do not have parental genes since they are generated through the accumulation of mutations in the ancestral protein-coding gene body (Zhang et al. 2010). These mutations can lead to loss of promoter signals or coding potential (through introduction of premature stop codons, frameshift mutations or splice site alterations), process known as "pseudogenization" (Gregório 2016).

Due to the emergence of genome-wide data and computational approaches, 18000-20000 pseudogenes were thought to exist in the human genome (Torrents et al. 2003; Svensson et al. 2006). However, a more recent and exhaustive studies reduced this number to 14000 pseudogenes in the human genome (Pei et al. 2012).

Although pseudogenes are very abundant in the mammalian genome, their parental genes represent only 16% of all protein-coding genes (Pei et al. 2012; Poliseno 2012). It is also known that some

parental genes originated a greater number of pseudogenes, namely ribosomal proteins (Tonner et al. 2012), olfactory receptors and metabolic enzymes, such as *GAPDH* (Liu et al. 2009; Zhang et al. 2003).

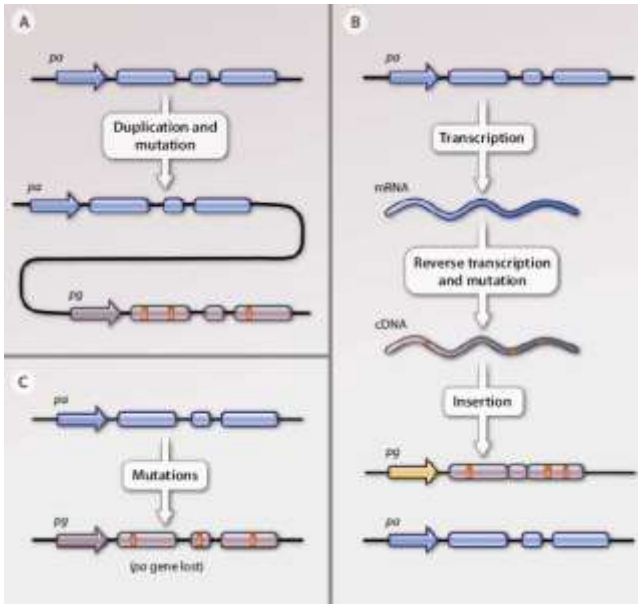

Figure 1.3 – **Pseudogene formation.** (A) Duplicated pseudogenes are copies of their parental protein-coding genes that through time acquired mutations that conditioned their function. (B) Retrotransposed pseudogenes are a consequence of the combination of the reverse transcription of a processed mRNA followed by the insertion of this structure in a random region of the genome. (C) In the bottom case, pseudogenes are formed by the degradation of the original gene structure through the accumulation of several mutations (adapted from Poliseno 2012).

1.2.2. Pseudogene Roles

The product of pseudogene transcription is thought to have a biological function in the regulation of their parental genes. Pseudogenes can affect the regulation of their parental genes mostly at RNA level although it has been described that pseudogenes can alter the structure of their parental genes at DNA level. For instance, *BRCA1* has an unprocessed pseudogene, *ΨBRCA1*, which is thought to be a potential recombination hotspot due to the extensive similarity between them. In families with breast and ovary cancer, this gene has been shown to be nonfunctional due to the loss of its promoter and initiation codon, providing a mechanism whereby this oncosuppressor gene can become inactivated in cancer (Puget et al. 2002). Another example is the pseudogene *CYP2A7*, which originates from the parental gene *CYP2A6* coding for an hepatic enzyme. *CYP2A6* gains a polymorphic site that stabilizes its mRNA leading to an augment in its abundance and stability. The enzyme resulting from this polymorphism metabolizes nicotine much faster and this genotype is often associated with an increased risk of developing lung cancer (Wang et al. 2006).

The major mechanisms through which pseudogenes can change the expression of their parental genes is through competition for microRNA sponges (Thomson and Dinger 2016), antisense transcripts or as lncRNA guides of chromatin remodeling complex proteins (Groen et al. 2014). *PTEN* is an example of a gene whose pseudogene shows a combination of several regulatory functions through the production of different RNA products from the same pseudogene, *PTENP1*. This pseudogene can, through several mechanisms, act as: 1) a lncRNA, when it is expressed in its sense

form acting as a microRNA sponge; 2) an antisense RNA, responsible for the recruitment of chromatin remodelers to the promoter of the *PTEN* gene; 3) an antisense RNA which binds to *PTENP1* sense transcript altering its stability and ability to act as a microRNA sponge (Johnsson et al. 2013). Another evidence of the great importance of this pseudogene's activity is the fact that cells induce cell-cycle arrest when the antisense RNA mechanism is disrupted (Johnsson et al. 2013).

Pseudogene expression can also have an influence on differentiation processes. One of these pseudogenes is *Oct4* pseudogene. *Oct4* is responsible for the maintenance of the undifferentiated state in embryonic stem cells. Conversely, when *Oct4* pseudogene is expressed, it is responsible for repressive chromatin rearrangements to the promoter regions of the *Oct4* gene leading to its decreased expression. Consequently, it allows the cell to enter the neural differentiation process (Liedtke et al. 2007).

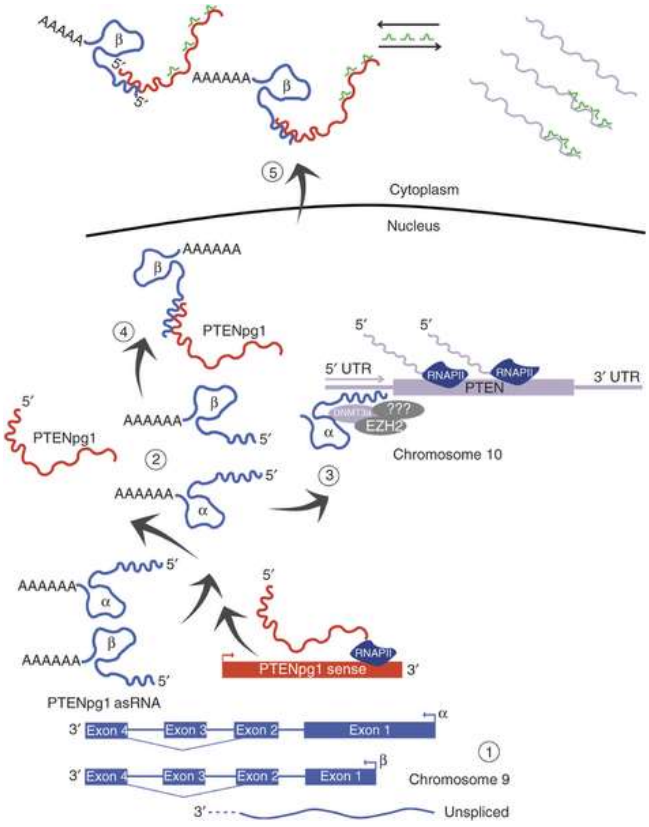

Figure 1.4 – Example of the effect of the *PTENP1* pseudogene transcription in the regulation of *PTEN* expression. *PTENP1* has two antisense RNAs, α and β. The α isoform is responsible for the epigenetic modulation of *PTEN* transcription. The β isoform pairs with PTENpg1 sense changing its stability and the interaction with microRNA sponges (adapted from Johnsson et al. 2013).

1.2.3. Pseudogene Transcriptional Regulation

Prior to the development of computational tools, the identification of pseudogene transcripts was conducted using PCR techniques which failed to be successful due to the similarity between pseudogenes and parental genes (Poliseno et al. 2010). Recent computational approaches have shown that pseudogenes are expressed (Pei et al. 2012) and that their expression can be tissue specific, not only in normal cells (Pei et al. 2012) but also in cancer (Han et al. 2014; Kalyana-Sundaram et al. 2012). Besides being tissue specific, pseudogene expression is particularly elevated in certain cell types, namely, testis, adrenal, oocytes and brain (Pei et al. 2012).

The specific patterns of pseudogene expression suggest a coordinated transcription (Han et al. 2014). However, not much is known regarding the mechanisms through which pseudogene expression is regulated. Since processed pseudogenes arise from the retrotransposition of processed mRNAs, lacking most promoter regulatory regions, their transcription regulation may differ from their cognate genes (Kandouz et al. 2004). Moreover, the upstream region of pseudogenes lack the regular transcription factor binding sites (Pei et al. 2012). On the contrary, unprocessed pseudogenes originate through genomic duplication, mechanism that preserves the genomic and regulatory features of their ancestors (Pink et al. 2011).

All these facts suggest that different mechanisms might be responsible for pseudogene transcription. One hypothesis relies on the fact that some pseudogenes are located within other loci and can be expressed by "hitchhiking" on the transcriptional machinery of the genes present there (Vinckenbosch et al. 2006). This mechanism suggests that mostly processed pseudogenes, which do not have upstream regulatory sequences, might become functional by being integrated in regions that favor transcription. This behavior is characteristic of retrogenes which integrate nearby genes, perhaps in the pursuit of the opportunity of being transcribed. An alternative theory proposes that pseudogenes transcription can be modulated at chromatin level. Globally, pseudogenes show the canonical histone modifications from transcribed (H3K4me3 and H3K36me3) and repressed (H3K27me3) protein-coding genes (Pei et al. 2012). However, recent studies described distinct features associated with pseudogenes transcription. First, the histone modification H3K36me3, prevalent in the gene-body of protein-coding genes, appear to be enriched near the transcription start-site of pseudogenes and long non coding RNAs (lncRNAs) (Pei et al. 2012; Sati et al. 2012). Also, expressed pseudogenes show an enrichment of H3K9me3, an histone modification usually associated with repressed regions (Guo et al. 2014). Moving apart from histone modifications, it is also known that DNA methylation can regulate gene activity (Schultz et al. 2015). Indeed, pseudogene expression is tissue-specific repressed by *de novo* methylation after the gene duplication event (Cortese et al. 2008). Also, the high expression of pseudogenes in testis appear to be a consequence of the transient demethylation during spermatogenesis (Grunau et al. 2000). Nevertheless, more studies are need to fully characterize the impact of DNA methylation in the regulation pseudogene expression.

1.3. High-throughput Sequencing Technology (HTS)

It was approximately 20 years between the discovery of DNA's double helix structure by Watson and Crick in 1953 (Watson and Crick 1953) and the first sequencing reactions in the 1970s developed by Sanger et al (Sanger et al. 1977). Due to its impact, Sanger's sequencing method was widely used to determine the DNA sequence of a given location in the genome for another 30 years. Until, in the 2000's, motivated by the need to increase throughput power, we were greeted with more automate and parallel processing technologies that became able to sequence the whole genome (Hattori 2005), culminating in the establishment of next generation sequencing (NGS) technologies. Presently, we find several technologies that present a diversified array of biological applications considering the goals to which these technologies are used.

Examples of NGS technologies are Roche 454, Illumina/Solexa, ABI-SOLiD and Ion Torrent. NGS sequencing protocols can be divided into three major steps as described in Figure 1.5, which are: library preparation, amplification and sequencing (Goodwin et al. 2016). The development of sequencing machines and protocols led to a significant reduction of the sequencing cost when compared to the cost of sequencing a human genome in 2004, meeting the goal to make this technology affordable and accessible (Van Dijk et al. 2014). Illumina is one of the most used sequencing platforms mainly due to their competitive cost, accuracy, and performance. Rapid

advances in the development of sequencing technologies in recent years have enabled an increasing number of applications in biology, namely in the fields of transcriptomics, epigenomics or metabolomics. In this project we used five types of sequencing data which are RNA-seq, GRO-seq, ChIP-seq, WGBS-seq and DNase hypersensitivity. Since the main goal of this project is to investigate how transcription can be affected by epigenetic features in pseudogenes, this data can be further divided into transcriptomic, which includes RNA-seq and GRO-seq, and epigenomic data, including ChIP-seq, WGBS-seq and DNase Hypersensitivity. RNA-seq evaluates the overall amount of polyadenilated RNAs inside the nucleus, therefore determining the relative amount of a given mRNA in the nucleus (Garber et al. 2011). GRO-seq determines active transcription through identifying genes with an engaged RNA polymerase (Core et al. 2008). ChIP-seq assesses the binding sites of proteins that bind to DNA (Mardis 2007). WGBS-seq uses bisulfite conversion to identify methylated cytosines in the genome (Ziller et al. 2015). DNase Hypersensitivy is a method that uses DNase I enzyme to identify regions that are sensitive to cleavage by this enzyme, thus assessing if the DNA found in this region can be accessible (Piper et al. 2013).

