UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Assessing Pseudogene Expression During Neural Differentiation by RNA-Seq

Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

Luís Carlos Pereira Simões

Dissertação orientada por:
Doutora Ana Rita Fialho Grosso
Instituto de Medicina Molecular

Professor Doutor Francisco José Moreira Couto
Faculdade de Ciências da Universidade de Lisboa

2015

Agradecimentos

À Doutora Ana Rita Fialho Grosso, não só por me ter orientado de forma excelente no meu trabalho, o que poderá ter sido uma tarefa bastante difícil, mas também por todo o suporte e disponibilidade quer a nível profissional, quer a nível pessoal, que me foi dando ao longo do tempo, o que ficará para sempre marcado na minha mente. Agradeço também pela inspiração profissional que será muito útil no futuro. Ficarei para sempre grato por tudo!

Ao Doutor Sérgio Almeida, Principal Investigador no Instituto de Medicina Molecular, por me abrir a porta do seu grupo de investigação e sempre se ter mostrado disponível e interessado, a todos os níveis. Também por me ter proporcionado conhecer todo um grupo de pessoas fantásticas, o que para mim se complementa com o excelente grupo de investigação a que pertencem e às quais não posso deixar de agradecer pela preocupação que tiveram comigo ao longo deste ano da minha vida e por todo o apoio que me deram. Muito obrigado a todos!

Ao Doutor Nuno Morais por ser uma inspiração a nível pessoal, profissional e desportivo e por sempre se ter mostrado prestável em todos estes campos. Agradeço também pelo espaço e material que me facultou ao longo do tempo. Muito obrigado.

À Doutora Ana Pombo do Max Delbrück Center for Molecular Medicine (Berlim), pela disponibilização dos seus dados para as minhas análises.

Aos meus verdadeiros amigos, por se terem cruzado no meu caminho e terem também vocês moldado um pouco (ou bastante) a pessoa que sou. Por terem acreditado em mim quando eu mesmo duvidei. Por terem sido sinceros ao longo do tempo, mesmo quando isso me magoou. Por todos os momentos de festa, em que todos os problemas e preocupações se apagaram das nossas mentes. Se continuasse aqui a escrever motivos pelos quais agradecer nunca mais parava. Se não fossem vocês e o vosso apoio, todo este ano teria sido muito mais complicado.

Por último, mas não em último, quero não só agradecer mas também demonstrar todo o meu amor e carinho pela minha família, pois esta representa os alicerces de tudo o que construí até agora. Em especial aos meus pais, pois sem vocês nunca estaria aqui e nunca teria tido tantas oportunidades na vida, fruto de muito sacrifício e suor constantes da vossa parte, desde que me lembro de ser gente. Agradeço também toda a dedicação investida na minha educação, e quando falo em educação não tenho em vista apenas o meu percurso académico, mas sim todos os ensinamentos e competências sociais que me foram transmitindo até aos dias de hoje. Eu sei que a vida não tem sido fácil, que nem tudo tem corrido como planeado, mas acredito que depois da tempestade vem a bonança, e se alguém a merece mais que tudo, são vocês. Obrigado por sempre terem lutado por me tornarem o melhor ser humano possível, mesmo quando eu não o quis, mesmo quando não compreendi os vossos pontos de vista, mesmo quando não aceitei os vossos conselhos, mesmo quando me revoltei por achar que algo não fazia sentido… Sei que tudo o que fizeram até hoje foi pelo meu bem, mas a evolução advém da diversidade, e isto não só é verdade para a biologia, como para os nossos relacionamentos interpessoais. As nossas diferenças fizeram-me evoluir e continuar no caminho que percorro, acreditando que é o correto. Nunca vos conseguirei agradecer nem retribuir tudo o que fizeram por mim. Resta-me só desejar que vos orgulhe e que isso tenha maior valor que um mero “obrigado”!

Resumo

Os pseudogenes são sequências genómicas que foram desprezadas ao longo do tempo, por se pensar que não passavam de réplicas ancestrais de genes codificantes de proteína. Esta visão tem sido desmistificada nos últimos anos e vários estudos têm vindo a surgir mostrando que os pseudogenes não são apenas meras cópias disfuncionais dos genes codificantes de proteína, pois também eles desempenham funções biológicas relevantes.

Existem 14,285 pseudogenes anotados no genoma humano pelo projeto GENCODE (versão 22, Outubro de 2014). O número tem vindo a aumentar ao longo dos anos devido ao desenvolvimento das tecnologias de sequenciação de nova geração (*next generation sequencing* - NGS) e de algoritmos para identificação de novos pseudogenes. No genoma de ratinho encontram-se anotados 8,526 pseudogenes pelo projeto GENCODE (versão M5, Dezembro de 2014).

A origem destas sequências que derivam de genes codificantes de proteína (genes parentais) dá-se na sua grande maioria por um de dois mecanismos de pseudogenização: duplicação genómica de um *locus* do gene parental (pseudogene duplicado ou não-processado); ou retrotransposição (pseudogene processado), onde um mRNA é reversamente transcrito em DNA de novo e inserido aleatoriamente no genoma. Este último é o processo pelo qual a maioria dos pseudogenes são originados, levando esta classe a ser a mais estudada. Inicialmente estas sequências podem ser operacionais, mas ao longo do tempo acumulam mutações deletérias que podem resultar na tradução de um codão *stop* prematuro, ou em mutações *frameshift* que levam a uma mudança na grelha de leitura, impedindo a expressão bem sucedida destas sequências. Existe ainda uma terceira classe de pseudogenes, os unitários, que não resultam de nenhum tipo de inserções genómicas, apenas de mutações pontuais, levando assim a que se torne um "gene vestigial".

O primeiro pseudogene foi descoberto em 1977, coincidindo com o aparecimento da primeira técnica de sequenciação. O desenvolvimento das tecnologias de NGS, bem como a redução de custo das mesmas, têm permitido que cada vez mais laboratórios em todo o mundo sequenciem as suas amostras e incorporem a sequenciação de moléculas de DNA (*Deoxyribonucleic acid*) ou RNA (*robonucleic acid*) na sua investigação. Uma destas tecnologias é a sequenciação do transcriptoma (RNA-seq) que representa uma forte alternativa ao uso de *microarrays* no estudo da expressão genética, sendo que para a análise destes dados são essenciais conhecimentos na área da Bioinformática e Biologia Computacional. Esta tecnologia tem permitido o estudo ao nível do transcriptoma não só de genes codificantes de proteína, como de outras sequências nucleotídicas não codificantes. Exemplos disto são os pseudogenes e ncRNAs (*non-coding RNAs*), permitindo assim compreender que também estes transcriptos não codificantes desempenham um papel importante em termos biológicos. Assim, a ideia de que estas sequências eram apenas zonas do genoma sem qualquer tipo de importância, sendo consideradas "lixo", tem sido desmistificada.

Os pseudogenes podem interagir com os seus genes parentais de diferentes formas: fonte de pequenos RNAs de interferência; transcriptos *antisense*; inibidores competitivos da tradução; RNAs endógenos competitivos; competindo pela ligação a microRNAs partilhados. Apresentam uma expressão específica de tecido para tecido,

sendo que o cérebro e os testículos são os tecidos que apresentam uma maior expressão de pseudogenes. Em 2014 foi descrito o caso de um gene (*OCT4A*) com um padrão de regulação associado a três dos seus pseudogenes durante a diferenciação neural de células estaminais humanas. Contudo, a extensão de pseudogenes envolvidos na regulação dos padrões de expressão associados com diferenciação nunca foram abordados globalmente. Deste modo, este trabalho tem o propósito de estudar a expressão dos pseudogenes na diferenciação de células estaminais em percursores neuronais, tanto em humano como em ratinho.

De modo a atingir o objetivo, foram analisados dados de transcripoma (RNA-Seq) de amostras obtidas ao longo da diferenciação neural em humano e ratinho. De modo a obter um catálogo completo de pseudogenes foram usadas três bases de dados (Ensembl, Yale e Noncode), perfazendo um total de 19444 pseudogenes no genoma de ratinho e 18061 no genoma de humano. Além disso foi também construída uma *pipeline* para descobrir novos potenciais pseudogenes, resultando num total de 130 (41 nas amostras de ratinho e 89 nas amostras de humano). Para obter os pseudogenes com expressão diferencial foram testados três métodos implementados em pacotes do R (DESeq e EdgeR), tendo-se optado por usar o pacote EdgeR para a análise final.

Devido à elevada semelhança entre as sequencias do pseudogene e o respetivo gene parental, os alinhamentos contemplaram apenas *reads* unicamente mapeadas. Assim, foi necessário estudar a mapabilidade dos pseudogenes, percebendo a singularidade dos mesmos e dessa forma filtrar os resultados. Após esta análise foi possível identificar 513 pseudogenes (92 de ratinho e 421 de humano) a variarem ao longo da diferenciação neural. A análise funcional dos respetivos genes parentais revelou 172 pseudogenes, potencialmente interessantes para a diferenciação celular e neural. A comparação dos resultados de ambos os organismos identificou um dos novos pseudogenes de ratinho como homólogo de um pseudogene humano anotado e contendo o ortólogo gene parental (*FAM205A*).

De modo a explorar a regulação dos genes parentais pelos seus pseudogenes, foi avaliada a correlação dos níveis de expressão para cada par de pseudogene-parental, sendo que em ambos os organismos foi encontrado um elevado número de pares de genes positivamente correlacionados.

Neste estudo, foram ainda incluídos os resultados com dados de transcriptoma ao nível de célula-única (*single-cell*). Devido ao nível baixo de sequenciação dos dados de célula-única a percentagem de pseudogenes expressos (com contagens processadas) foi reduzida e não permitiu detetar os pseudogenes da análise de transcriptoma global. Contudo, a análise destes dados revelaram quatro pseudogenes diferencialmente expressos cujos parentais têm funções relevantes a nível da diferenciação neuronal e envolvidos nas vias de sinalização de doenças neurodegenerativas.

Concluindo, o presente trabalho permitiu obter um conjunto de pseudogenes que variam ao longo da diferenciação neural e com o potencial para regularem genes parentais associados com diferenciação celular, neural ou doenças neurais. Além disso, os resultados realçam a importância da utilização de dados das tecnologias de sequenciação em larga-escala na descoberta de novos transcritos.

Palavras chave: pseudogene; pseudogenização; RPKM; diferenciação neuronal; RNA-Seq

Abstract

Pseudogenes are nucleotide sequences that were been neglected since they were discovered. In the last years this point of view is changing, and their functions have been studied and there are been annotated more pseudogenes than the estimated number in human genome.

Pseudogenization process can occur essentially by two mechanisms: genomic duplication of a parental gene *locus* (duplicated pseudogene); or retrotransposition (processed pseudogene) where an mRNA is reversely transcribed and randomly inserted in the genome. Initially these type of sequences can operate as a normal protein coding gene, but after some time and accumulation of deleterious mutations the open reading frame is modified, preventing their well-succeeded expression.

RNA-Seq technology allows studying the transcriptome of all type of biologic sequences, not only protein coding genes, giving an idea of the roles performed by them and demystifying the idea that they are genomic "junk".

There are evidences of interaction with their parental genes as: source of endogenous siRNAs; antisense transcripts; competitive inhibitors of translation; competitive endogenous RNAs (ceRNAs); competing for binding to shared miRNAs. In 2014 (last year) was described that a protein coding gene (*OCT4*) with a regulation patern associated with three of its pseudogenes in human stem cells differentiation.

To achieve our goal, were analyzed transcriptome sequencing (RNA-Seq) of neural differentiation datasets of human and mouse. Three databases were merged and a pipeline was constructed in order to find new possible pseudogenes, resulting in a total 130 in two dataset. Differential expression analysis was performed with two R packages (DESeq and EdgeR) with three different approaches, and after comparison, EdgeR pair wise analysis was selected as the best for our study.

Because of the high similarity between pseudogenes and their cognates, we only allowed reads uniquely mapped. Thus, it was necessary to study mappability of pseudogenes, realizing the uniqueness of them and therefore filter results. After this analysis, was possible to identify 513 pseudogenes varying along differentiation (92 in mouse dataset and 421 in human dataset). Functional analysis allowed to identify 172 potentially interesting pseudogenes in neural differentiation. Comparison of results between organisms identified one new putative mouse pseudogene homologous of anannotated pseudogene in human, with an ortholog parental (*FAM205A*).

To assess regulation of cognates by their pseudogenes, expression values were evaluated with Pearson's coefficient and there were found many pairs with significant correlation.

Processed data from single-cell experiments were analyzed too and there were highlighted four differentially expressed pseudogenes associated with neurodegenerative diseases and neural differentiation.

Finally, this work enabled to obtain a set of pseudogenes varying along neural differentiation with regulatory potential of parental genes associated with cell and neural differentiation or neurodegenerative diseases. The results highlight the importance of high throughput sequencing in discovery of new transcripts.