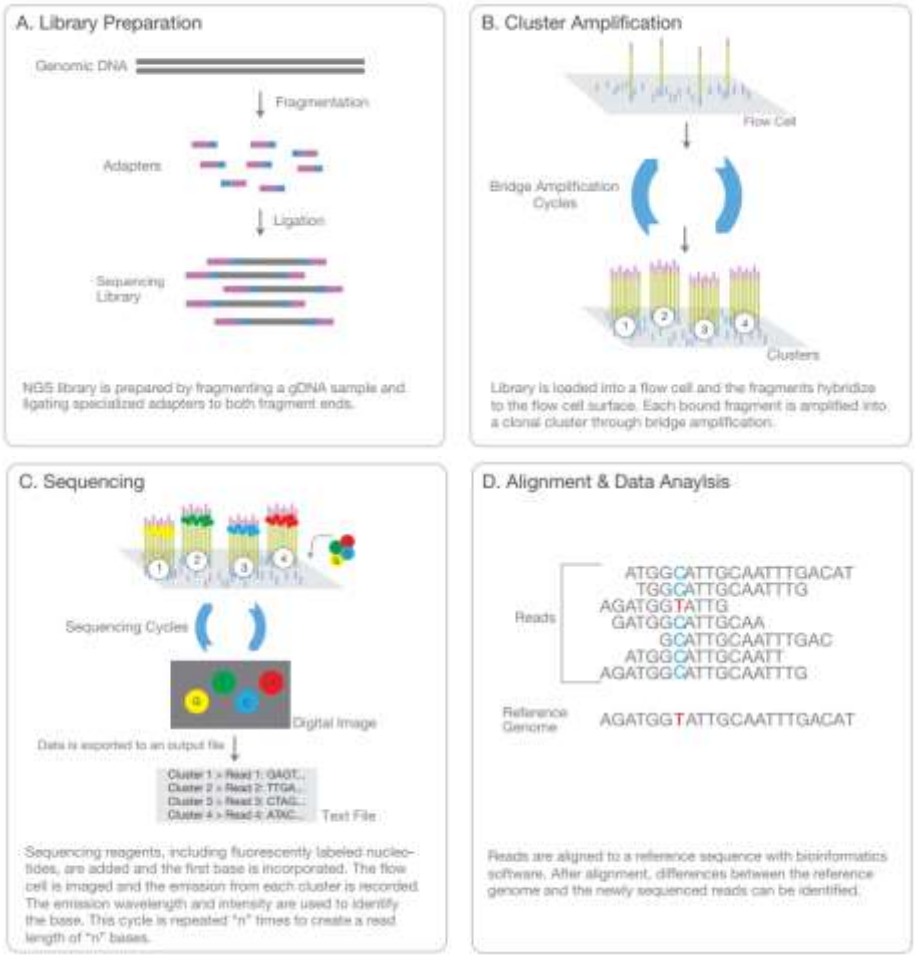

Figure 1.5 – **Illumina's next generation sequencing steps** (adapted from http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf).

1.3.1. The NIH Roadmap Epigenomics Mapping Consortium

The NIH Roadmap Epigenomics Mapping Consortium was developed with the main goal of producing a public database that contained mostly epigenomic sequencing data from normal tissues and cell lines (Bernstein et al. 2010). Presently, it has sequencing information regarding DNA methylation, histone modifications, chromatin accessibility and transcriptome from human cells and tissues, with a total of 127 both human tissues and cell lines. Besides sequencing data, the Consortium also aims to standardise the protocols used for both sequencing and analysis steps in order to allow the scientific community to increase data uniformity, through the usage of the same guidelines. One of the major goals of the consortium is the public dissemination of raw sequence data, processed data and integrated data maps (Romanoski et al. 2015). The NIH Roadmap Epigenomics Consortium is part of a bigger consortium, the International Human Epigenome Consortium (http://ihec-epigenomes.org/), which gathers sequencing information for several others consortiums, like ENCODE (https://www.encodeproject.org/) and Blueprint (http://www.blueprint-epigenome.eu/). Similarly to NIH Roadmap, the International Human Epigenome Consortium has, as main objectives, the generation of reference maps of human epigenomes in health and disease.

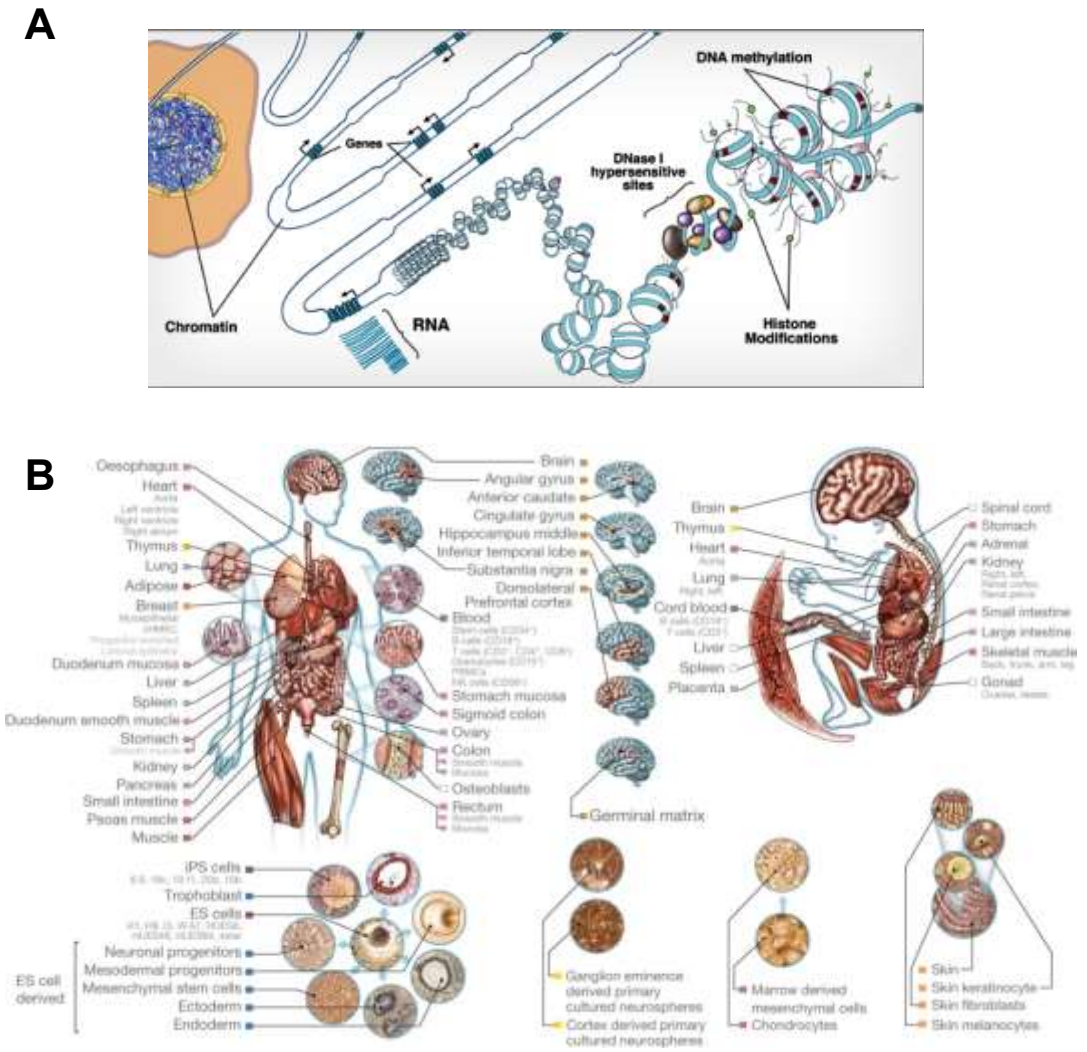

Figure 1.6 – NIH Roadmap Epigenomics Mapping Consortium Data (A) Covered epigenomic information contained in each of the data sets (B) All cell lines and tissues found in this database (adapted from http://www.roadmapepigenomics.org/).

1.4. Background of the thesis and Aims

The development of high-throughput sequencing technologies, their applications and the gerneration of public databases allowed the disclosure of tissue-specific expression of pseudogenes, suggesting a regulated transcription. However, the mechanisms driving pseudogene transcription are incompletely understood.

Thus, the main goal of this thesis was to identify the epigenomic features that coordinate pseudogene transcription. To achieve this goal, a comprehensive analysis was applied assessing chromatin accessibility (DNase Hypersensitivity), DNA methylation (BS-seq), histone modifications (ChIP-seq) and transcriptome (RNA-seq and GRO-seq) of pseudogenes. The genome-wide data was retrieved from the NIH Roadmap Epigenomics Consortium covering 72 samples and 194 replicates. Due to the high expression and role of pseudogenes in brain, we surveyed pseudogene features throughout *in vitro* neural differentiation of human embryonic stem cells (H1) into neuronal progenitor cultured cells (H1N). We also included also H1 derived mesenchymal embryonic cells to understand if the epigenetic patterns in neural differentiation are unique. Hence, through the assessment of an epigenome-wide map, we aim to disclose the epigenomic "active" and "repressive" states ruling pseudogene transcription during neural differentiation.

2. Methods

2.1. Database and Samples

Samples. High-throughput sequencing (HTS) profiles for 194 replicates (corresponding to 72 samples) were produced by the NIH Roadmap Epigenomics Mapping Consortium (Consortium et al. 2015) and the raw data was collected from the GEO database (https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/) (Supplementary Table 1.1). In addition, GRO-seq data for H1 cell line (GEO accession no: GSM1006728) was also included to assess nascent RNA.

Cell Lines. The 3 cell lines chosen for this analysis were human embryonic stem cell lines H1, (H1), H1 derived mesenchymal stem cells (H1M) and H1 derived neuronal progenitor cultured cells (H1N).

Genome Annotation. Gene coordinates were obtained from GENCODE Annotation v23 for GRCh38. This version was chosen over GRCh37 because it contained a higher number of pseudogenes.

2.2. Quality Assessment

All runs were converted from SRA files to FASTQ files and a data quality analysis was performed using FastQC software (Andrews 2015). Due to the large number of samples, a workflow was used to parse the FASTQC output files and filter out the bad quality samples. FastQC outputs several quality measurements being the most important: per base sequence quality; per sequence quality scores; per base sequence content; per sequence GC content and overrepresented sequences. Per base sequence quality defines the range of quality values at each nucleotide position and is usually the measure where most HTS samples show problems. To overcome this issue and recover data, samples were split according to quality criteria and reads from problematic samples were trimmed (Table 2.1).

Table 2.1 – Criteria for Quality Analysis applied to all replicates. All sequencing files labelled as "pass" were automatically included in the analysis. All sequencing files labelled as "check" were trimmed in length according to the first position in which the median value is below 20. "Fail" sequencing files were automatically excluded from the analysis.