Key words: pseudogene; pseudogenization; RPKM; neural differentiation; RNA-Seq

Index

List of Tables

Supplementary Tables

1. Human new putative pseudogenes
2. Differentially Expression Analysis
3. Expression values for pseudogenes which cognates have relevant functions
4. Correlation between pseudogenes and their cognates
5. Single-Cell data Expression Analysis

List of Figures

List of abbreviations

ceRNA – competitive endogenous RNA

DEA – Differential Expression Analysis

DEG – Differentially Expressed Gene

DEP – Differentially Expressed Pseudogene

DNA – Deoxyribonucleic Acid

ds-RNA – double-strand RNA

endo-siRNA – endogenous small inferring RNA

ESC – Embryonic Stem Cell

FC – Fold-Change

FDR – False Discovery Rate

GEO – Gene Expression Omnibus

miRNA – micro RNA

mRNA – messenger RNA

NGS – Next-Generation Sequencing

NOS – Nitric Oxide Synthase

NPC – Neural Precursor Cell

OR – Olfactory Receptor

PCA – Principal Component Analysis

PCR – Polymerase Chain Reaction

PGM – Personal Genome Machine

RNA – Ribonucleic acid

RNA-Seq – RNA sequencing

RPKM – Reads Per Kilobase per Million

RPM – Reads Per Million

SRA – Sequence Read Archive

SMRT – Single Molecule Real Time sequencing

TPM – Transcripts Per Million

UCSC – University of California Santa Cruz

UTR – Untranslated Reagion

1. Introduction

1.1. Next Generation Sequencing

1.1.1. Technologies

In 1944 DNA (Deoxyribonucleic acid) was described as genetic material by Oswald Theodore Avery. Nine years later (1953) Watson and Crick determined the structure of DNA as we know nowadays, a double-helix structure, defined by sequences of four nucleotide bases. These events were crucial to the evolution of molecular biology and they led scientific community to go deeper in the understanding of the. In order to make it happen, there was a need to achieve a method that could sequence and tell us the order that those bases appear in specific DNA sequences or in entire genome.

The first step in sequencing technologies was driven by Sanger, with the development of a "first generation" technology based on chain-termination method (Sanger et al., 1977), able to determine which base is in a specific position of a certain region of genome. This method was used worldwide and represents a revolutionary technique that allowed scientists to understand more and more about DNA.

Since 1977 until now, a lot of techniques were develop and the first automatic sequencing machine (AB370) appear ten years later, developed by Applied Biosystems, based on capillary electrophoresis technique, allowing a faster and more accurate sequencing, ten years later from Sanger's discovery (Liu et al., 2012). The main goals after this turning point in the sequencing technology were increasing speed and accuracy while reducing cost. This became more evident with the Human Genome Project (2001). To achieve that were develop several "second generation" sequencing technologies - **Next Generation Sequencing** (NGS).

The first NGS technology commercialized was Roche 454 (2005), followed by Solexa/Illumina (2006) and SOLiD (2007) (van Dijk et al., 2014). Steps of fragmentation and amplification by Polymerase Chain Reaction (PCR) of genetic material are required in this type of sequencing before detection of specific nucleotides using fluorescence and camera scanning. In 2010, Ion Torrent released the Personal Genome Machine (PGM) that uses semiconductor technology instead of fluorescence. More recently, new technologies emerged (called "third generation") that allow the sequence in real time of single molecules without previous DNA amplification. One of the most used third generation methods was developed by PacBio in 2010 (van Dijk et al., 2014).

Although very expensive, the cost of sequencing using NGS has been decreasing along the past years (**Figure 1.1**) giving us great perspectives that these technologies will be even more reachable worldwide.

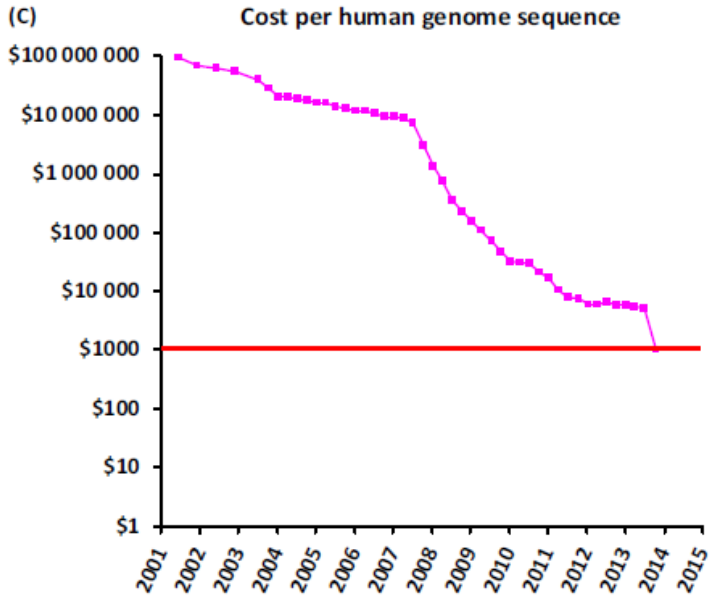

Figure 1.1.: Evolution of cost per human genome sequencing. Graphic from van Dijk et al., 2014. Red line represents the 1000$ cost threshold per human genome sequence, one of the goals achieved.

1.1.2. Applications

Evolution and wide availability of NGS technologies allowed, over the past decade, the development of several assays to answer different biological questions, focused on: transcription; translation; replication; post-transcriptional modifications; methylation; nucleic acids interactions; chromatin structure.

RNA-Seq (RNA sequencing) is the most widely used method to study transcription. In this methodology a population of RNA is converted to cDNA fragments (library preparation) with adaptors attached to one or both ends, and the sequenced (Wang et al., 2009). There are many other technologies focused on transcription. With **NET-Seq** (Native Elongating Transcript Sequencing) is possible to monitor transcription at nucleotide resolution, by deep sequencing 3' ends of nascent transcripts (Churchman et al., 2011). In **3P-Seq** (Poly(A)-Position Profiling by Sequencing) application, a RNA:DNA oligonucleotide is hybridized with thymines and ligated to the mRNA polyadenylated tail to prevent internal priming (Jan et al., 2011). **GRO-Seq** (Global Run-On Sequencing) methodology is applied to map position, amount, and orientation of transcriptionally engaged RNA polymerases by nuclear run-on RNA molecules (Core et al., 2008).

For determine chromatin structure the number of methodologies possible to use is very large. In FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) assays, chromatin and formaldehyde are cross linked *in vivo*, and together with massive parallel sequencing (**FAIRE-Seq**) is a methodology used to study the relationship between chromatin structures (Giresi et al., 2007; Yang et al., 2013). For mapping DNase I hypersensitive sites, **DNase-Seq** (DNase I Hypersensitive Sites Sequencing) is a method that selectively digests nucleosome-depleted DNA followed

by high-throughput sequencing (Song and Crawford, 2010). These two technologies are used to find "open chromatin" regions with regulatory activity corresponding to nucleosome-depleted regions (NDRs) (Song et al., 2011). With **ChIA-PET** (Chromatin Interaction Analysis by Paired-End Tag Sequencing) is a method for studying long-range chromatin interactions in a three-dimensional mode and provides a more trustworthy way to determine transcription factor binding sites and identify chromatin interactions (Li et al., 2014).

An important question in biology is the manner how proteins interact with nucleic acids. **ChIP-Seq** (Chromatin Immunoprecipitation Sequencing) is widely used to study protein-DNA interactions. It allows mapping genomic locations of transcription factors binding and histone modifications, after a protocol of chromatin immunoprecipitation and high throughput sequencing (Stephen et al., 2012). Focused on protein-RNA interactions, **CLIP-Seq** (Cross-Linking Immunoprecipitation) is an effective strategy by stringent purification of RNAs bound to a protein of interest in living cells (Murigneux et al., 2013).

In methylation studies **BS-Seq** (Bisulfite Sequencing) allows to measure cytosine methylation on a genome-wide scale within specific sequence contexts after bisulphite treatment of DNA (Cokus et al., 2008).

Translation is other of the biological questions that massive sequencing allows to understand. The basis of another NGS application, **Ribo-Seq** (Ribosome profiling), is the isolation of messenger RNA (mRNA) fragments protected by ribosomes followed by massively parallel sequencing. This methodology allows us to measure ribosome density along all mRNA transcripts (Michel et al., 2014).

In the next subchapter RNA-Seq methodology is more detailed, because is the one used in our study.

1.1.3. RNA-Seq overview

RNA sequencing (RNA-seq) is a NGS method specially developed for characterization and quantification of the transcriptome. First, it is necessary a construction of cDNA fragments library, from a RNA population, with adaptors attached to the sequence's ends. Each molecule is sequence (an amplification step may be necessary) resulting in millions of single-end or paired-end (RNA fragments sequenced from one end or both ends, respectively) as demonstrated in **Figure 1.2** (Wang et al, 2009).

Paired-end sequencing is more efficient dealing with multi-mapping which can be a serious problem at the alignment step, because both ends of each cDNA fragment should map nearby on the transcriptome, allowing solving this type of ambiguity in most part of the times, unlike single-end reads.

This method has many advantages when compared with microarrays hybridization-based approaches that had been the most widely used methodology to quantify the transcriptome. Specially, RNA-Seq shows higher sensitivity and dynamic range and lower technical variation (Oshlack et al., 2010). Furthermore, the technology does not require a previous knowledge of existing genomic sequences, unlike microarrays methodology, making RNA-Seq an interesting method to detect transcripts in non-model organisms (Wang et al, 2009).

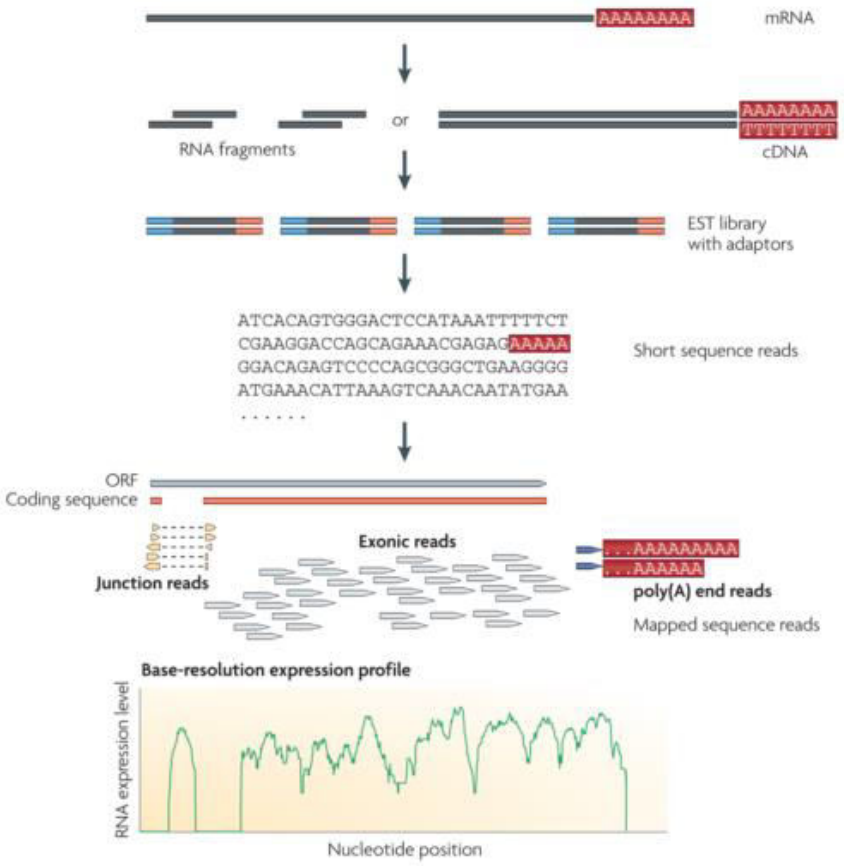

Figure 1.2: Summarization of a RNA-Seq experiment. Figure from Wang et al., 2009.

1.2. Pseudogenes

The first **pseudogene** was discovered by Jacq and colleagues (1977), for the oocyte-type 5S RNA gene in the genome of a model organism, *Xenopus laevis*. Since then, pseudogenes have been described as a relic of evolutionary selection by the scientific community (Muro et al., 2011).

Only some years after, Korneev and colleagues (1999) identified the first pseudogene with a relevant biological function, suggesting that the **nitric oxide synthase** (NOS) pseudogene transcript acts like an **antisense regulator** of neural NOS, its parental.

Due to the development of new approaches, several functional pseudogenes have been described for the past years, demystifying the idea that pseudogenes represent a category of “junk-DNA” or “genetic fossils”. This topic will be discussed in the next subchapters.

With the emerging knowledge of pseudogenes and their functions, arises the need to identify and annotate them genome-wide. Three years ago, GENCODE using simulations estimated that human genome contains, approximately 14000 pseudogenes (Pei et al., 2012). Now, GENCODE has more than 14000 pseudogenes

(**Table 1.1**) and with the evolution of sequencing technologies and analysis methods, this number is expected to increase.

Table 1.1.: GENCODE project annotated pseudogenes summarization for mouse and human genomes. Data obtained from the GENCODE site (gencodegenes.org/) on 13rd August 2015.

	Mouse (Version M6, GRCm38)	Human (Version 23, GRCm38)
# Total Annotated Pseudogenes	8787	14477
# Processed Pseudogenes	6097	10727
# Unprocessed Pseudogenes	2272	3271
# Unitary pseudogenes	15	172

1.2.1. Origin

Pseudogenes can be divided into some categories, depending on its origin. Essentially, there are three classes of pseudogenes: **processed**, **duplicated** and **unitary** (Pei et al., 2012). The first two classes are derived from genomic insertion events, while the third one just depends on accumulation of punctual mutations on the parental gene nucleotide sequence, leading to a "vestigial gene".

Duplicated pseudogenes derived from incomplete gene duplication events in a cell, as indicated by their names, usually in a near locus of the parental gene. This pseudogene suffers a **pseudogenization** process, losing its function as protein coding gene (Mighell et al., 2000).

Processed pseudogenes are the most studied class of pseudogenes, and a reason why that happens is because they are the most abundant. Their formation is originated by the reverse transcription of a mature mRNA (already spliced) back into DNA that is randomly inserted into the genome. Because of that, they lack promoter and introns and typically have 3'UTR (untranslated region) and poly-A tails. Normally these pseudogenes are found on different chromosomes of parental gene (Torrents et al., 2003; Pei et al., 2012).