	Criteria - FastQC
Pass	Per Base Sequence Quality = "PASS" OR Per Base Sequence Quality = "WARN" OR If Per Base Sequence Quality = "FAIL" Lower Quartile > 20 (for the first 30 positions) AND Median > 20 (for the length of the sequence except last position)
Check	Per Base Sequence Quality = "FAIL" Lower Quartile > 20 (for the first 30 positions) AND Median < 20 (any position except last) Report: 1st position in which Median < 20
Fail	Per Base Sequence Quality = "FAIL" AND Lower Quartile < 20 (any of the 1st 30 positions)

2.3. Genome Mappability

The genome mappability (also known as uniqueness) improves with increased read length and generally shows an inverse correlation with the presence of genomic repeats. Genome mappability is an important feature to determine mapping depth which allows the identification of noncomplex or repetitive regions. It is assessed by fragmentation of the genome in K-mers (Kbps sliding windows of 1 bp step size) and determination of their frequency in the genome (after alignment). Due to lack of mappability tracks for the latest human genome version (GRCh38), we generated mappability tracks using the software GEM (Derrien et al. 2012). GEM software outputs a mappability score for each position of the genome (ranging from 0 to 1) which is calculated based on the fragmentation of genome in K-mers followed by alignment of the generated k-mers. Since the genome mappability varies with the read length, we determined the mappability for K-mers of 36bps and 101 bps (read lengths of the ChIP-seq and RNA-seq data, respectively).

Gene mappability was defined as the sum of the mappability value of each base pair, divided by the total length of the gene. For the epigenomic analysis, only genes with a mappablity higher than 80% were considered.

2.4 Expression Data: RNA-seq and GRO-seq

Transcriptomic data was aligned using Kallisto (Bray et al, 2016) and transcripts per million (TPMs) were calculated. For RNA-seq, TPMs for each gene were defined as the mean of TPMs across replicates for each cell line. All genes with TPMs > 1 were defined as expressed. In order to compare RNA-seq and GRO-seq samples (Figure 3.2, B-D), RPKMs (normalized reads per kilobase per million mapped reads) were obtained using normalized library size, as implemented in edgeR R package (Robinson et al, 2010). Bayesian analysis was applied to determine differentially expressed genes using limma R package (Ritchie et al, 2015) and the following thresholds: B-value > 0 and fold-change > 2.

2.5 Epigenomic Data: ChIP-seq, WGBS-seq, DNase Hypersensitivity

ChIP-seq Data Analysis. To reduce redundancy, improve data quality and achieve uniformity required for our integrative analysis, we decided to trim all ChIP-seq sequences to 36 bp and merge all replicates from the same sample in order to increase the number of reads per sample thus augmenting the coverage of the analysis. Because discrepancies in the number of reads across samples can compromise the quality of the analysis, we decided to subsample all replicates that had more than 20M reads to 20M reads (Consortium et al. 2015). Bowtie software (Langmead et al. 2009) was used to align ChIP-seq replicates to the new reference genome GhRC38, reporting only uniquely mapped reads. PCR duplicates were removed using Picard software (http://picard.sourceforge.net). Histone modification enriched regions (peaks) were identified using MACS2 software with the options –broadpeaks –broadcutoff 0.1. Peaks with minimum FDR value (qvalue) < 0.5 were defined as highly significant peaks and used for the identification of enriched sequences. For quantitative calculation and profiles for all genes, uniquely mapped reads were extended in the 3’ direction to reach 150 nt with the Pyicos (Althammer et al. 2011). Only read counts that overlapped enriched regions identified above were considered.

WGBS-seq Data Analysis. BS-seq reads were mapped to the reference human genome (GRCh38) using Bismark (Krueger and Andrews 2011), that aligns bisulfite converted sequence reads and determines cytosine methylation states.

DNase Hypersensivity Analysis. DNase Hypersensitive reads were aligned similar to the ChIP-seq data. DNase Hypersensitive regions are identified using a peak calling software Homer (Heinz et al. 2010). These regions were subjected to further analysis using the software pyDNase (Piper et al. 2013) for a more accurate identification of DNase Hypersensitive regions

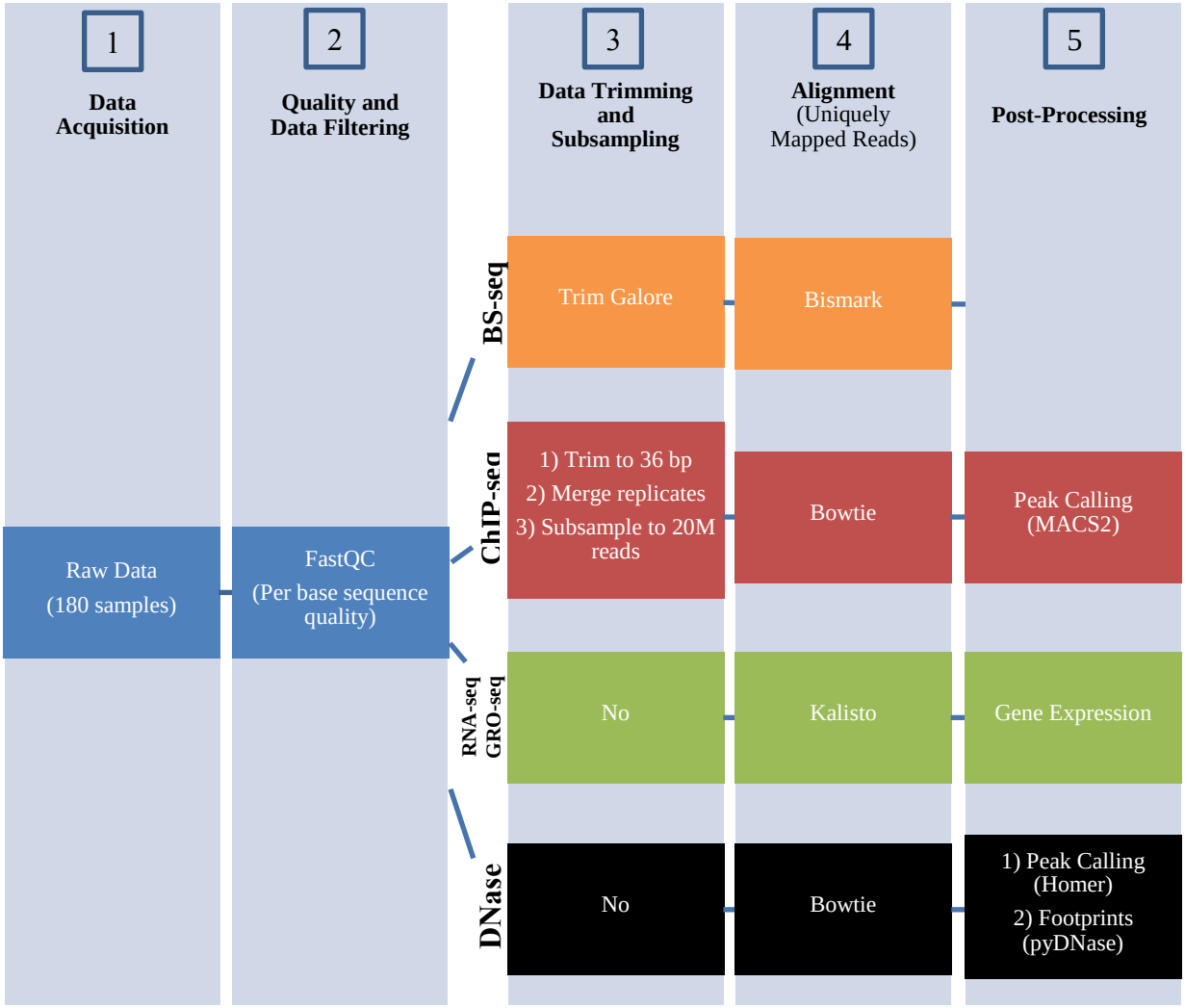

Figure 2.1 – Analysis pipeline according to each type of dataset from NIH Roadmap Epigenomics Data.

ChromHMM. ChromHMM (Ernst and Kellis 2012) was used to assess chromatin states based on histone modifications (ChIP-seq) and chromatin accessibility (DNase-seq) data. Two chromatin states models were inferred as previously described, comprising: 18 states (5 core chromatin marks H3K4me1, H3K4me3, H3K36me3, H3K9me3,H3K27me3 and DNase Hypersensitive regions) and 51 states (18 chromatin marks and DNase Hypersensitive regions). These models were used afterwards to map the epigenomes and for chromatin state enrichment analyses. These downstream analyses were performed using ChromHMM tools to detect chromatin state enrichment in gene regions (Overlap enrichment function) and to infer chromatin state enrichment around anchor positions (TSS) (neighbourhood function).

Other tools. The SAMtools utility for storing large nucleotide sequence alignments and manipulating alignments (Li et al. 2009) and BEDtools software for the comparison of genomic

features (Quinlan and Hall 2010) were used for filtering steps and file format conversion. Finally, processed data was plotted and visualized using software of the R project for statistical computing (Team 2011). For the metagene profile, genes were aligned at the first and last nucleotides of the annotated transcripts and read counts were scaled as follows: the 5' end (2 kb upstream of the transcription start site) and the 3' end (2 kb downstream of the transcription termination site) were unscaled and averaged in a 50-bp window, and the remainder of the gene was scaled to 200 windows using cubic spline interpolation (so that all genes seem to have the same length). Individual profiles were produced using a 50-bp window. All profiles were plotted on a normalized reads per kilobase per million mapped reads (RPKMs). For the epigenomic analysis, we defined a subset of genes considered for the analysis based on the criteria present on Table 2.2.

Table 2.2 – Filtering criteria to define genes used in epigenomic analysis.

Number of Genes After Filter (60498 genes in Gencode Annotation)	**Filter**
14915	Remove overlapping genes (2kb before TSS and after TTS)
8206	Remove genes less than 80% mappable
8152	Remove genes with one alternative intron
8149	Remove genes that are translated pseudogenes or lncRNAs

3. Results

3.1 Public high-throughput sequencing data and quality issues

After downloading the raw data from the NIH Epigenetics Roadmap, we assessed the read quality using FastQC software (Andrews, 2010). Due to the high number of sequencing files (runs) that presented poor quality, mainly due to low per base sequence quality and adapter contamination (Figure 3.1 - panels A and B), we were forced to define a set of criteria in order to rigorously identify which samples could be considered for further analysis, as explained in methods. Files with quality problems were either subjected to post-processing (namely trimming and removing adapter sequences) or discarded from the analysis, in the latest case when the quality check was very poor. As seen in Figure 3.1C, around 40% of all ChIP-seq data was discarded due to poor quality as well as around 30% of BS-seq data. In contrast, around 80% and 90% of DNase Hypersensitivity and RNA-seq data were classified as high quality data sets. Thus, overall around 150 sequencing replicates from NIH Epigenetics Roadmap displayed very low levels of quality and could not be included in the downstream analyses. Due to quality problems and complex post-processing steps in WGBS-seq that require more complex analysis, we decided not to integrate this data in the analysis.