Figure 1.4 summarizes the different pathways to generate a pseudogene starting from a protein coding gene (parental) for each class of pseudogene.

Curiously, the **olfactory receptors** (OR) genes is one of the families more pseudogenized in humans (Olender et al., 2008; Niimura, 2009). It was hypothesized that this phenomenon was related with the development of vision and decreasing of odor sensing. Olfactation results from the combination of different ORs, thus the use of pseudogenes to modify OR genes could be less dispendious (Olender et al., 2008).

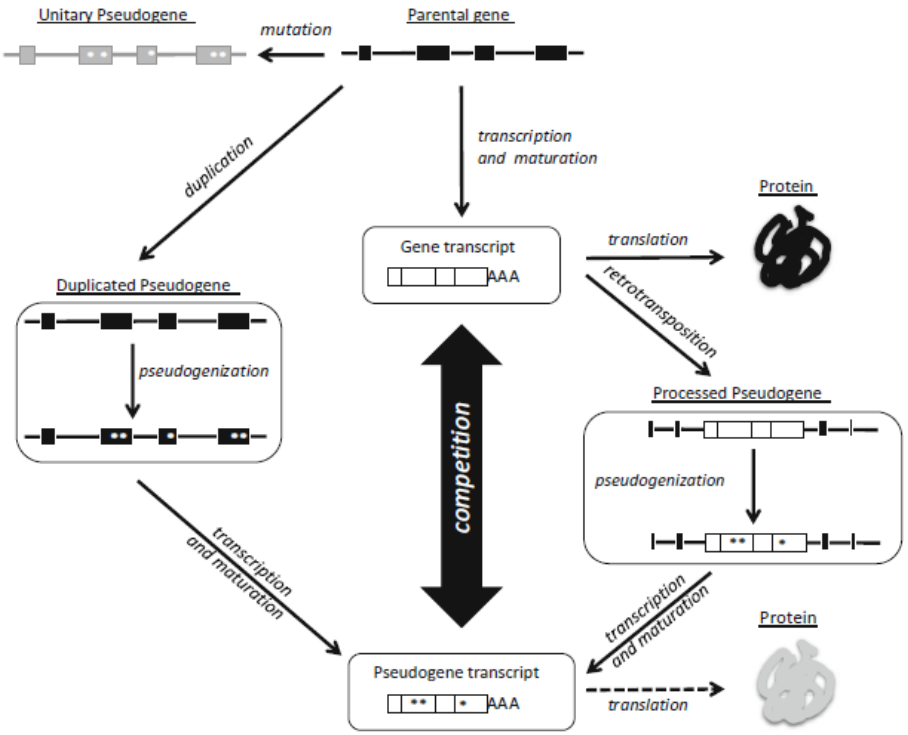

Figure 1.3: Pseudogenization processes for different classes of pseudogenes, deriving from a protein coding gene (parental gene). Figure from Laura Poliseno, Functions and Protocols, Pseudogenes book (2014).

1.2.2. Regulation of cognate genes

Pseudogenes can regulate their cognates (parental genes) by some different ways: **antisense transcripts**; **competitive inhibitors of translation**; **source of endogenous small inferring RNAs** (endo-siRNAs); **competitive endogenous RNAs** (ceRNAs).

The first evidences that pseudogenes regulate their cognates are from 1999, when it was understood that the **antisense** region of a pseudogene **transcript** prevented the mRNA translation of the respective parental (Korneev et al., 1999).

Connexin43 pseudogenes can inhibit their parental expression while they are expressed (Kandouz et al., 2004), acting as **competitive inhibitors of translation**.

Transcripts of pseudogenes can form double-strand RNAs (dsRNAs) by interacting with mRNAs resulting from protein-coding genes and then processed into **endo-siRNAs** (Tam et al., 2008).

Some transcribed pseudogenes, because of the high similarity with their cognate genes appear to be strong **ceRNAs** (Welch et al., 2015). Thus, most of the transcribed pseudogenes can regulate their parental genes by competing for miRNA biding (Poliseno et al., 2010).

The processes that lead regulation of protein coding genes by their pseudogenes are still being explored and described. The understanding of these genomic regions previously believed “genetic fossils” evolved in the last years and tends to increase with more studies focused on them.

1.2.3. Roles in neural differentiation

Neural differentiation from embryonic stem cells advances are very promising for nervous system therapies and neural tissue repair (Abranches et al., 2009) and it is a potential strategy for neurodegenerative diseases treatment (Felfly et al., 2011). Neural differentiation resulting in NPC formation shows specific regulation and gene expression patterns for clusters of genes (Zimer et al., 2011).

Pseudogene expression is tissue-specific, with testis and brain showing the highest levels of their transcription (Soumillon et al., 2013). Recent findings suggest that pseudogenes may play functional role on neural differentiation. For instance, the pluripotency regulator *OCT4* is a transcriptional activator of genes involved in maintenance of undifferentiated state and as a repressor of differentiation-specific genes. Notably, expression of *OCT4* and of its several pseudogenes was found to follow a developmentally regulated pattern in differentiating human embryonic stem cells (ESCs), suggesting that a tight regulatory relationship between them drives specific cellular functions (Jez et al., 2014). Pseudogenes widespread expression gives clues that these sequences may play also an important role in cancer, being some of them cancer-specific (Kalyana-Sundaram et al., 2012).

1.3. Objectives

Transcriptomic alterations during cell differentiation have been extensively characterized for protein-coding genes and non-coding RNAs, however the changes of other classes of non-coding genes, such as pseudogenes, have been largely unexplored. Hence, we decided to identify putative pseudogenes involved in neural differentiation and to assess their pattern conservation between different species.

Specifically, this works aims to:

1. Identify potentially new pseudogenes transcribed during neural differentiation;
2. Characterize the pseudogene transcriptome and determine the significant expression alterations;
3. Determine the functional role of pseudogenes, by monitoring the expression of the parental genes and their involvement in pathways relevant for neural differentiation;
4. Assess the conservation of pseudogene regulatory patterns between species.

To assess this, we will use RNA-Seq data obtained from different stages of **embryonic stem cells** (ESCs) differentiation into **neural precursor cells** (NPC), for mouse and human.

2. Methods

2.1. RNA-seq Data and Preprocessing

The present work involved the analysis of RNA-seq datasets for two different organisms, mouse and human. We gathered whole-transcriptome data from different stages of neural differentiation of mouse embryonic stem cells (ESCs) through collaboration with Dr. Ana Pombo from the Max Delbrück Center for Molecular Medicine (Berlin). Sequenced samples were collected for 5 time points (days 0, 1, 2, 3 and 4) along differentiation of ESCs to **neural percursors** (NPs) as previously described (Abranches et al., 2009). Transcriptomic data for human differentiation was recently published (Sauvageau et al., 2013) and was obtained from **Gene Expression Omnibus** (GEO) database (http://ncbi.nlm.nih.gov/geo/, Acc.Number **GSE56785**). Samples from NPC induced differentiation from stem cells, were collected over 7 time points (days 0, 1, 2, 4, 5, 11 and 18) assayed in triplicate cultures. Both transcriptomes contained paired-end sequencing reads with 100 bp (mouse) and 101 bp (human).

The first step of preprocessing, consisted in the conversion of the files downloaded from GEO in *.sra* format to *.fastq*, using the **SRA** (Sequence Read Archive) **Toolkit**, as showed by the following command:

- fastq-dump --split-files GEODownloaded.sra.

Fastq-dump command with --split-files option allows converting a *.sra* file to two *.fastq* format files, with matched mate-pair reads.

Second, data quality was assessed using **FASTQC tool** over *.fastq* files. This tool evaluates the quality of NGS data for the following characteristics: duplication levels; *k-mer* profiles; per base GC content; per base n content; per base quality; per base sequence content; per sequence GC content; per sequence quality; sequence length distribution.

2.2. Reads alignment

Paired-end reads were aligned against mouse and human reference genome (mm10 and hg19, respectively) with **TopHat2** (Kim et al., 2013), allowing 2 hits maximum and with a 100 bp mean distance between mates and a 50 bp standard deviation, as demonstrated in this example:

- tophat2 -g 2 --fusion-search --mate-inner-dist 100 --mate-std-dev 50 --output-dir OutputDirectory --GTF KnownTranscriptomeFile.gtf Bowtie2Index InputFile1.fastq InputFile2.fastq.

Since pseudogenes have high similarity with their cognates, the same read can align perfectly to both genes. Hence, to avoid this technical bias and assign each read correctly, only uniquely aligned reads were considered for downstream analyses. Thus,

after alignment reads mapped more than once were excluded, keeping only reads with "NH:i:1" flag in *.bam* output files.

2.3. New pseudogenes discovery

To assess the pseudogene expression profiles, along the last years there were designed several pipelines to identify new pseudogenes. Here, we aim also to detect new putative pseudogenes expressed along neural differentiation.

The method used to identify potential novel pseudogenes was based on two pipelines developed before (Kalyana-Sundaram et al., 2012; Zheng and Gerstein, 2006) and its schematically view represented in **Figure 2.1**. First, uniquely mapped reads were sorted by read name with **samtools sort** tool. Second, **bedtools pairtobed** tool was used to establish which paired-end reads do not intersect with annotated regions in neither ends and are located in the same chromosome, as showed by the command:

- pairToBed -abam SortedFile.bam -b KnownRegionExonsFile.bed -type neither -bedpe | awk '$1==$4' - > OutputFile.bedpe.

This step created a *.bed* file with the coordinates for each paired-end read located in intergenic regions. Next, the duplicated reads were removed. Then, the overlapping paired-end fragments are clustering using **bedtools merge** tool. Finally, only clusters longer than 40 bp, shorter than 5000 bp, taking into account that less than 10% of our human annotated pseudogenes are bigger than this threshold and supported by more than 1000 paired-end reads were selected:

- sort -V -k1,1 -k2,2 PairRangeFile.bed | awk '!x[$1,$2,$3]++' - | mergeBed -i stdin -c 6,1 -o distinct,count | awk '$5>100' - | awk ' $3-$2>40' - | awk 'BEGIN{FS=OFS="\t"} {print $1,$2,$3, $1 "_" $2 "_" $3, $5, $4}' - > MetaClustersFile.bed.

Then, we proceed to the identification of the putative parental genes for the discovered pseudogenes. First, we obtained the meta-clusters sequences using **bedtools getfasta**:

- fastaFromBed -name -fi ReferenceGenome -bed FinalMetaClustersFile.bed -fo FinalMetaClustersFile.fasta.

Since the sequence was determined, **BLAT** (Kent, 2002) against annotated protein coding genes was performed to determine putative parental genes for these meta-clusters, filtering the results with a similarity higher than 95% and a E-value lower than $1E^{-4}$.

After all, we obtained a *.gtf* file with the coordinates of the potential pseudogenes for each sample that were merged to obtain a single annotation file with all new discovered pseudogenes.

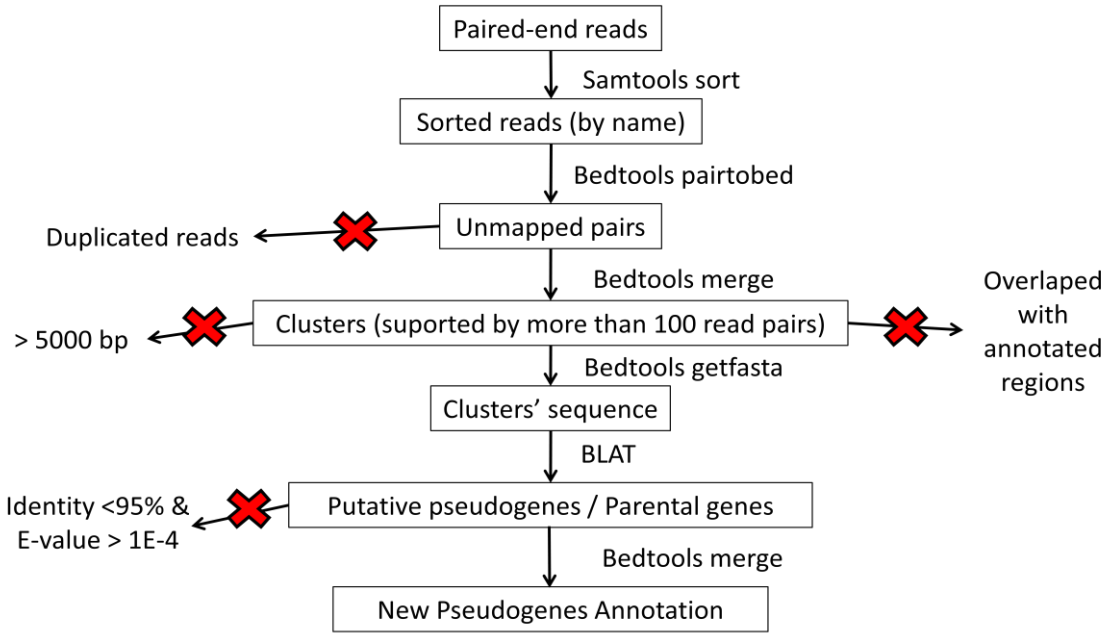

Figure 2.1.: Schematic view of pipeline to identification of new pseudogenes.

2.4. Reference genome annotation

The reference gene annotations used in further analyzes for both organisms are result of compilation and merging of all exons transcripts from three databases, **Ensembl**, **Yale** and **Noncode**. Versions of each annotation are Ensembl 75, Yale 74 and Noncode V4u1 for human (hg19 assembly) and Ensembl 76, Yale 76 and Noncode V4 (mm10 assembly). For Noncode V4 annotation was necessary to do a LiftOver step from mm9 assembly to mm10.