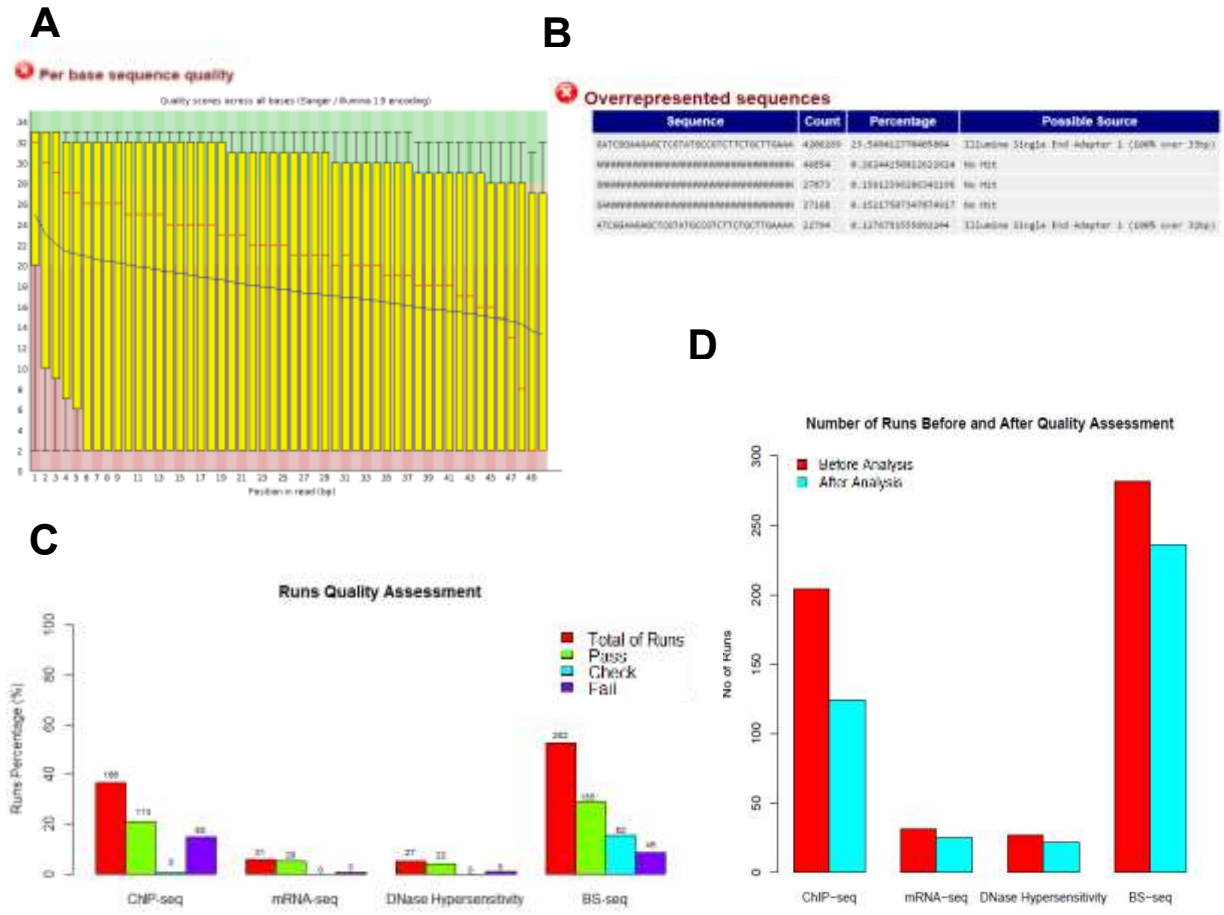

Figure 3.1 Data Quality Analysis for NIH Roadmap Project data (A) Per base sequence quality from FastQC report. In this box whisker it is shown the quality score associated with each base pair from the reads in a representative replicate. The score associated with each base pair is drastically decreasing throughout the length of the read indicating a bad replicate. (B) Overrepresented sequences in FastQC report. Example of replicate with high amount of overrepresented sequences (C) Quality Number of sequencing files (runs) within each dataset (D) Number of sequencing files before and after Quality Assessment for each dataset.

3.2 Pseudogene transcription

To assess pseudogenes expression levels, we used transcriptome data (RNA-seq) from H1, H1N and H1M cell lines. First, we assessed the expression alterations during neural and mesenchymal differentiation to verify if the brain-specific expression of pseudogenes could be established during early development stages (Pei et al. 2012; Guo et al. 2014). Overall, pseudogenes and lincRNAs expression decreases throughout neural and mesenchymal differentiation (Figure 3.2A), which might be explained by the absence of repressive histone modifications in early embryonic stages (Zhu et al. 2013). However, we observed a higher number of pseudogenes being upregulated in neural differentiation relatively to mesenchymal differentiation (Figure 3.2A). The same was also observed for lincRNAs which suggests that not only pseudogenes but other non-coding RNAs might be upregulated in neuronal differentiation. Second, we evaluated the expression levels to define sets of transcribed pseudogenes on each cell type. However, only around 500 pseudogenes could be identified as expressed (TPMs > 1) in H1 (Figure 3.2B). Although, the neural progenitor cell line (H1N) containing the highest number of expressed pseudogenes, only 587 could be detected. Since RNA-seq technology determines steady-state RNA levels (dependent of transcription activity and RNA stability) we decided to assess pseudogenes expression by directly measuring nascent RNA production (GRO-seq).

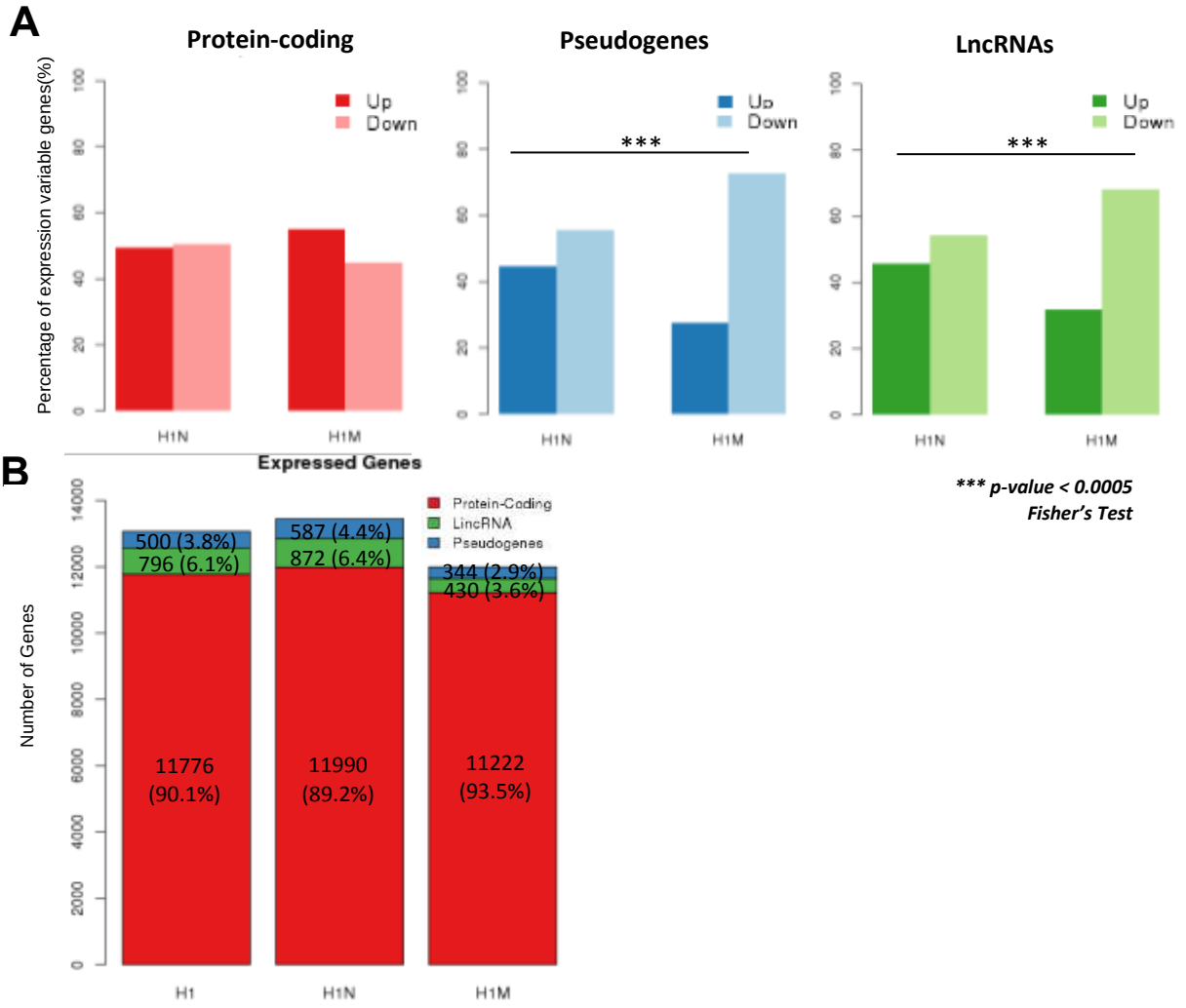

Figure 3.2. Quantification of gene expression in neural (H1-H1N) and mesenchymal differentiation (H1-H1M). (A) Differentially expressed genes according to gene type in neural and mesenchymal differentiation. (B) Expressed genes divided according to gene type in each cell line.

Thus, we assessed the number of expressed genes detected by RNA-seq and GRO-seq in H1 cell line. The overall number of expressed genes was higher in GRO-seq, as observed in panel A from Figure 3.3. However, the proportion of pseudogenes (14.7%) and lincRNAs (14.45%) increased relative to all expressed genes (Fisher's Exact Test p-value < 0.001). Second, we compared the RNA-seq and GRO-seq expression levels (normalized RPKMs) for each gene set (Figure 3.3B). Higher association was obtained for protein-coding genes, whereas pseudogenes showed the lowest correlation value. Moreover, a large fraction of pseudogenes and lincRNAs appear to be actively transcribed but with low final transcript levels, which might suggest a fast degradation of these RNA species. Overall these results are in agreement of previous studies showing general lower RNA stability for pseudogenes and lincRNAs (Thomson and Dinger 2016).

Since we aim to assess the regulatory features of actively transcribed pseudogenes, the expressed genes for downstream analyses were determined using GRO-seq data. However, due to the unavailability of GRO-seq data for H1N and H1M cell lines, expressed genes were determined using RNA-seq data.

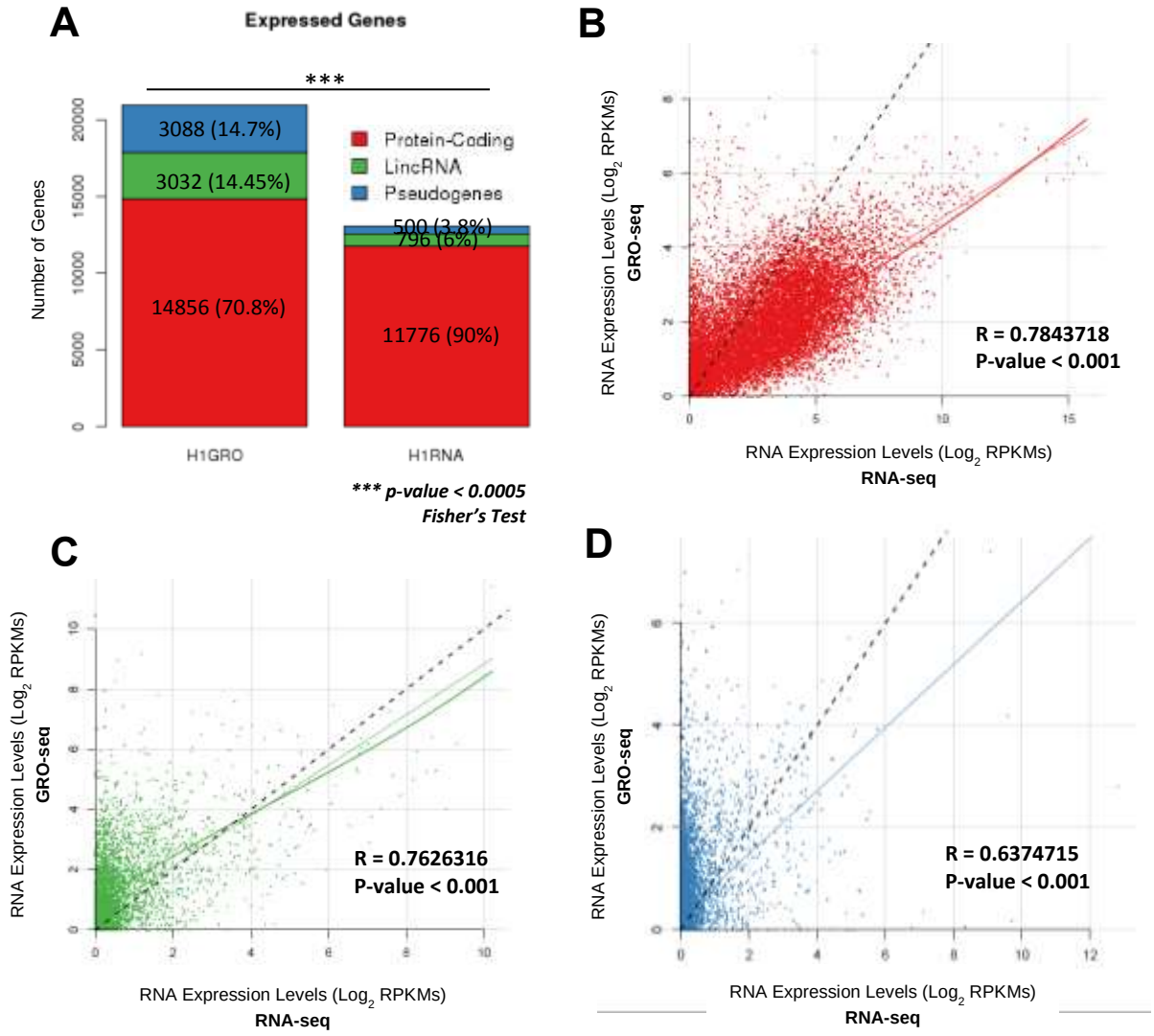

Figure 3.3. Gene transcription defined using GRO-seq and RNA-seq (A) Number of expressed genes divided by gene type defined by GRO-seq or RNA-seq. Fisher's test was performed comparing protein-coding genes with both pseudogenes and lincRNAs in GRO-seq and RNA-seq. (B) (C) (D) Comparison between RPKMs for all protein-coding, lincRNAs and pseudogenes respectively. Protein-coding genes are represented in red, pseudogenes in blue and lncRNA in green. The dashed line represents y = x correlation. Estimated correlation coefficients were obtained using Pearson's correlation.

3.3 Canonical Histone Modifications in Pseudogenes

To understand if transcribed pseudogenes bear the canonical histone modifications associated with transcription (Black et al. 2012), we examined H3K4me3, H3K9me3, H3K27me3 and H3K36me3. However, to guarantee that the effect of these histone modifications is specific to the subset of genes we defined, we decided to filter genes that overlapped with other genes to a minimum of 2 kb before or after the TSS or TTS, respectively. Then, due to the problem of pseudogenes being highly similar to their parental genes, we only selected genes at least 80% mappable. In the final filtering stages, we removed all genes that had an alternative 5'SS splice site and possible translated pseudogenes, as described by Ji et al. (Zhe Ji et al. 2015). These filtering steps are mentioned in Table 2.2 in methods. These genes were then divided according to gene type and expression as seen in Table 3.1.

Table 3.1 – Number of genes divided according to gene type and expression after filtering.

	Protein-Coding Genes	**Pseudogenes**	**LincRNAs**
Initial Genome Count	19815	14505	7674
After filtering	2650	3083	2416
Divided by expression group			
Expressed	1452	201	596
Silent	1198	2882	1820

First, we studied the histone modification H3K4me3, typical of the promoter region in actively transcribed genes. As described before (Pei et al. 2012), pseudogenes presented a subtle peak of H3K4me3 around the TSS, smaller that the promoter mark in lincRNAs and protein-coding genes. Then, we explored the H3K9me3 and H3K27me3 histone modifications, that are defined as repressive marks and correlate inversely with transcription (Pérez-Lluch et al. 2015). Indeed, we observed a notorious enrichment in both H3K27me3 and H3K9me3 for silent protein-coding genes. However, for pseudogenes and lincRNAs, the enrichment of presence of repressive signals was not so obvious. Indeed, expressed pseudogenes appear to have an H3K9me3 enrichment right after the TSS region, as previously described (Guo et al. 2014). The same results were obtained for H1N and H1M (data not shown).

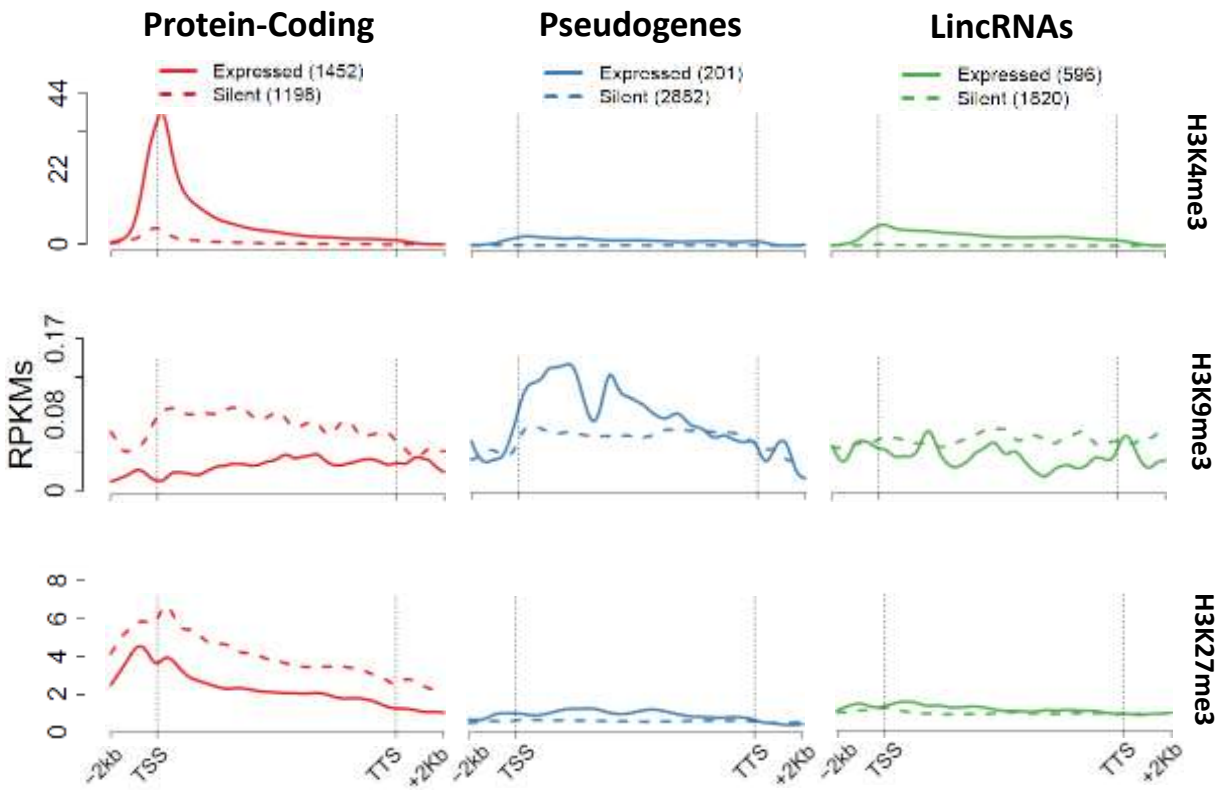

Figure 3.4. Canonical histone modifications distribution in the gene body of expressed (full line) and silent (dashed line) protein-coding genes, pseudogenes and lincRNA in H1.

Second, we explored the H3K36me3 histone modification that is found to be enriched in the body of intron-containing protein-coding genes, correlating with expression levels (de Almeida et al. 2011). Thus, we segregated our genes sets based on the presence and absence of introns. As expected, intron-containing protein-coding genes showed an enrichment of H3K36me3 towards the end of the gene body, as opposed to intronless protein-coding genes (Figure 3.5A). Surprisingly, H3K36me3 levels were higher in intronless pseudogenes. To deeply explore the distribution of H3K36me3 in pseudogenes, we evaluated individual profiles for all the expressed pseudogenes. Notably, some pseudogenes revealed transcription activity upstream and downstream of the annotated region (Figure 3.5B). Indeed, the patterns of nascent transcription suggested that the transcription initiated in the upstream gene *GALTN1* and proceeded throughout the pseudogene. Moreover, the absence of regulatory features (DNase hypersensitive sites and H3K4me3) in the pseudogene promoter supports this hypothesis. Overall, all these results suggest that some pseudogenes are transcribed by "hitchhiking" the transcriptional machinery of the upstream genes. This hypothesis can explain the low levels of the promoter histone mark H3K4me3 observed in pseudogenes.

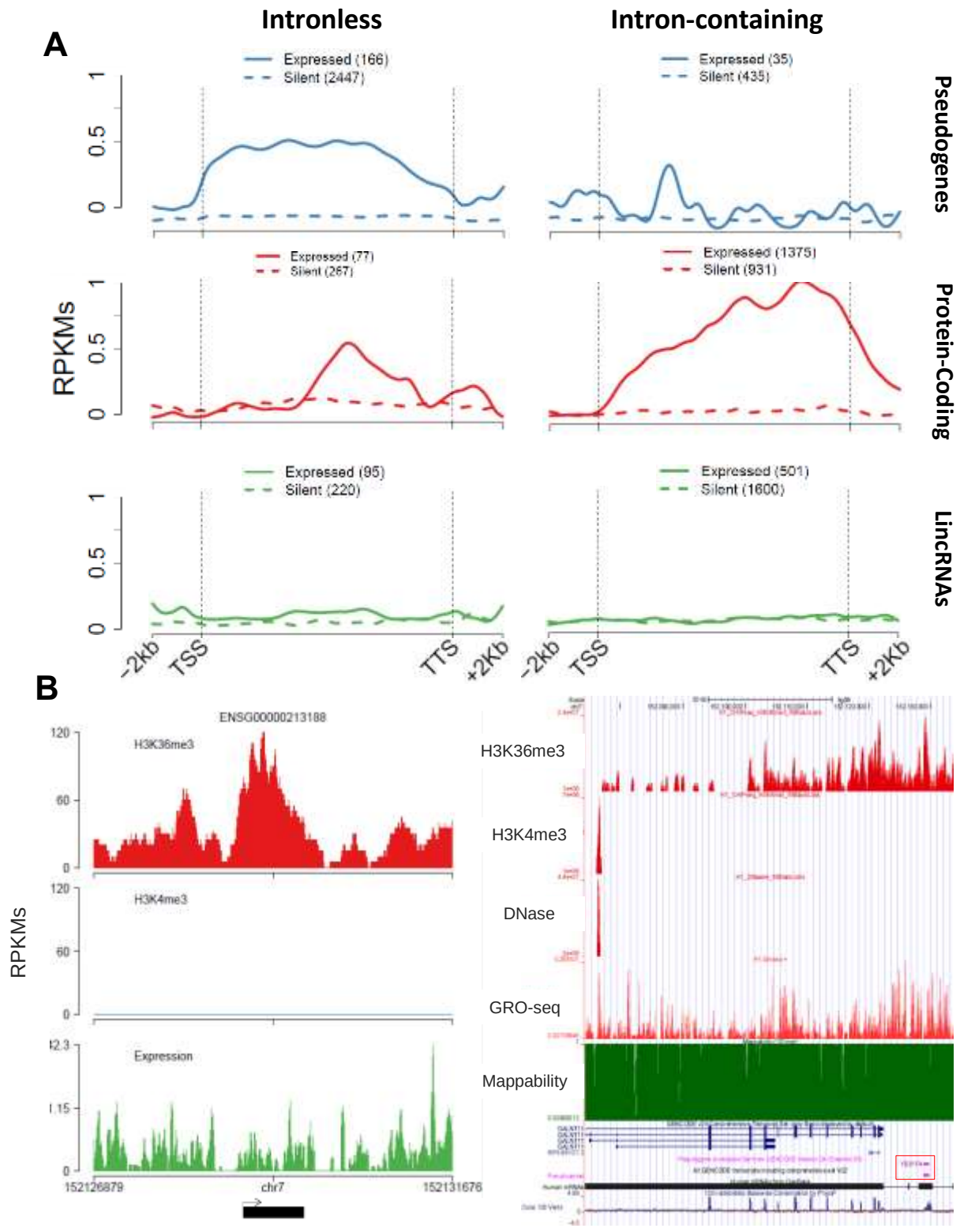

Figure 3.5. H3K36me3 is present in expressed pseudogenes (A) Distribution in the gene body of expressed (full line) and silent (dashed line) protein-coding genes, pseudogenes and lncRNAs (B) Individual profile of H3K36me3, H3K4me3 and expression of an expressed pseudogene (right) and genome browser screenshot of the same pseudogene (left).