For downstream analysis, the new pseudogenes were concatenated to the reference genome annotation.

Table 2.1.: Summarization of merged annotations in human and mouse.

	Human (hg19)	Mouse (mm10)
# Protein Coding Genes	18915	21510
# Pseudogenes	18061	19444
# Other Non-coding RNAs	38126	42994

2.5. Gene summarization and normalization

In order to determine the expression levels of each transcript, **bedtools multicov** tool was run for all *.bam* files:

- bedtools multicov -split -bams File1.bam File2.bam (…) FileX.bam -bed MergedAnnotation.gtf > multicovOutputFile.txt.

This command results in a tab delimited *.txt* file, where the first columns are equal to *.gtf* file given as input, but with more X columns, being X the number of *.bam* files or, in other words, the number of samples.

Next, read counts were normalized calculating **RPKM** (reads per kilobase per million) for each gene in each sample, a measure to estimate gene expression, following this formula: RPKM = (10^9 * C)/(N * L); where C represents the number of reads mapping a gene, N the total number of reads and L the exon length for a gene, in a specific sample. This normalization step takes into account biases within lane (scaling for gene length) and between lane (adjust for total number of reads) and it was used for following unsupervised analyses (**Section 2.6**).

To normalize read counts for gene expression analysis (**Section 2.7**) within lanes and between them was used an R package, ***EDASeq***, because R packages do not accept RPKM values.

2.6. Unsupervised analysis

Exploratory analysis of RNA-Seq data usually involves unsupervised analysis to find hidden patterns or grouping data. One of the methods often used is **hierarchical clustering**. This method consists in an algorithm that agglomerates the data by a specific distance method chosen, ordering (n-1) subtrees, where n is the number of samples. Practically, the Euclidean distance between the RPKM values of each sample are estimated using dist() R function. Then, the distance matrix is submitted to hclust() function, to obtain the cluster dendogram graphical representation.

Other technique to achieve this is **principal component analysis** (PCA). PCA is a dimensionality reduction method, resulting into a new coordinate system where the first axis corresponds to the first principal component (PC). PCs are a new set of variables that can be interpreted as the direction that resumes the variation among them. The first few PCs normally capture most of the variation in original data, and the last few only capture "noise" (Yeung and Ruzzo, 2001). For this analysis the RPKM values were first standardized using the stdize() R function, implemented in *pls* package. Then, the standardized matrices were submitted to princomp() R function.

These two analyses were performed splitting pseudogenes and protein coding genes, in order to compare clustering of samples with different features.

2.7. Gene expression analysis

As pseudogenes are very low expressed compared to their cognates, there was a need to compare several methods to perform **differential expression analysis** (DEA). Three methods were perform with mouse's dataset, with two R libraries, ***EdgeR***

(Robinson et al., 2010) **pairwise comparison** and ***DESeq*** (Anders and Huber, 2010) **time series analysis** and **pairwise comparison**, where data is modeled as negative binomial distributed.

EdgeR pairwise comparison consists in using this R package estimateGLMCommonDisp() function to estimate negative binomial dispersion parameter for our gene expression datasets. Next, using glmFit() function and dispersion calculated we fitted a negative binomial generalized log-linear model for our counts. Finally, with fitted model, a statistical test is performed with glmLRT() function, for each day against day 0 of differentiation.

DESeq time series analysis uses fitNbinomGLMs() function to fit the generalized linear model of our counts dataset for the two hypothesis: fits to the formula that the time course is a covariate or fits the null hypothesis. Then, p-values where calculated with nbinomGLMTest() function, correcting this values with p.adjust() function.

The last method, *DESeq* pairwise comparison of read counts only uses nbinomTest() function for all days against day 0 of differentiation.

The comparison of these three methods is presented in chapter 3.

2.8. Pseudogenes Mappability

The use of uniquely aligned reads to avoid fragments mapping to multiple similar regions, such as pseudogenes and their cognates, will create a bias in the mappability of pseudogenes. Hence, the pseudogenes with high similarity to the parental genes will have long regions without reads aligned (i.e, unmappable). Thus, these pseudogenes will have false low expression values due the normalization of the read counts by the total gene length.

To overcome this problem we assessed the mappability/uniqueness analysis for pseudogenes *loci*. Resuming, this analysis consists in the alignment of the pseudogenes sequences, which were fragmented in *k-mers* of the original reads length. So, a *.gtf* file with *k*-mers of 100 nt was created for all pseudogenes annotated in mouse genome (extending 100bp upstream and downstream from annotation). For human genome, the process was the same but with 101nt (because reads have 101bp). Then these sequences were passed to *.fasta* format with **bedtools getfasta** tool and aligned to the reference genome with **bowtie2** tool. Non-unique aligned reads were excluded from the output *.bam* file and a **bedgraph** was generated with **bedtools genomecov** tool:

- genomeCoverageBed -ibam UniquelyMapped100mersFile.sam -g GenomeFile.txt -bg -split -scale 0.01 >> Unscaled.bedgraph.

The bedgraph was submitted to **UCSC Genome Browser** in *.bw* format and this track can be visualize as shown in **Figure 2.2.** Regions with higher mappability will represent unique regions in the genome, with maximum value being one.

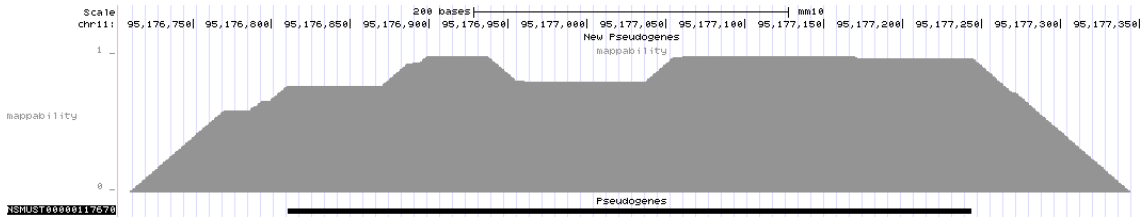

Figure 2.2.: UCSC Genome Browser mappability example track of pseudogene ENSMUSG00000083548.

Mappability varies significantly between different genomic regions, so we should take this in account especially for pseudogenes. Thus, we considered for analysis only the pseudogenes with mappability higher than 0.5 in at least 50% of total gene length.

2.9. Pseudogenes vs. parental genes

As discussed previously pseudogenes can regulate their cognates by some different ways (**Section 1.2.2**). To investigate this hypothesis, we compared the expression patterns of pseudogene and respective parental along neural differentiation. Thus, we performed a correlation test for each pair pseudogene/cognate using logarithmic RPKM values (log2) and **Pearson's coefficient** method implemented in cor.test() R function. Then, p-values were corrected for multiplicity problem using False Discovery Rate approach implemented in p.adjust() R function.

Finally, for each pair pseudogene/cognate with a significant correlation was produced a scatter plot, showing the expression for the pseudogene and respective parental gene (example in **Figure 2.3**).

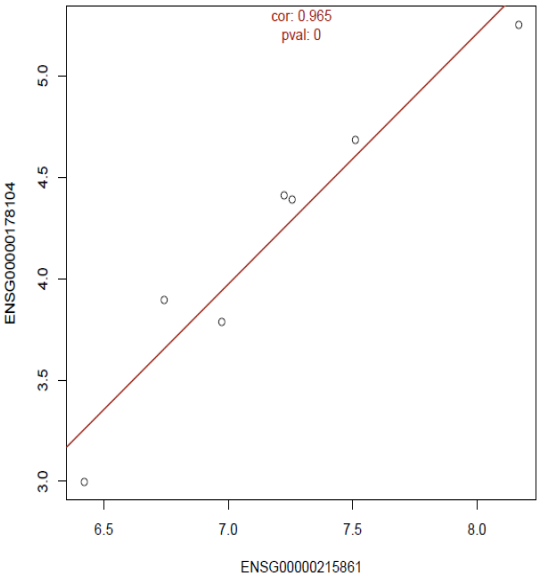

Figure 2.3.: Example of Correlation between pseudogene and its parental plot. In this case genes are positively correlated.

2.10. Functional/Pathway enrichment analysis

To assess which functions and pathways were significant altered during neural differentiation, we performed an enrichment analysis using **DAVID's functional annotation tool** (Huang et al., 2007)**.** We performed the analysis for all the genes differentially expressed and also for the parental genes of the pseudogenes with significant expression alterations.

For the analysis and functional assignment we used the following annotations: Gene Ontology Term for Biological Process (GOTERM_BP_FAT); Gene Ontology Term for Molecular Function (GOTERM_MF_FAT); Kyoto Encyclopedia of Genes and Genomes pathways (KEGG_PATHWAY).

Enrichment pathway analysis was performed with DAVID Functional Annotation Chart tools, using Fisher Exact statistics and default parameters. Only pathways/functions with Benjamini's corrected p-value lower than 0.05 were selected.

To identify the differentially expressed pseudogenes functionally relevant in neural differentiation, we assessed the function of respective parental genes. First, the DAVID Functional Annotation Table tool was used to obtain the function and pathways for each parental gene. Second, we searched for terms related to neural differentiation, cell differentiation, cell cycle and neurodegenerative diseases.

To visualize the expression patterns of functional relevant pseudogene/parental gene pairs, there were produced **heatmaps** with variations (log_2 fold-change) of pseudogenes differentially expressed and their respective cognates using heatmap() R function. The colors used in these heatmaps were created with brewer.pal() function from *RColorBrewer* package.

2.11. Species Comparison

Orthologs are genes evolved from a common ancestral by speciation and generally have the same function between species.

In order to identify conserved pseudogenes altered consistently in mouse and human neural differentiation we performed two different approaches. First, we compared the parental genes for which pseudogenes showed significant expression alterations along cell differentiation. We used **Ensembl Biomart** tool (ensembl.org/biomart/) to identify the orthologous genes between mouse and human. **Figure 2.4** shows the filters and attributes used. The output reports, for each mouse's parental gene Ensembl ID, the respective human Ensembl ID, the common ancestor, percentage identities with query gene and human gene and a binary value of orthology confidence. Only gene pairs with an orthology confidence value equals to one were considered.

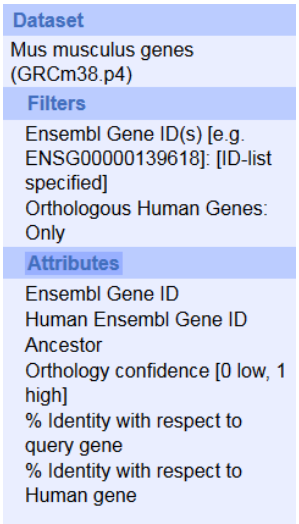

Figure 2.4: Ensembl Biomart orthology analysis summarization. Printed from Ensembl Biomart tool.

Second, we used **liftOver** tool from **UCSC Genome Bioinformatics** to identify homologous regions of the pseudogenes between different organisms. This was performed to convert the genomic coordinates of new pseudogenes discovered and annotated DEP in mouse dataset to human genomic coordinates and vice versa. The results from coordinates conversion were submitted to **bedtools intersect** tool in order to compare these converted positions against the genes annotated in the other organism.

2.12. Single-cell data comparison

To assess transcription heterogeneity of neural differentiated from a mouse ESC population we gathered processed high-throughput single-cell transcriptomic data recently published (Kumar et al., 2014). Processed data was obtained from Supplementary Information of the original study, which contained values of gene expression for each NPC and ESC cell, differential expression analysis between these two types of cells and some relevant statistics.

In order to estimate which genes are differentially expressed in this dataset, logarithmic expression values, measured in **transcripts per million** (TPM), of all ESCs and NPCs were compared using **Student's t-test** and correcting the p-values with p.adjust() R function, with default method, "Holm". A gene was defined as differentially expressed if the adjusted p-value of the statistic test was lower than 0.05. Finally, DEP found using the single-cell data were compared to the results of initial transcriptomic data.

3. Results and discussion

3.1. RNA-Seq Data Preprocessing

3.1.1. Data quality

Both transcriptomic datasets presented in general good quality for all parameters tested. One of the most important parameters tested is the per base sequence content (**Figure 3.1**) that represents the quality scores (y-axis) per each position or some position range. The background of the graph is divided in three per base quality groups: very good (green); reasonable (orange); poor (red). The data showed good quality through the entire read, thus not requiring any read trimming

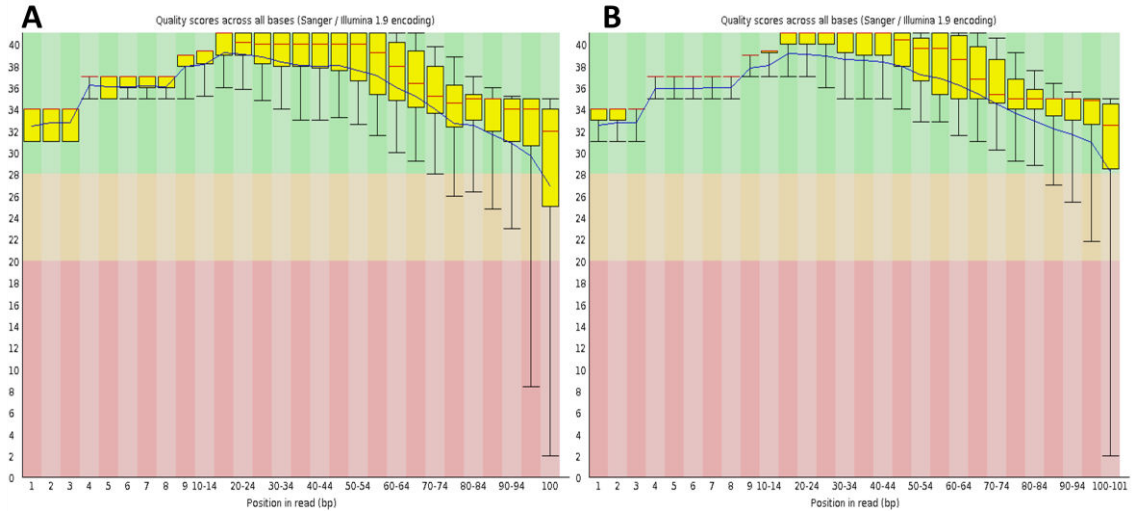

Figure 3.1.: Per base quality example plots (output from FastQC tool). (A) Mouse dataset, day 0, mates 1 (B) Human dataset, day0, replicate 2, mates 2.