3.4. Chromatin States and Dynamics of Pseudogenes

In order to assess the epigenetic features associated with pseudogenes transcription, we used a more complex approach that identifies chromatin states based on a multivariate hidden Markov models (implemented in ChromHMM). The simplest model contained 18-states (defined by five core chromatin marks H3K4me1, H3K4me3, H3K36me3, H3K9me3, H3K27me3 and DNase Hypersensitive regions), revealed an enrichment of the bivalent marks H3K4me3 and H3K27me3 in TSS of expressed pseudogenes (state 10 in H1 and state 4 in H1N) (Figure 3.6). Bivalent chromatin domains were previously associated to developmental genes in embryonic stem cells and to genes expressed at low levels (Bernstein, et al 2006, Cell). Notably, the region close to the TSS also possessed DNase hypersensitive sites (state 15) in H1, not observed for the 200-400nt downstream region of the TSS (state 10). More striking, the chromatin states approach confirmed the enrichment of H3K36me3 and H3K9me3 close to the TSS of expressed pseudogenes for both H1 and H1N (state 4 in H1 and state 14 in H1N). This association was not found for lincRNAs or protein-coding genes. Relative to the silent pseudogenes, an overall enrichment of H3K9me3 was observed for the entire loci and flanking regions. Overall, expressed pseudogenes present a more diversified chromatin arrangement in the transcription initiation regulatory region, when compared to silent pseudogenes, in which the most significant trait is the isolated presence of H3K9me3, as observed in Figure 3.6. Additionally, when compared to protein-coding and lncRNA genes, states attributed to active genes in the three defined classes resemble each other more than when compared to silent genes, in which the enrichment of H3K9me3 appears to be exclusive of pseudogenes.

Finally, we extended the model to all histone modifications available and build a 51-state model (defined by 18 chromatin marks and DNase Hypersensitive regions) (Figures 3.7 and 3.8). Similarly, to the results for the 18-states, expressed pseudogenes were mostly associated with the presence of bivalent marks H3K4me3 and H3K27me3 (states 22 in H1 and 11 in H1N). In addition, we could observe an enrichment of a chromatin state containing subtle levels of H3K4me1 and several histone acetylations (state 15 in H1). Silent pseudogenes were mostly associated to H3K9me3 in both H1 and H1N, coherent to the results from 18-state model.

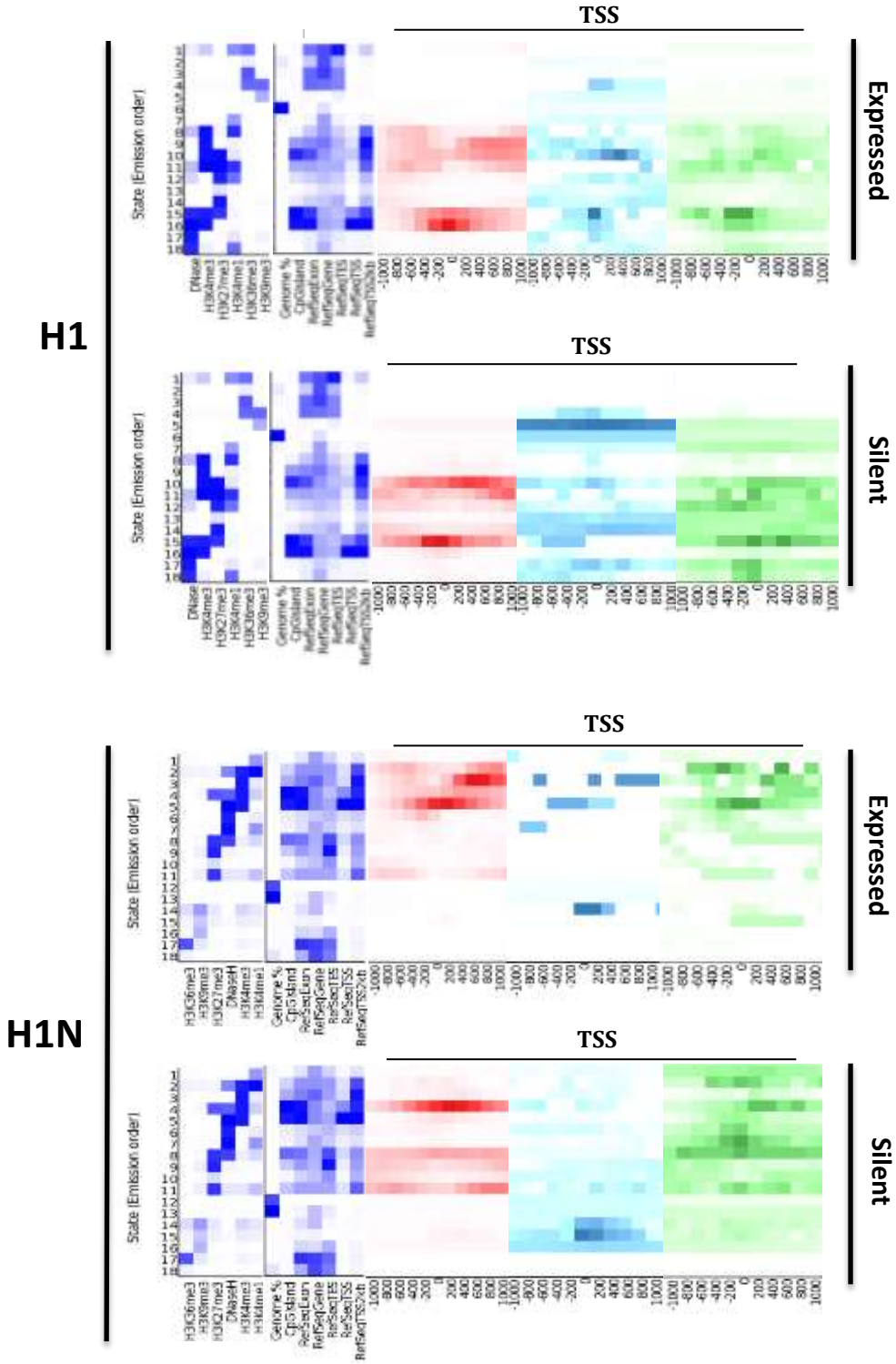

Figure 3.6. 18-state ChromHMM model for expressed genes and silent genes divided according to gene type (protein-coding genes in red, pseudogenes in blue and lncRNAs in green). The first 2 heatmaps are a reference built by the program which associates each histone modification and each genomic regions to a specific state (row), respectively. The following heatmaps depict the overall state enrichment for each of the three gene groups. Heatmap 7-12 represent the overall state enrichment centered on the TSS of all three gene groups.

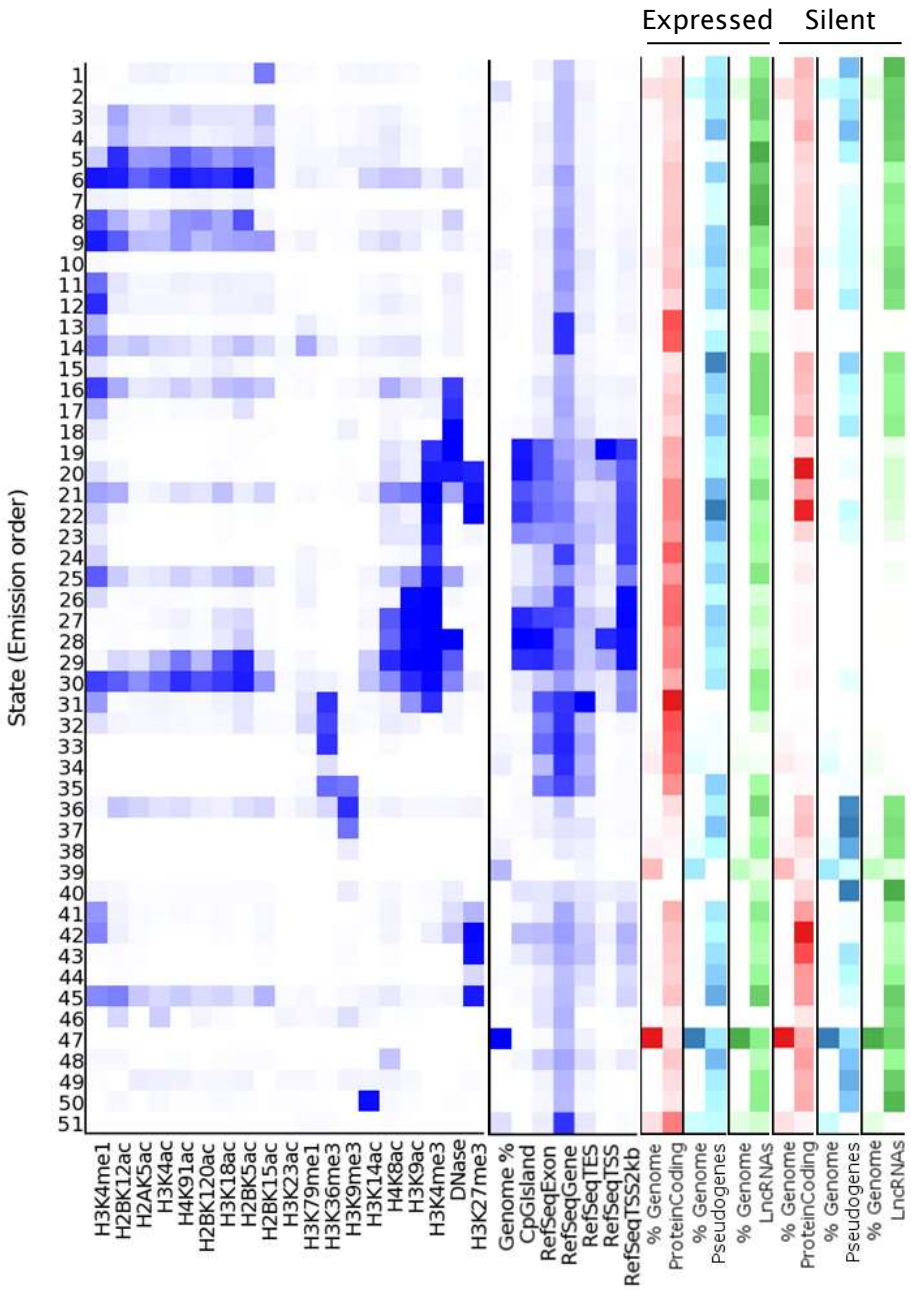

Figure 3.7. 51-state ChromHMM model displaying overall state enrichment for expressed genes and silent genes divided according to gene type (protein-coding genes in red, pseudogenes in blue and lncRNAs in green) and expression level in H1 cell line. The first 2 heatmaps are a reference built by the program which associates each histone modification and each genomic regions to a specific state (row), respectively. The following heatmaps depict the overall state enrichment for each of the three gene groups divided according to expression level.

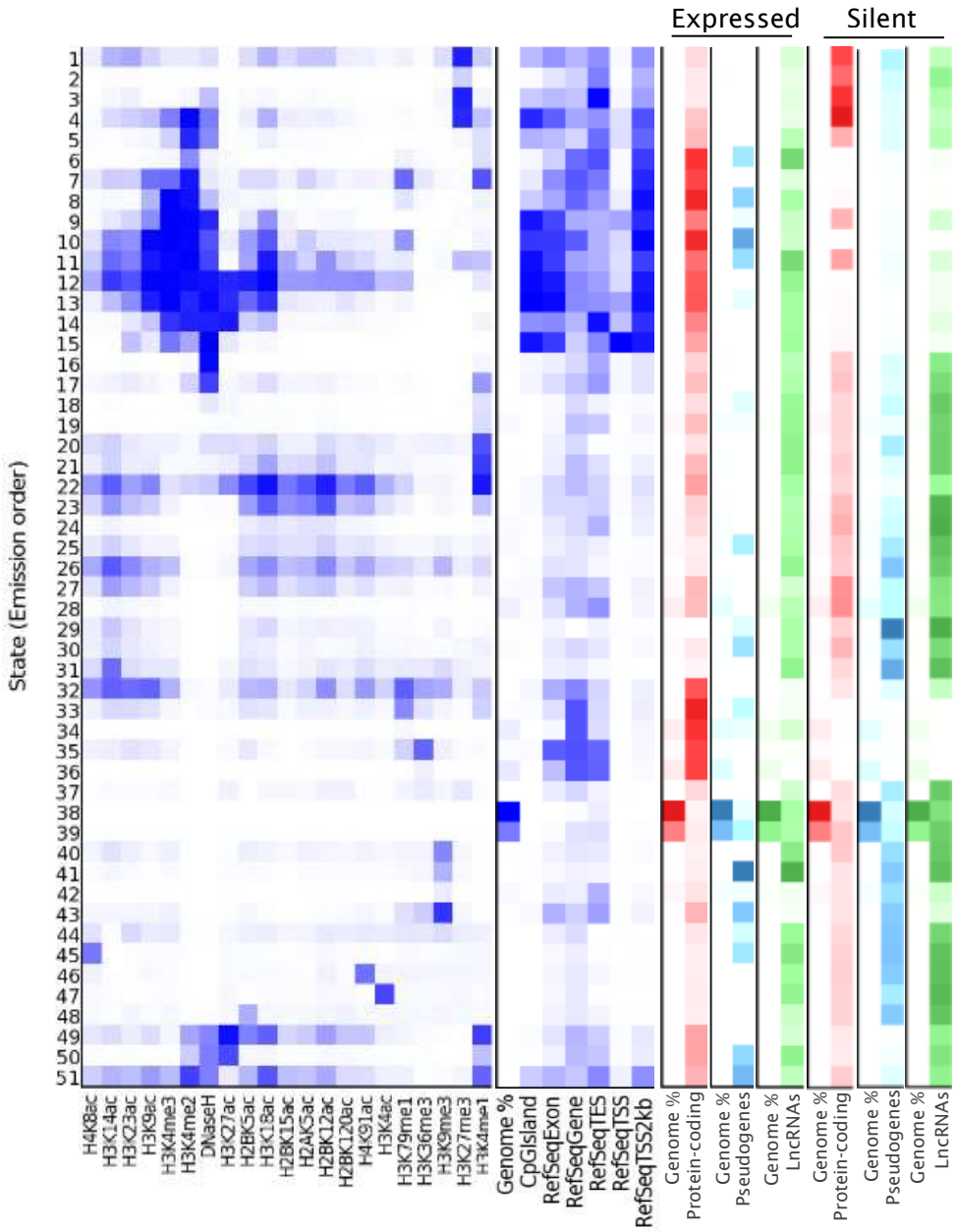

Figure 3.8. 51-state ChromHMM model displaying overall state enrichment for expressed genes and silent genes divided according to gene type (protein-coding genes in red, pseudogenes in blue and lncRNAs in green) and expression level in H1N cell line. The first 2 heatmaps are a reference built by the program which associates each histone modification and each genomic regions to a specific state (row), respectively. The following heatmaps depict the overall state enrichment for each of the three gene groups divided according to expression level.

4. Discussion and Concluding Remarks

Pseudogenes are copies of ancestral protein-coding genes that exhibit evolutionary conservation (Balakirev and Ayala 2003) and that, through the process of pseudogenization, accumulated mutations leading to the loss of coding function. However, recent evidence has shown that pseudogenes are transcribed and their transcription seems to play a fundamental biological role (Ye et al. 2015; Milligan and Lipovich 2015; Kalyana-Sundaram et al. 2012). In this project we aimed to understand which epigenetic mechanisms were responsible for the regulation of pseudogene expression. We characterized pseudogene transcription in neural differentiation and looked at the histone modifications present in expressed and silent pseudogenes.

This project relied mostly on high-throughput sequencing data, a technology that largely pushed the emancipation of the field of bioinformatics. These advancements lead to the exponentially growth of the computational approaches aiming to analyse and simplify the enormous sets of genomic information, which otherwise would not be manually looked through. Throughout the past decade, a lot of effort has been put in the development of new consortiums responsible for the maintenance of public databases that are specialized in the storage of genome-wide datasets, such as the NIH Roadmap Epigenomics Mapping Consortium. However, sometimes public data presents low quality and further steps of the analysis might become compromised. When we started this analysis, we encountered several samples with poor quality, demanding the establishment of quality filters to process the data. Faced with the same problem in the analysis of 111 reference epigenomes (Consortium et al. 2015), the NIH Roadmap Epigenomics Mapping Consortium decided to merge data for the same sample for some datasets in order to improve data coherency, quality and decrease the possibility of noise or contamination in the analysis. This calls the attention to the importance of quality assessment steps when analysing sequencing data and challenges these genome-wide data consortiums to establish more stringent measurements regarding the quality of the available data.

Previous studies have shown that pseudogenes are highly expressed in the brain (Pei et al. 2012) and play important roles in neural differentiation (Scarola et al. 2015; Poursani et al. 2016; Echols et al. 2002). Thus, we decided to explore if this brain-specific expression could be established during early stages of neural differentiation. Our analyses revealed higher number of pseudogenes transcripts in the neural progenitor cells (H1N) cell line and an increased expression of pseudogenes during neural differentiation. However, lncRNAs also seem to be strongly up-regulated, suggesting also an important role in neuronal development, which has also been described (Ng et al. 2012).

The role of epigenetics in the regulation of gene expression is of great relevance in normal protein-coding genes. Canonical histone modifications are associated with specific patterns and correlate either positively or negatively with transcription activity (Barth and Imhof 2010). In this project our main goal was to assess and understand the importance of epigenetic mechanisms in the regulation of pseudogene transcription. In order to evaluate if pseudogene transcription can be regulated through specific histone modifications, we looked at four canonical histones modifications which are known to be associated with transcription regulation (Black et al. 2012): H3K36me3, H3K4me3, H3K9me3 and H3K27me3. As described previously (Pei et al. 2012), H3K4me3 seems to be present in expressed pseudogenes, displaying the same profile as in protein-coding genes. However, expressed pseudogenes showed a strong association with bivalent chromatin domains (H3K4me3 and H3K27me3). The presence of these bivalent domains in expressed pseudogenes further contributes to the hypothesis that pseudogene transcription is indeed regulated and that can play important roles in differentiation steps, since the existence of a bivalent promoter opens the possibility for the pseudogene to be transcribed, similarly to many protein-coding genes. Additionally, expressed pseudogenes also showed a more bizarre behaviour for H3K36me3 and H3K9me3 profiles. H3K36me3 is known to be present throughout the gene body of intron-containing protein-coding

genes, in a gradual enrichment towards the transcription termination site in a transcription-dependent manner (de Almeida et al. 2011). Intriguingly, in pseudogenes had been described an enrichment near the transcription initiation site (Pei et al. 2012) as well as for lncRNAs (Sati et al. 2012) but in none of these analyses the metagene profiles were normalized for gene size. Due to the smaller size of pseudogenes when compared to protein-coding genes, the apparent TSS enrichment may belong to terminal region of the pseudogenes. Nonetheless, when we plotted the length-normalized profile of H3K36me3 for expressed pseudogenes, we observed a supplementation of this histone modification through the length of the pseudogene body, especially for intronless pseudogenes. Intrigued, we looked individually for some pseudogenes using the genome browser (Kent et al. 2002) and found that pseudogenes seem to be transcribed as consequence of transcription read-through. Accordingly, these pseudogenes showed higher level of H3K36me3, similar to last exons of intron-containing genes. This expression could have two explanations: 1) as described before (Vinckenbosch et al. 2006), pseudogenes could be inserted in other loci or contributing, for example, as a 3' terminal exon of the gene upstream, therefore taking advantage of the transcription machinery of these loci; 2) annotation problems could result in the wrong classification of these pseudogenes, that could be, in fact, other genomic elements. Besides H3K36me3, also H3K9me3 seems to have an enrichment in both expressed and silent pseudogenes. This association has been previously described in a transcription dependent manner and more intensely in pseudogenes that produced small-RNAs (Guo et al. 2014). A similar relationship was observed in expressed lncRNAs (Melé et al. 2016). These findings suggest that small-RNAs derived from pseudogene transcription might modulate the repression of pseudogene transcription. Another explanation may be related to the cell heterogeneity from genome-wide profiles produced from bulk cell population. In this case, pseudogenes could be actively transcribed in some cells and repressed in others. One way to test this would be to use single cell technology which could allow us to relate expression and epigenomic data from the same cell (Angermueller et al. 2016). In addition, one cannot discard the possibility that H3K9me3 acts repressively in the non-transcribed allele. Further analysis of histone modifications using more complex approaches are necessary to identify chromatin states identified an association between expressed pseudogenes and chromatin features.

Additional work is needed, namely to deeply characterize the pseudogenes promoter regions according to the presence of other regulatory features, namely transcription factor binding sites. Overall, the transcription activation signals are located around the TSS, sometimes invading the first exon and introns. Since processed pseudogenes arise from retrotransposition events, the transcription of pseudogenes could depend on the presence of the regulatory regions downstream of the TSS. Also, the role of DNA methylation in the regulation of pseudogene expression is still in need of clarification and analysis, however there is the indication that this modification might have a very important role in pseudogene transcription (Grunau et al. 2000).

The study of pseudogenes presents many challenges and limitations. First, mapping of high-throughput sequencing data in pseudogenes genomic regions is technically demanding. The fact that pseudogenes sequences are very similar to their parental genes can produce valid alignments where the reads from the parental gene are mistakenly aligned to the pseudogene region. In this project, we bypassed this problem by accepting the reads with unique alignments (only one equal sequence in the genome) and we filtered out pseudogenes with low mappability. These strict filters reduce drastically the number of feasible pseudogenes available for the study and limited the downstream analyses. One way to overcome this could be by the generation of longer reads in genome-wide technologies. Second, considering that pseudogenes resemble their parental genes and therefore can act as microRNA sponges which are consequently degraded by the RNA-induced silencing complex (RISC) (Thomson and Dinger 2016), using RNA-seq data to determine pseudogene expression might lead to an underestimation of the transcription levels of pseudogenes. Indeed, when we compared steady-state

(RNA-seq) and the nascent transcript levels (GRO-seq), we observed higher number and expression levels of pseudogenes in the active transcription dataset. Thus, the study of low expressed and unstable transcripts derived from repeated genomic regions demands new advances in library protocols and sequencing technologies. Finally, during the current work we could identify pseudogenes that are being transcribed by transcription read-through that continues beyond the TTS of the upstream gene. This observation could highlight the existence of annotation issues even in the latest database versions, such as Gencode v23, claiming the need of throughout effort to deeply define the genome.

Pseudogene tissue specific expression (Kalyana-Sundaram et al. 2012) puts forward the idea that, indeed, pseudogenes are important in the epigenetic regulation of other genes themselves contracting the "pseudo-" suffix in pseudogenes, at least functionally. Although this contribution seems to be clear and this work has tried to confirm these observations, the epigenetic regulation of pseudogene transcription still needs further clarification in order to fully understand how epigenetic mechanisms can contribute to the regulation of pseudogenes.

5. References

Allis CD, Jenuwein T. 2016. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **1**: 1–14.

Andrews S. 2015. FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Inst*.

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* **13**: 229–32.

Balakirev ES, Ayala FJ. 2003. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37**: 123–51.

Barth TK, Imhof A. 2010. Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem Sci* **35**: 618–626.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045–8.

Black JC, Van Rechem C, Whetstine JR. 2012. Histone Lysine Methylation Dynamics: Establishment, Regulation, and Biological Impact. *Mol Cell* **48**: 491–507.

Cheng X, Blumenthal RM. 2008. Mammalian DNA Methyltransferases: A Structural Perspective. *Structure* **16**: 341–350.

Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**: 1845–8.