3.1.2. Alignment

The genomes are full of repetitive sequences, and this can be a problem when we are mapping reads (Treangen et al., 2012). Since the goal of this work is to study pseudogenes with high similarity to the parental genes, we included only reads uniquely aligned. In both datasets the percentage of mapped read across the samples ranged between 80% and 90% (**Table 3.1**).

Other interesting result is that human dataset has an overall lower percentage of mapped reads when two hits are allowed. A higher percentage of uniquely mapped reads suggests that mouse genome has more repetitive regions than the human genome and that is concordant with other authors previous results (Haubold and Wiehe, 2006).

Table 3.1.: Alignment summary for all samples of both datasets.

Data set	Sample	# Mapped Reads (max_hits=2)	# Uniquely mapped Reads	# Total Reads	%Mapped Reads (max_hits=2)	%Uniquely Mapped Reads
Mouse	Day0	92337051	83001177	101304040	91.1	81.9
	Day1	88248264	78215517	96581066	91.4	81
	Day2	97631234	86654482	107000680	91.2	81
	Day3	90788837	82795456	98524424	92.1	84
	Day4	88634463	81590842	95975150	92.4	85
Human	Day0_rep0	45859736	45282356	50481512	90.8	89.7
	Day0_rep1	55063026	54397540	60463680	91.1	90
	Day0_rep2	63598813	62787961	70484132	90.2	89.1
	Day1_rep0	90493038	89071897	100526120	90	88.6
	Day1_rep1	81649316	80442231	91587496	89.1	87.8
	Day1_rep2	77550464	76388544	87427942	88.7	87.4
	Day2_rep0	42776570	42203886	46882574	91.2	90
	Day2_rep1	42681836	42106598	46922910	91	89.7
	Day2_rep2	42563041	42042856	47695440	89.2	88.1
	Day4_rep0	92074470	90744382	102437988	89.9	88.6
	Day4_rep1	89022617	87745660	98885808	90	88.7
	Day4_rep2	70518445	69535378	79234844	89	87.8
	Day5_rep0	153534079	151319987	171841758	89.3	88.1
	Day5_rep1	75743542	74659369	84068468	90.1	88.8
	Day11_rep0	111273149	109583911	133940020	83.1	81.8
	Day11_rep1	116764536	115093727	141720146	82.4	81.2
	Day11_rep2	125656336	123755819	151379400	83	81.8
	Day18_rep0	102438671	100960490	115552342	88.7	87.4
	Day18_rep1	93981413	92591605	106219460	88.5	87.2
	Day18_rep2	86559391	85301216	97691602	88.6	87.3

3.2. Identification of New Pseudogenes

The workflow described to identify new pseudogenes (**Section 2.3**) was applied to mouse and human datasets. Analysis revealed 41 putative pseudogenes for mouse (**Table 3.2**) and 89 for human (**Supplementary Table 1**). The difference may be due to the higher number of samples in human dataset.

Table 3.2.: New putative pseudogenes discovered following the pipeline described at section 2.3 in mouse dataset.

Gene ID/coordinate	Parental Gene Name
chr1:24050708-24051679	*Gm10160*
chr1:32486416-32486746	*Lypla1*
chr1:88271268-88272541	*Ccdc79*
chr1:153082942-153083145	*1700008F21Rik*

chr11:3318094-3319118	*Svop*
chr11:95864992-95865643	*Gm10160*
chr11:106994798-106995686	*Timeless*
chr12:11239030-11239552	*Gm11032*
chr12:38868568-38869078	*Gm10563*
chr13:22835511-22837112	*Zranb3*
chr13:23325311-23326004	*Cd59b*
chr13:100782127-100782952	*Svop*
chr13:112881216-112882447	*Snf8*
chr15:33221975-33222671	*Gm10491*
chr15:76888487-76888750	*Timeless*
chr15:99471290-99472055	*Mybl2*
chr16:13976683-13977868	*Ifitm7*
chr16:30955238-30955539	*Sec14l4*
chr16:38831711-38832823	*Cd59b*
chr17:12895759-12896800	*Arhgef40*
chr17:32314259-32315309	*Fance*
chr18:44735655-44735823	*1700008F21Rik*
chr19:38950878-38951669	*Gm10160*
chr2:26639789-26639995	*Bnc2*
chr2:156993480-156994037	*Limk2*
chr2:177089204-177089784	*Dpp6*
chr3:88351639-88352590	*Timeless*
chr3:123265980-123266769	*Slc17a9*
chr4:42716171-42717210	*Gm13298*
chr4:129727966-129728671	*Letm1*
chr4:140714527-140715809	*Dlg1*
chr4:152327691-152328501	*Fance*
chr5:110772559-110773175	*Amotl2*
chr7:138889831-138890521	*Gtdc1*
chr8:12476765-12477829	*Slc17a9*
chr8:22174052-22174555	*Atg13*
chr8:105845572-105846232	*Lman2l*
chr9:13843163-13843916	*Zfp280d*
chr9:15315649-15316132	*Letm1*
chr9:57507830-57508505	*Drosha*
chrX:52741373-52741588	*Gm10491*

Some putative pseudogenes possessed the same parental genes such as *Timeless*, *Gtdc1*, *Fance* and *Cd59b*, so this result may be generated because of the repetitiveness of specific sequences in genome that align these protein coding genes.

In **Figure 3.2** it is possible to see the tracks in UCSC Genome Browser for two examples of new pseudogenes discovered in both datasets. Pseudogene annotation is represented at the bottom of each subfigure.

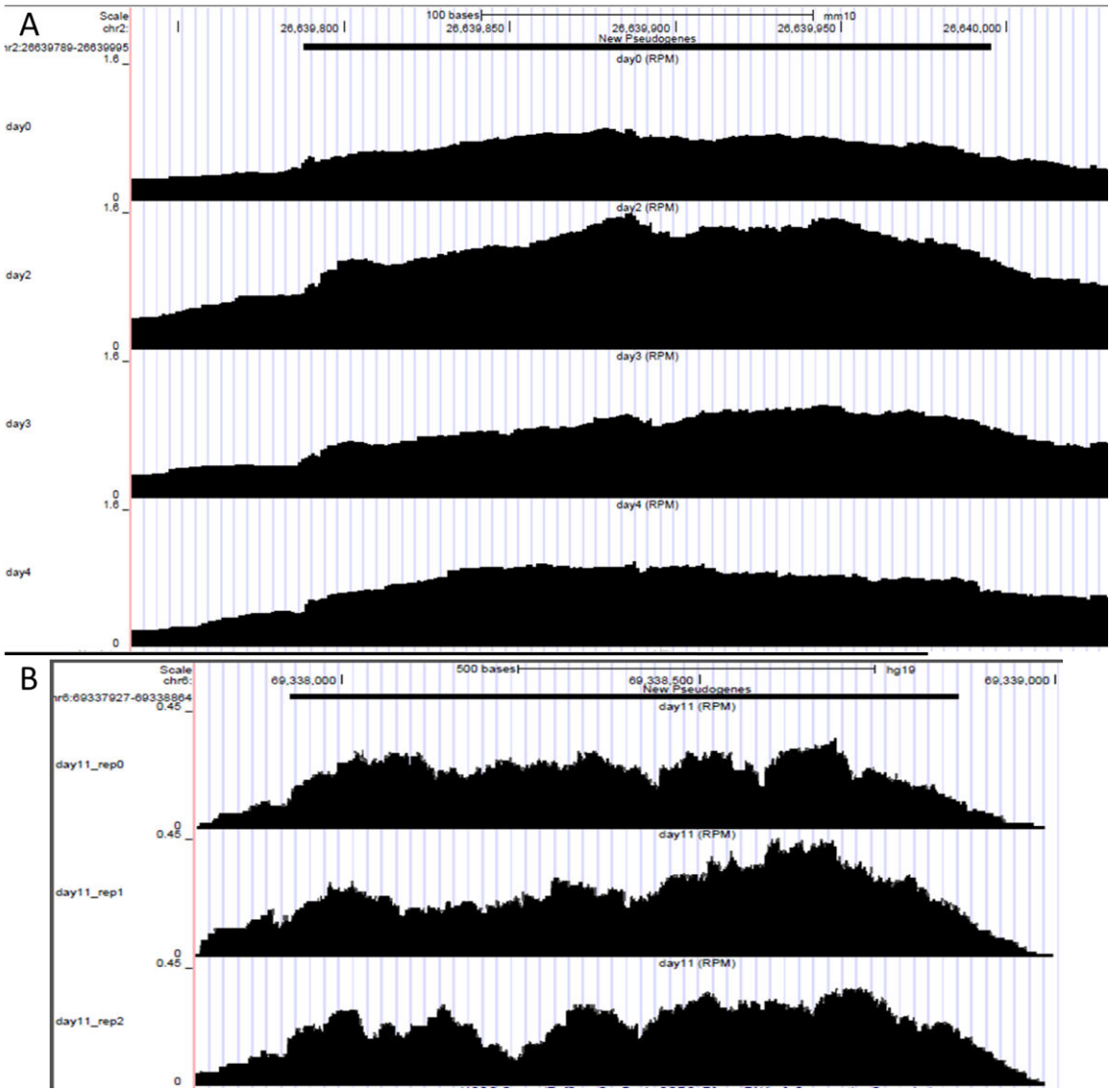

Figure 3.2.: UCSC Genome Browser tracks for new pseudogene examples. (A) Mouse new pseudogene in coordinates chr2:26639789-26639995 (parental gene – Bnc2; involved in regulation of transcription) in all time-points. (B) Human new pseudogene in coordinates chr6:69337927-69338864 (parental gene - REV3L; involved in DNA repair) in day 11, triplicate.

3.3. Unsupervised Clustering Analysis

To characterize the pseudogene transcriptome we applied unsupervised methods, such as hierarchical clustering and PCA (**Figure 3.3**). They suggest that log(RPKM) values, in general, separated samples by time. In the case of human triplicate samples, they are clustered essentially according the time-point, with few exceptions. There are highlighted 3 examples (days 0, 2 and 4, **Figure 3.3 C**) where that is very clear. For mouse differentiation the two first PCs of pseudogene expression could explain half of the variance (54.6%), whereas for human data the percentage of variation decreased to 16.21%.

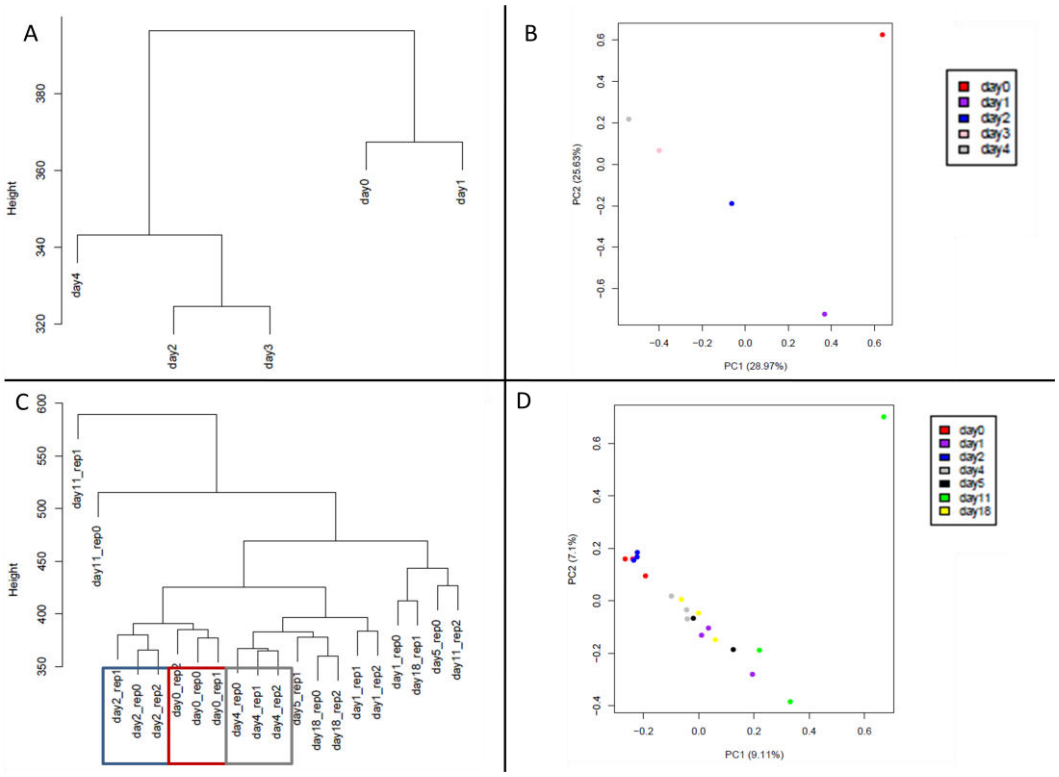

Figure 3.3.: Unsupervised clustering analysis of pseudogenes. (A) Clustering dendogram of mouse samples. (B) PCA plot of mouse samples. (C) Clustering dendogram of human samples. (D) PCA plot of human samples.