Cortese R, Krispin M, Weiss G, Berlin K, Eckhardt F. 2008. DNA methylation profiling of pseudogene-parental gene pairs and two gene families. *Genomics* **91**: 492–502.

de Almeida SF, Grosso AR, Koch F, Fenouil R, Carvalho S, Andrade J, Levezinho H, Gut M, Eick D, Gut I, et al. 2011. Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nat Struct Mol Biol* **18**: 977–983.

Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* **7**.

Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein MB. 2002. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* **30**: 2515–23.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin state discovery and characterization. *Nat Methods* **9**: 215–216.

Fillingham J, Keogh M-C, Krogan NJ. 2006. γH2AX and its role in DNA double-strand break repair. *Biochem Cell Biol* **84**: 490–494.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**: 469–477.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation

sequencing technologies. *Nat Rev Genet* **17**: 333–351.

Gregório LK. 2016. Publisher main menu Not so pseudo : the evolutionary history of protein phosphatase 1 regulatory subunit 2 and related pseudogenes. 1–14.

Groen JN, Capraro D, Morris K V. 2014. The emerging role of pseudogene expressed non-coding RNAs in cellular functions. *Int J Biochem Cell Biol* **54**: 350–355.

Grunau C, Hindermann W, Rosenthal a. 2000. Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes. *Hum Mol Genet* **9**: 2651–2663.

Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. 2014. Characterization of human pseudogene-derived non-coding RNAs for functional potential. *PLoS One* **9**.

Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, Diao L, Xu Y, Verhaak RGW, Liang H. 2014. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun* **5**: 3963.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res* **33**: 2374–2383.

Hattori M. 2005. Finishing the euchromatic sequence of the human genome. *Tanpakushitsu Kakusan Koso* **50**: 162–168.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.

Holliday R, Pugh JE. 1975. DNA modification mechanisms and gene activity during development. *Science (80-)* **187**: 226–232.

Jacq C, Miller JR, Brownlee GG. 1977. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell* **12**: 109–120.

Johnsson P, Ackley A, Vidarsdottir L, Lui W-O, Corcoran M, Grandér D, Morris K V. 2013. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nat Struct Mol Biol* **20**: 440–6.

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani I a., Kothari V, Prensner JR, Lonigro RJ, et al. 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149**: 1622–1634.

Kandouz M, Bier A, Carystinos GD, Alaoui-Jamali M a, Batist G. 2004. Connexin43 pseudogene is expressed in tumor cells and inhibits growth. *Oncogene* **23**: 4763–4770.

Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. 2010. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**: 2926–2931.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Kornberg RD, Lorch Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**: 285–294.

Kornblihtt AR, Schor IE, Allo M, Blencowe BJ. 2009. When chromatin meets splicing. *Nat Struct Mol Biol* **16**: 902–903.

Kouzarides T. 2007. Chromatin Modifications and Their Function. *Cell* **128**: 693–705.

Krueger F, Andrews SR. 2011. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Law JA, Jacobsen SE. 2010. Establishing , maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**: 204–220.

Li B, Carey M, Workman JL. 2007. The Role of Chromatin during Transcription. *Cell* **128**: 707–719.

Liedtke S, Enczmann J, Waclawczyk S, Wernet P, Kögler G. 2007. Oct4 and Its Pseudogenes Confuse Stem Cell Research. *Cell Stem Cell* **1**: 364–366.

Liu Y-J, Zheng D, Balasubramanian S, Carriero N, Khurana E, Robilotto R, Gerstein MB. 2009. Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity. *BMC Genomics* **10**: 480.

Mardis ER. 2007. ChIP-seq: welcome to the new frontier. *Nat Methods* **4**: 613–4.

Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. 2016. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res* 1–11.

Milligan MJ, Harvey E, Yu A, Morgan AL, Smith DL, Zhang E, Berengut J, Sivananthan J, Subramaniam R, Skoric A, et al. 2016. Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. *Front Genet* **7**: 26.

Milligan MJ, Lipovich L. 2015. Pseudogene-derived lncRNAs: Emerging regulators of gene expression. *Front Genet* **6**: 1–7.

Ng S-Y, Johnson R, Stanton LW. 2012. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J* **31**: 522–33.

Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N. 2003. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* **4**: R74.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu X, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biol* **13**: R51.

Pérez-Lluch S, Blanco E, Tilgner H, Curado J, Ruiz-Romero M, Corominas M, Guigó R. 2015. Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat Genet* **47**: 1158–67.

Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Raul D, Carter F. 2011. Pseudogenes : Pseudo-functional or key regulators in health and disease. 792–798.

Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. 2013. Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res* **41**.

Poliseno L. 2012. Pseudogenes: newly discovered players in human cancer. *Sci Signal* **5**: re5.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**: 1033–8.

Poursani EM, Mohammad Soltani B, Mowla SJ. 2016. Differential Expression of OCT4 Pseudogenes in Pluripotent and Tumor Cell Lines. *Cell J* **18**: 28–36.

Puget N, Gad S, Perrin-Vidoz L, Sinilnikova OM, Stoppa-Lyonnet D, Lenoir GM, Mazoyer S. 2002. Distinct BRCA1 Rearrangements Involving the BRCA1 Pseudogene Suggest the Existence of a Recombination Hot Spot. *Am J Hum Genet* **70**: 858–865.

Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. 2015. Epigenomics: Roadmap for regulation. *Nature* **518**: 314–316.

Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T. 2007. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene* **389**: 196–203.

Sanger F, Nicklen S, Coulson a R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463–7.

Sati S, Ghosh S, Jain V, Scaria V, Sengupta S. 2012. Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. *Nucleic Acids Res* **40**: 10018–10031.

Scarola M, Comisso E, Pascolo R, Chiaradia R, Maria Marion R, Schneider C, Blasco M a, Schoeftner S, Benetti R. 2015. Epigenetic silencing of Oct4 by a complex containing SUV39H1 and Oct4 pseudogene lncRNA. *Nat Commun* **6**: 7631.

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**: 212–216.

Smith ZD, Meissner A. 2013. REVIEWS DNA methylation : roles in mammalian development. *Nat Rev Genet* **14**: 204–220.

Soboleva TA, Nekrasov M, Ryan DP, Tremethick DJ. 2014. Histone variants at the transcription start-site. *Trends Genet* **30**: 199–208.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465–76.

Svensson Ö, Arvestad L, Lagergren J. 2006. Genome-wide survey for biologically functional pseudogenes. *PLoS Comput Biol* **2**: 358–369.

Thomson DW, Dinger ME. 2016. Endogenous microRNA sponges: evidence and controversy. *Nat Rev Genet* **17**: 272–283.

Tonner P, Srinivasasainagendra V, Zhang S, Zhi D. 2012. Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics* **13**: 412. BMC Genomics.

Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. *Genome Res* **13**: 2559–2567.

Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet* **30**.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* **103**: 3220–5.

Voigt P, Tee WW, Reinberg D. 2013. A double take on bivalent promoters. *Genes Dev* **27**: 1318–1338.

Waddington CH. 1942a. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150**: 563–565.

Waddington CH. 1942b. The Epigenotype. *Endeavour* 18–20.

Wang J, Pitarque M, Ingelman-Sundberg M. 2006. 3′-UTR polymorphism in the human CYP2A6 gene affects mRNA stability and enzyme expression. *Biochem Biophys Res Commun* **340**: 491–497.

Watson JD, Crick FHC. 1953. Molecular structure of nucleic acids. *Nature* **171**: 737–738.

Ye X, Fan F, Bhattacharya R, Bellister S, Boulbes DR, Wang R, Xia L, Ivan C, Zheng X, Calin GA, et al. 2015. VEGFR-1 Pseudogene Expression and Regulatory Function in Human Colorectal Cancer Cells. *Mol Cancer Res* **13**: 1274–1282.

Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13**: 2541–2558.

Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* **11**: R26.

Zhe Ji, Ruisheng Song, Aviv Regev KS. 2015. Kevin Struhl Many lncRNAs , 5 ' UTRs , and pseudogenes are translated and some are likely to express functional proteins. Department of Biological Chemistry and Molecular Pharmacology , Harvard Medical Howard Hughes Medical Institute , Department of Biolog.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**: 7–18.

Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, et al. 2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**: 642–654.

Ziller MJ, Hansen KD, Meissner A, Aryee MJ. 2015. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* **12**: 230–2, 1 p following 232.

6. Supplementary Material

Table 6.1. All replicates initially processed for this analysis identified by the respective cell line, library type and mark (for ChIP-seq data). The GEO IDs to assess all of these samples are GSE16256, GSE16368 and GSE18927.

Cell Line	Library type	Chromatin Modification	Number of Replicates
H1	BS-seq	-	5
H1N	BS-seq	-	5
H1M	BS-seq	-	2
H1	ChIP-seq Input	-	13
H1M	ChIP-seq Input	-	1
H1N	ChIP-seq Input	-	4
H1	ChIP-seq	H2AK5ac	2
H1M	ChIP-seq	H2AK5ac	2
H1N	ChIP-seq	H2AK5ac	2
H1	ChIP-seq	H2BK120ac	3
H1M	ChIP-seq	H2BK120ac	1
H1N	ChIP-seq	H2BK120ac	2
H1	ChIP-seq	H2BK12ac	2
H1M	ChIP-seq	H2BK12ac	2
H1N	ChIP-seq	H2BK12ac	1
H1	ChIP-seq	H2BK15ac	2
H1M	ChIP-seq	H2BK15ac	1
H1N	ChIP-seq	H2BK15ac	2
H1	ChIP-seq	H2BK5ac	2
H1M	ChIP-seq	H2BK5ac	2
H1N	ChIP-seq	H2BK5ac	1
H1	ChIP-seq	H3K14ac	2
H1M	ChIP-seq	H3K14ac	2
H1N	ChIP-seq	H3K14ac	1
H1	ChIP-seq	H3K18ac	2
H1M	ChIP-seq	H3K18ac	2
H1N	ChIP-seq	H3K18ac	2
H1	ChIP-seq	H3K23ac	2
H1M	ChIP-seq	H3K23ac	2
H1N	ChIP-seq	H3K23ac	3
H1	ChIP-seq	H3K27ac	2
H1M	ChIP-seq	H3K27ac	2
H1N	ChIP-seq	H3K27ac	5
H1	ChIP-seq	H3K27me3	7
H1M	ChIP-seq	H3K27me3	2
H1N	ChIP-seq	H3K27me3	4
H1	ChIP-seq	H3K36me3	7
H1M	ChIP-seq	H3K36me3	2

H1N	ChIP-seq	H3K36me3	3
H1	ChIP-seq	H3K4ac	2
H1M	ChIP-seq	H3K4ac	2
H1N	ChIP-seq	H3K4ac	2
H1	ChIP-seq	H3K4me1	6
H1M	ChIP-seq	H3K4me1	2
H1N	ChIP-seq	H3K4me1	3
H1	ChIP-seq	H3K4me2	2
H1M	ChIP-seq	H3K4me2	2
H1N	ChIP-seq	H3K4me2	1
H1	ChIP-seq	H3K4me3	8
H1M	ChIP-seq	H3K4me3	2
H1N	ChIP-seq	H3K4me3	4
H1	ChIP-seq	H3K79me1	3
H1M	ChIP-seq	H3K79me1	2
H1N	ChIP-seq	H3K79me1	1
H1	ChIP-seq	H3K9ac	5
H1M	ChIP-seq	H3K9ac	2
H1N	ChIP-seq	H3K9ac	1
H1	ChIP-seq	H3K9me3	7
H1M	ChIP-seq	H3K9me3	2
H1N	ChIP-seq	H3K9me3	3
H1	ChIP-seq	H4K8ac	2
H1M	ChIP-seq	H4K8ac	2
H1N	ChIP-seq	H4K8ac	1
H1	ChIP-seq	H4K91ac	2
H1M	ChIP-seq	H4K91ac	1
H1N	ChIP-seq	H4K91ac	1
H1	DNase	-	2
H1N	DNase	-	2
H1M	DNase	-	2
H1	mRNA-seq	-	4
H1M	mRNA-seq	-	2
H1N	mRNA-seq	-	2