The same analysis was performed for protein coding genes, in order to assess the impact of the transcriptome in samples clustering. Both analyses revealed also time-series grouping (**Figure 3.4**), with a slight increase in the variance percentage represented by the two first PCs (60.25% for mouse and 18.44% for human).

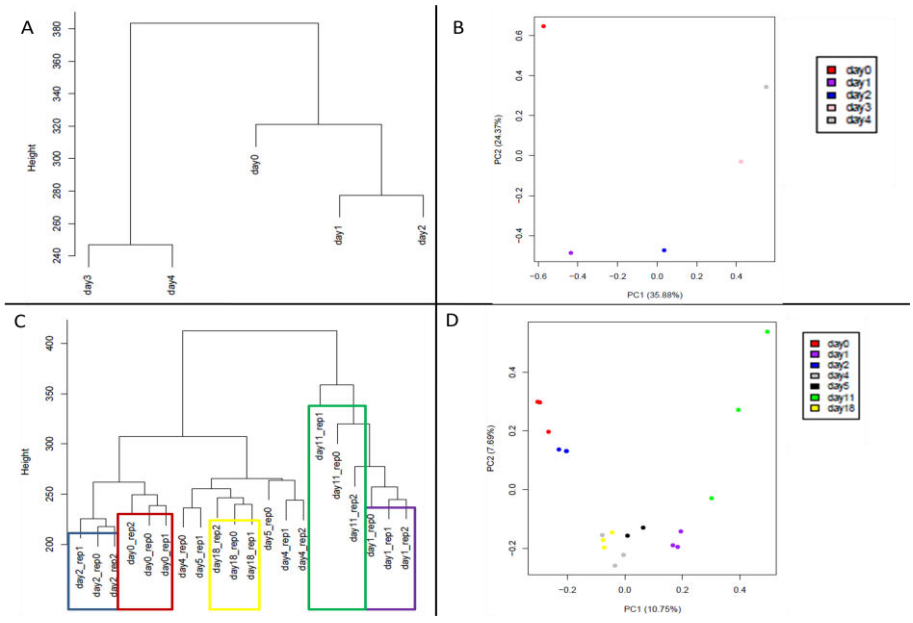

Figure 3.4.: Unsupervised clustering analysis of protein coding genes. (A) Clustering dendogram of mouse samples. (B) PCA plot of mouse samples. (C) Clustering dendogram of human samples. (D) PCA plot of human samples.

This difference between the pseudogenes and protein-coding transcriptome can be explained by the overall lower and more variable expression of pseudogenes.

3.4. Comparison of Statistic Methods for Differential Expression Analysis

Differential expression analysis was performed using three different methods (EdgeR pairwise comparison, DESeq pairwise comparison and DESeq time-series analysis) in order to choose the best alternative for our study.

All methods show that **differentially expressed genes** (DEG) number increases along differentiation (**Figure 3.5**). DESeq pairwise comparison shows the lower number of DEG. DESeq time series analysis produces the high number of DEG. EdgeR pairwise comparison shows higher percentage of **differentially expressed pseudogenes** (DEP). One limitation of DESeq time series analysis is that this method does not provide a DEG list for each time point. This method only reports a p-value that indicates if a gene expression changes in any time, but not which time-point is that, despite retrieving fold-change for each time. We assessed that a gene was differentially expressed if have a p-value lower than 0.05 and an absolute logarithmic fold-change for that specific time-point higher than 0.58.

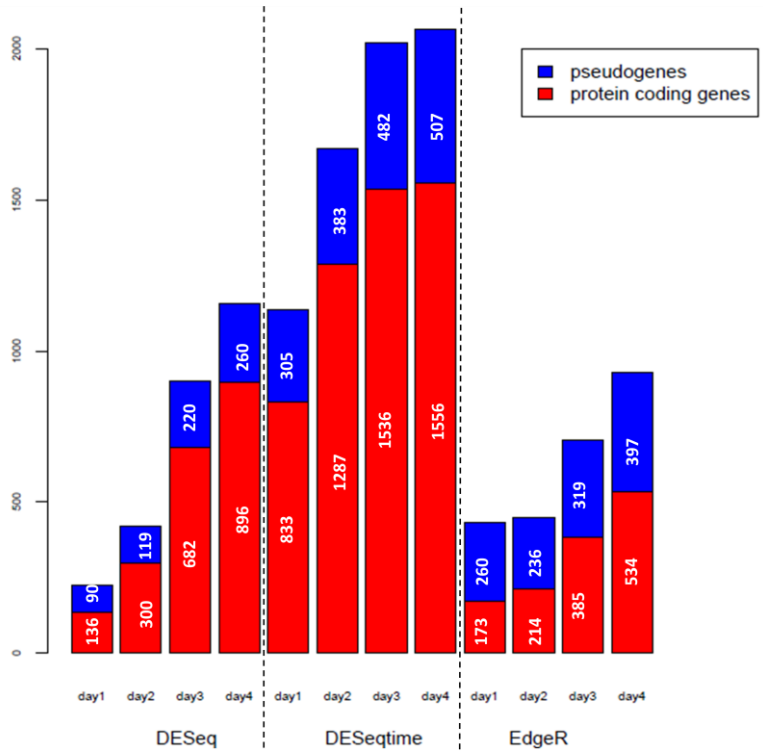

Figure 3.5: Gene expression analysis summarization for three distinct methods (DESeq pairwise comparison, DESeq time series analysis and EdgeR pairwise comparison). Numbers represents the number of differentially expressed genes (FDR or adjusted p-value < 0.05 and |log(Fc)|>0.58).

Only few DEP were consistently identified by all statistical (**Figure 3.6**), with higher number of common DEP found between the two DESeq pipelines. However, the DESeq-time workflow revealed to produce a higher number of DEP not found in any of

the other analyses. Although the low number of common pseudogenes, the two methods with the larger number of DEP (EdgeR pairwise comparison and DESeq time-series analysis), showed consistent expression fold-changes (**Figure 3.7**).

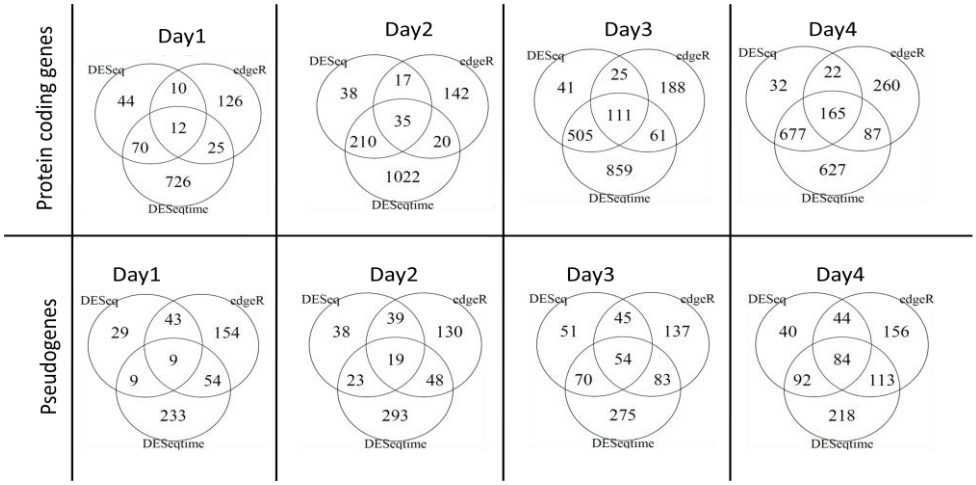

Figure 3.6.: Venn Diagrams for all time-points of differentially expressed protein coding genes and pseudogenes, between all methods.

Thus, these two methods differed essentially in the estimated p-values, where EdgeR package calculates a p-value for each pseudogene on each time-point and DESeq-time reports if a specific pseudogene expression varies over time. This do not allows us to make a certain decision about differential expression in specific time-points, so this is a reason why this analysis produces the highest number of DEP and differentially expressed protein coding genes. The DESeq-pairwise analysis appeared to be too strict, since it only identified a small number of DEP.

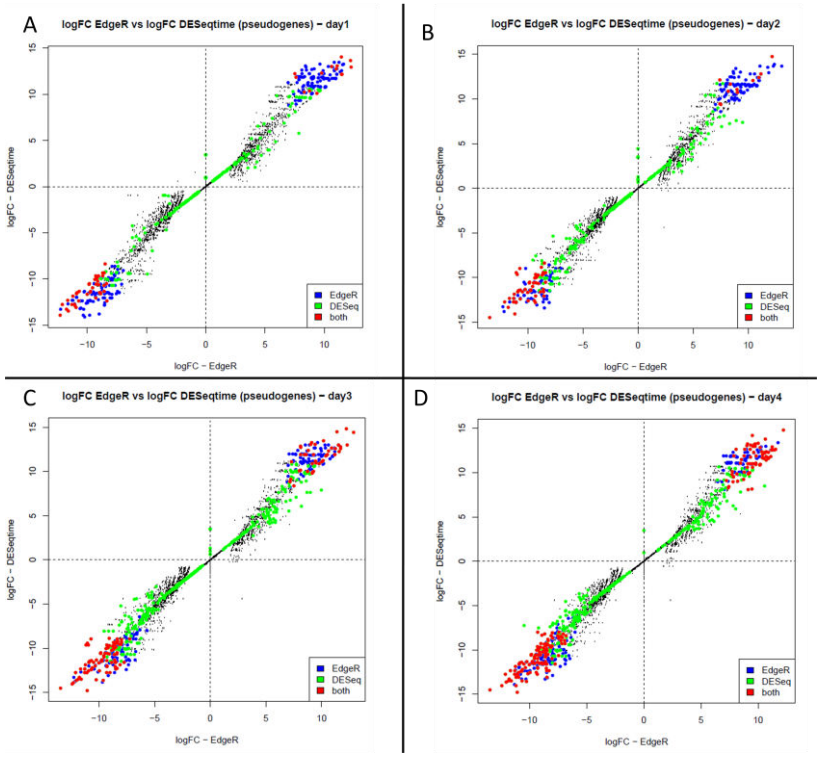

Figure 3.7.: Fold change comparison of differentially expressed pseudogenes between two different methods (EdgeR and DESeq time series analysis), for all time-points.

So, to obtain a confident reasonable amount of DEP for each time-point we performed the downstream differential expression analyses using the EdgeR method.

3.5. Differential Expression Analysis and Pseudogene Mappability

The differential expressed pseudogenes were determined comparing each time-point to initial stage with EdgeR method, using a **false discovery rate** (FDR) cut-off value of 0.05 and a absolute value of **logarithmic fold change** (log_2) higher than 0.58.

To filter low covered or noisy pseudogenes in our samples, only those that present 60% of read coverage and appeared as differentially expressed for, at least, two time-points were considered for downstream analyses (**Supplementary Table 2**).

Comparison of volcano plots for both datasets revealed that the human differentiation showed higher amplitude of expression fold-changes (**Figure 3.8 C**). Due to the absence of replicates in mouse dataset, only pseudogenes with higher fold-changes were considered significant (**Figure 3.8 A**). Comparison of the fold-change with the mean expression level revealed that for mouse dataset, pseudogenes and protein-coding genes were equally distributed (**Figure 3.8 B**). In opposition, human pseudogenes showed lower mean expression levels but largest fold-changes (**Figure 3.8 D**). This was consistently observed for all time-points (data not shown).

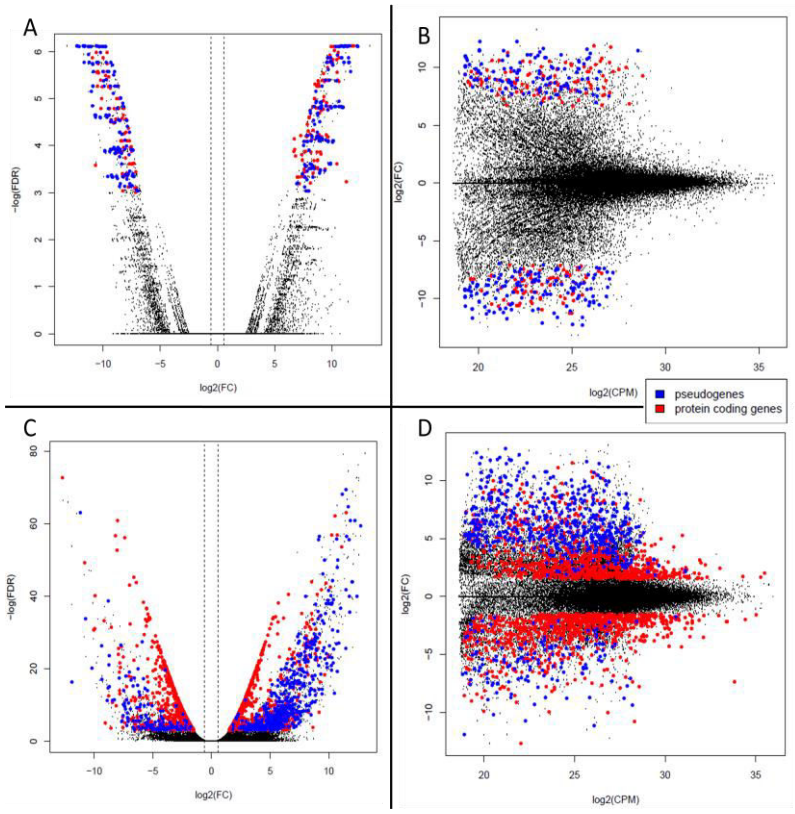

Figure 3.8.: Day 1 expression view. Blue dots represent differentially expressed pseudogenes and red dots represent differentially expressed protein coding genes. (A) Volcano plot of mouse data. (B) MAplot of mouse data. (C) Volcano plot of human data. (D) MA plot of human data.

Figure 3.9 shows, for each dataset, an example of low read covered pseudogenes. It may be due to the fact of pseudogenes have low uniqueness because they are very like their parental genes. To assess this hypothesis, the process was the one described in **Section 2.8.** To ensure the pseudogenes were being expressed and not "noise" we required that at least 90% of the mappable region should contain one read aligned.

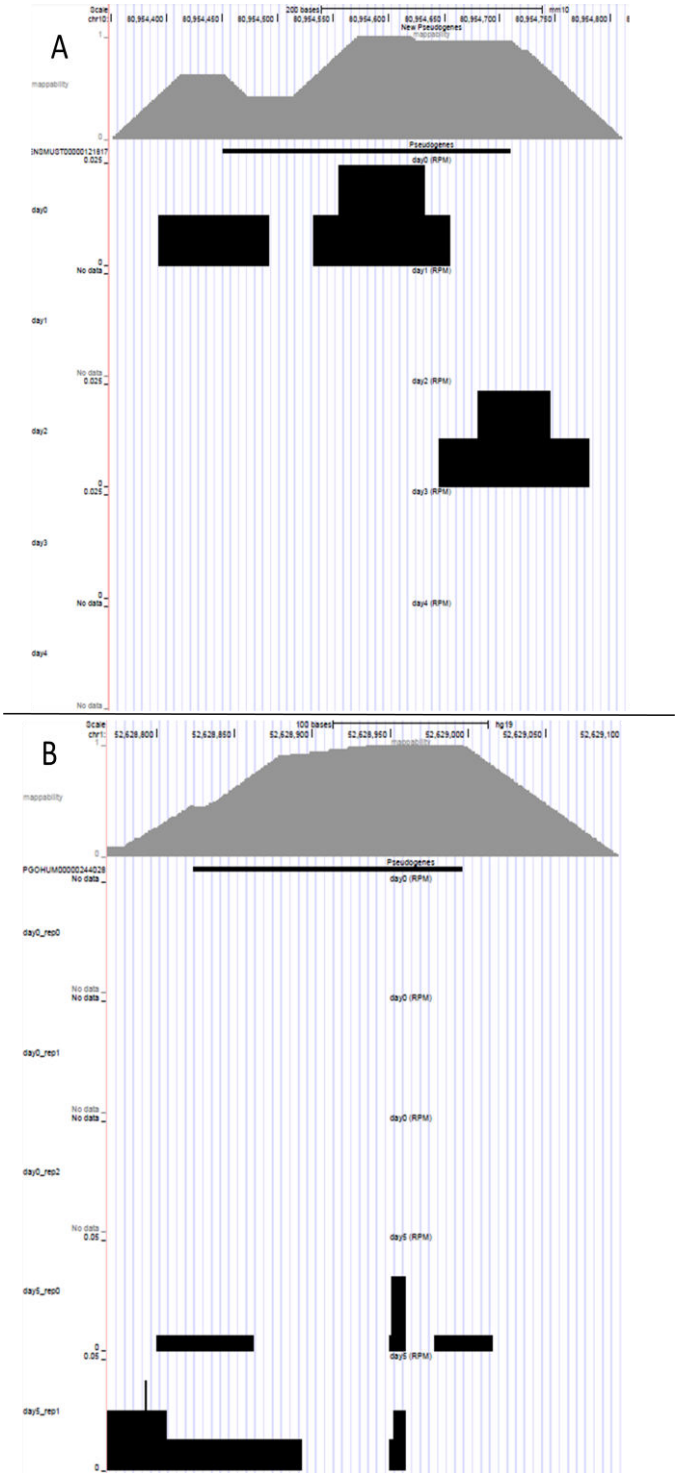

Figure 3.9.: UCSC Genome Browser example tracks for low covered DEP. (A) ENSMUSG00000083808 (mouse pseudogene), all days. (B) PGOHUM00000244028 (human pseudogene), day 0 triplicate and day 5 duplicate.

Only 20% of the DEP showed mappability higher than 90% (**Table 3.3**). We decided to apply this filter only after the statistical analysis, to not eliminate permanently candidate pseudogenes just because of their percentage of similarity with their cognates. Thus, although showing low mappability, these pseudogenes could be followed in the future using other experimental assays.

Table 3.3.: Troubleshooting summary.

	Total number of pseudogenes	Mappable> 90%	#DEP	Mappable > 90% ∩ DEP
Mouse	19444	18920	721	92
Human	18061	16562	1786	421

3.6. Pseudogenes and Neural Differentiation

When applying the statistical method we observed that the proportion of differentially expressed pseudogenes was very similar to the protein-coding genes. For three time-points (mouse data – days 1 and 2; human – day 18) the proportion of DEP was even higher than the number of differentially expressed protein-coding genes (**Figure 3.10**). These results may indicate the importance of pseudogenes in neural differentiation.

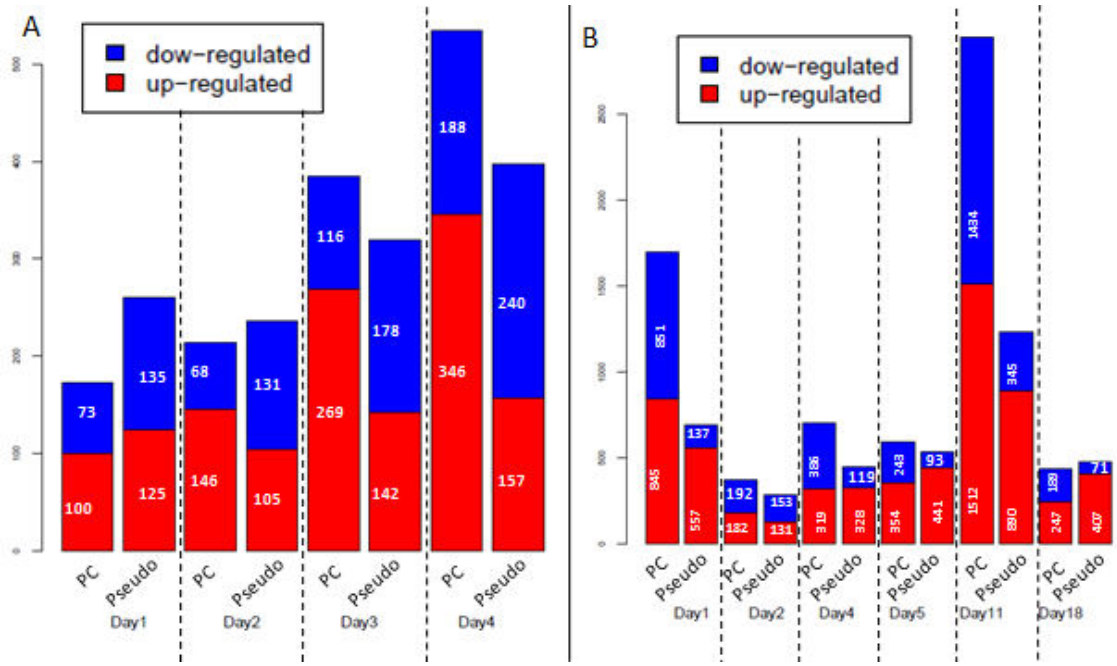

Figure 3.10.: Expression summary. Bar plots show the number of protein coding genes and pseudogenes differentially expressed. (A) Mouse data. (B) Human data.

Other interesting result is that day 11 of human data appears to have an abnormal expression of protein-coding genes and pseudogenes comparing to other time-points.

From all new pseudogenes discovered from mouse transcriptomic data, only one appeared as differentially expressed in mouse dataset at day 3 (chr9:57507830-57508505). Its parental gene was ***Drosha***, involved in **miRNA processing** (Lee et al., 2003). The human transcriptomic dataset also revealed only one new differentially

expressed pseudogene (chr6:69337927-69338864). For this case, parental was REV3L, known as required for common fragile sites (CFSs) stability (Bhat et al., 2013). CFSs are "hot spots" of genomic instability (Debatisse et al., 2012).

After applying all filters of coverage and mappability, the numbers of pseudogenes differentially expressed decreased drastically (**Figure 3.11**). Overall, mouse data revealed similar proportions of pseudogenes up and down-regulated, whereas for human differentiation mostly genes were up-regulated. Besides that, for common times in two datasets, the number of pseudogenes down-regulated is essentially higher in human, with the exception of day4. The amount of pseudogenes up-regulated in human dataset is clearly larger and comparing with down-regulated pseudogenes, they represent an abnormal percentage. These differences between different organisms may be due to the fact that the two experiments were not designed with the same procedure, only human dataset contained replicates reinforcing statistical analysis.

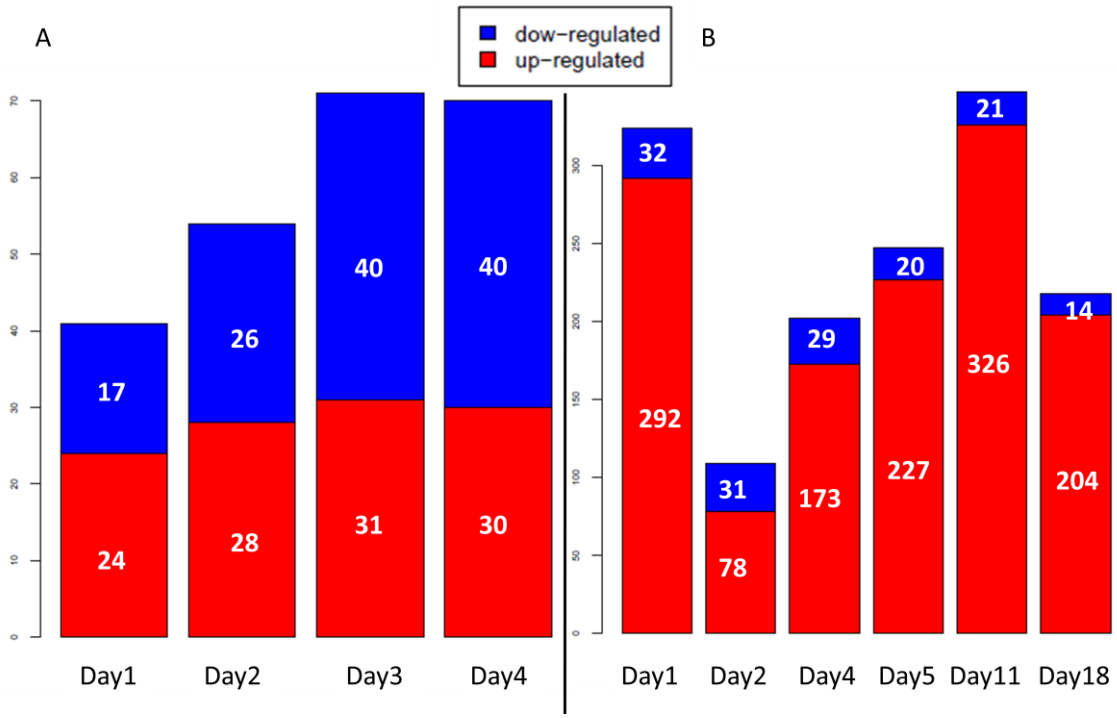

Figure 3.11.: Expression summary of filtered DEP. Bar plots for mouse data (A) and human data (B).

All DEP that passed all the thresholds imposed and the respective parental gene identifiers were submitted to DAVID as described at **Section 2.10**, to assess their functions and possible pathways where they are involved (**Table 3.4**). In human dataset is possible to see three neurodegenerative diseases pathways enriched (highlighted in red).

Table 3.4.: Functional/Pathway analysis. Adapted from Functional Annotation Chart output from DAVID Functional Annotation online tool (david.ncifcrf.gov/).

Organism	Term	Count	Benjamini adjusted p-value
Mouse	Ribosome	17	1.91E-19
	structural constituent of ribosome	16	8.38E-17
	translation	16	1.56E-11
	structural molecule activity	17	2.98E-11
Human	translational elongation	42	2.69E-47
	Ribosome	42	7.83E-43
	structural constituent of ribosome	44	4.30E-40
	translation	50	8.53E-34
	structural molecule activity	48	2.87E-19
	Parkinson's disease	22	1.32E-10
	Oxidative phosphorylation	22	1.21E-10
	RNA binding	38	7.04E-10
	hydrogen ion transmembrane transporter activity	14	3.65E-08
	oxidative phosphorylation	14	4.01E-07
	monovalent inorganic cation transmembrane transporter activity	14	1.83E-07
	inorganic cation transmembrane transporter activity	15	1.85E-06
	ribonucleoprotein complex biogenesis	16	1.02E-05
	Huntington's disease	20	1.64E-06
	generation of precursor metabolites and energy	20	2.06E-05
	oxidoreductase activity, acting on heme group of donors, oxygen as acceptor	8	4.95E-06
	cytochrome-c oxidase activity	8	4.95E-06
	heme-copper terminal oxidase activity	8	4.95E-06
	oxidoreductase activity, acting on heme group of donors	8	4.95E-06
	ribosome biogenesis	13	2.70E-05
	ribosomal small subunit biogenesis	6	3.57E-05
	Alzheimer's disease	18	8.19E-06
	rRNA processing	11	8.93E-05
	rRNA metabolic process	11	1.18E-04
	mitochondrial ATP synthesis coupled electron transport	9	1.18E-04
	ATP synthesis coupled electron transport	9	1.18E-04
	respiratory electron transport chain	9	3.03E-04
	cellular respiration	10	8.20E-04
	mitochondrial electron transport, NADH to ubiquinone	7	0.002102197
	NADH dehydrogenase (quinone) activity	7	0.001342639
	NADH dehydrogenase activity	7	0.001342639
	NADH dehydrogenase (ubiquinone) activity	7	0.001342639
	electron transport chain	10	0.002591404
	oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor	7	0.002548872
	ncRNA processing	12	0.00503617
	ncRNA metabolic process	13	0.007054583

	energy derivation by oxidation of organic compounds	10	0.012959922
	ribosomal large subunit biogenesis	4	0.018492917
	Cardiac muscle contraction	9	0.010239628
	oxidation reduction	21	0.039928389
	oxidoreductase activity, acting on NADH or NADPH	7	0.040748874

A heatmap was produced to see the patterns of regulation for a total of 172 DEP (146 from human genome and 26 from mouse genome) that passed all filters and which respective parental genes have relevant functions on differentiation or neurodegenerative diseases (**Figures 3.12 and 3.13; Supplementary Table 3**).

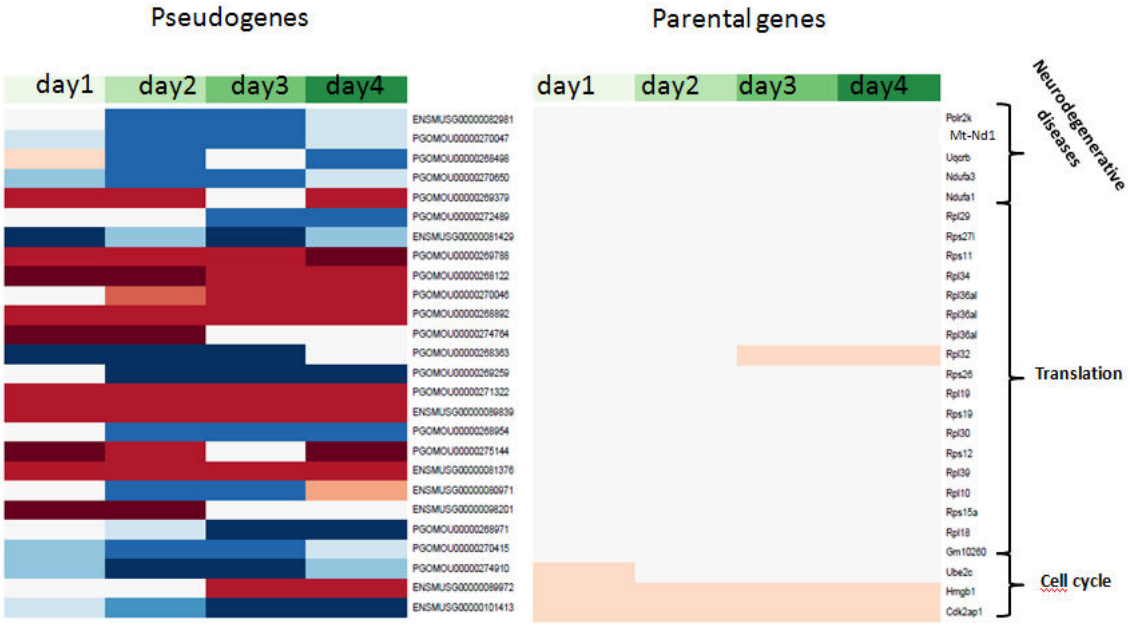

Figure 3.12.: Heatmap with mouse pseudogenes and their cognates with relevant functions in neurodegenerative diseases pathways or involved in translation and cell cycle processes.

It is possible to see that there are two families of parental genes very represented here: **small ribosomal proteins** (RPS gene family) and **large ribosomal proteins** (RPL gene family). Ribosomal proteins activity may control gene expression and mammalian development (Kondrashov et al., 2011) and some of them are reference genes for neuronal differentiation (Zhou et al., 2010). Our analysis showed that two of these ribosomal proteins, ***RPS15A*** and ***RPL18***, possessed pseudogenes with significant transcriptome alterations.

Another interesting result was the presence of five genes of **cytochrome c oxidase** (COX) subunits as cognates of DEP. Cytochrome c oxidase is a bigenomic enzyme, a rare case, accounting that are only four of this type, resulting in a combination of three mitochondria-encoded subunits and ten nucleus-encoded subunits. As described previously, neurons depend on COX for their survival and proper functional development, being its regulatory mechanisms been explored along the past years. As result there are some clues but not a fully comprehensive conclusion of how this process occurs (Wong-Riley, 2012; Dhar et al., 2013). What is known until now is that all nucleous-encoded subunits expression is regulate by **nuclear respiratory factors** (*NRF-1* and *NRF-2*). These two factors are regulated by neuronal activity and respond to it (Dhar et al., 2008). Mitochondrial-encoded subunits

expression is indirectly regulated by them too, because they activate transcription factors A and B (*TFAM, TFB1M* and *TFB2M*) (Gleyzer et al., 2005). Sp1 transcription factor was described as a bigenomically regulator of all COX subunits genes two years ago and the expression of this gene is dependent of neuronal activity too (Dhar et al., 2013).

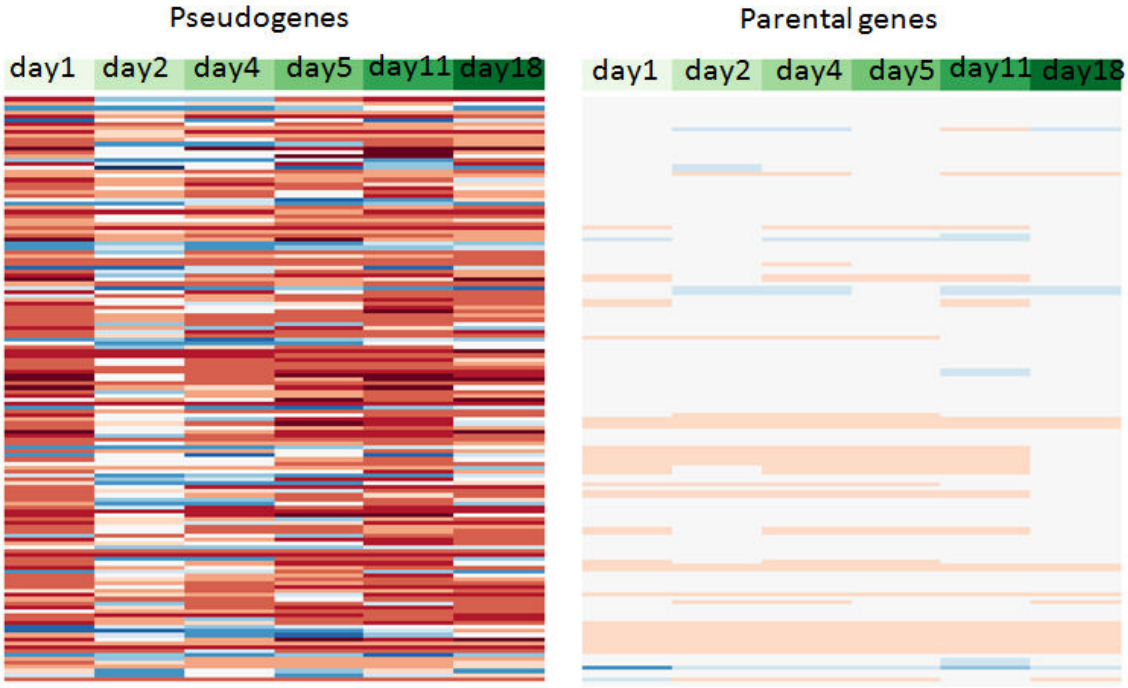

Figure 3.13.: Heatmap with human pseudogenes and their cognates with relevant functions in neurodegenerative diseases pathways or involved in translation, neuron differentiation and cell cycle processes. Because of the large number of pseudogenes it is not possible to show pseudogene ID and parental gene names. These results are in Supplementary Table 3.

Comparison of the expression levels of pseudogenes and respective parental gene along differentiation revealed that 312 mouse pseudogenes were statistically correlated (267 positively and 45 negatively correlated). For human dataset were found 1294 statistically correlated pairs, wherein 904 were positively correlated and 390 are negatively correlated (**Supplementary Table 4**). In both cases it seems that pseudogenes regulate their cognates, in general, in the same direction. None of these pseudogenes significantly correlated with their parental genes (p-value < 0.05) was differentially expressed and passes mappability filter, simultaneously.

In summary, pseudogenes seem to regulate their cognates, but experimental assays and more computational approaches are needed to clarify the mechanism of regulation.

3.7. Species Comparison

Proceeding as described in **Section 2.11**, we could not identify orthologues pseudogenes differentially expressed in both mouse and human neural differentiations. However, one of the new pseudogenes discovered in chr4:42716171-42717210 mouse genomic coordinates, when performing a **liftOver**, falls in a human

genomic region that belongs to an annotated pseudogene with a parental ortholog (*FAM205A* - **transmembrane protein** C9orf144B) (**Table 3.5**), which gives a higher confidence that the pseudogenization process occurred in these two species from a common ancestral sequence.

Table 3.5.: New mouse pseudogene orthology analysis.

		New Pseudogene
Mouse	ID/Coordinates	chr4:42716171-42717210
	Parental Gene Name	Fam205a2
Human	Coordinates from Liftover	chr9:34894173-34895331
	ID	ENSG00000187791
	Coordinates	chr9:34889060-34895775
	Parental Gene Name	FAM205A

Although not detected as DEP, this new possible mouse pseudogene is expressed in all samples and its expression seems to slightly increase along differentiation (**Figure 3.14**).

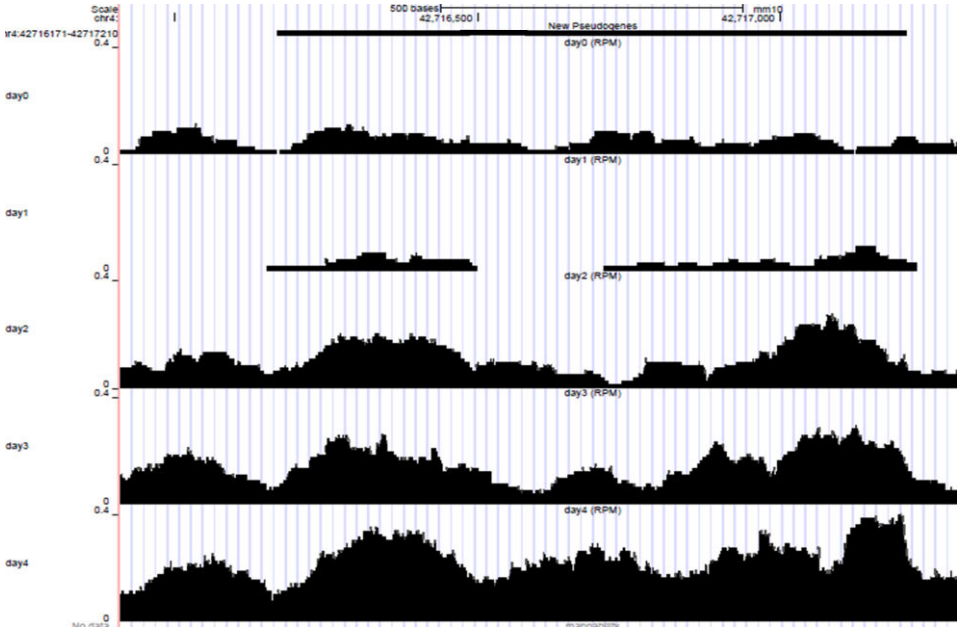

Figure 3.14.: UCSC Genome Browser tracks for pseudogene in coordinates chr4:42716171-42717210, with Fam205a2 as its parental gene.

Orthology studies and comparison of similar genomic regions between species did not show an overlap. This suggests that pseudogenization processes are specific for each organism and it is a process that goes along evolution and speciation.

3.8. Comparison with Single-Cell Data

The mouse transcriptomic single-cell data previously published (Kumar et al., 2014) was pre-processed by the authors using the Ensembl annotation, resulting in a smaller set of pseudogenes relative to our study. Comparing the pre-processed expression

levels, we could identify 39 pseudogenes differentially expressed (**Supplementary Table 5**). Probably caused by the different annotations used, none of them was in common with the results from our global transcriptomic data. **Table 3.6** summarizes the pseudogenes differentially expressed in single-cell dataset.

Table 3.6.: DEP in Single Cell dataset.

Pseudogene ID	Pseudogene Name	Parental Gene Name	log2(NPC/ESC)	Adjusted p-value
ENSMUSG00000081249	*Gm11517*	*Kxd1*	-4.953	1.44E-33
ENSMUSG00000057990	*E030024N20Rik*	*Ppia*	-1.693	1.01E-14
ENSMUSG00000098065	*Gm5177*	*Gapdh*	-1.839	6.17E-11

From this table is possible to see, three pseudogenes which their cognates have relevant functions for neural differentiation. The pseudogene with a cognate involved in neural differentiation was *E030024N20Rik* (*Ppia*). The other two pseudogenes showed parental genes associated with the neurodegenerative diseases: *Gm5177* (*Gapdh*) for Alzheimer's and *Gm11517* (*Kxd1*) for Parkinson's disease **Figure 3.15** shows that these three pseudogenes are down-regulated in neural differentiation. This suggests that these pseudogenes are, in general, required in the beginning of differentiation to regulate their cognates and after this event, their expression decays.

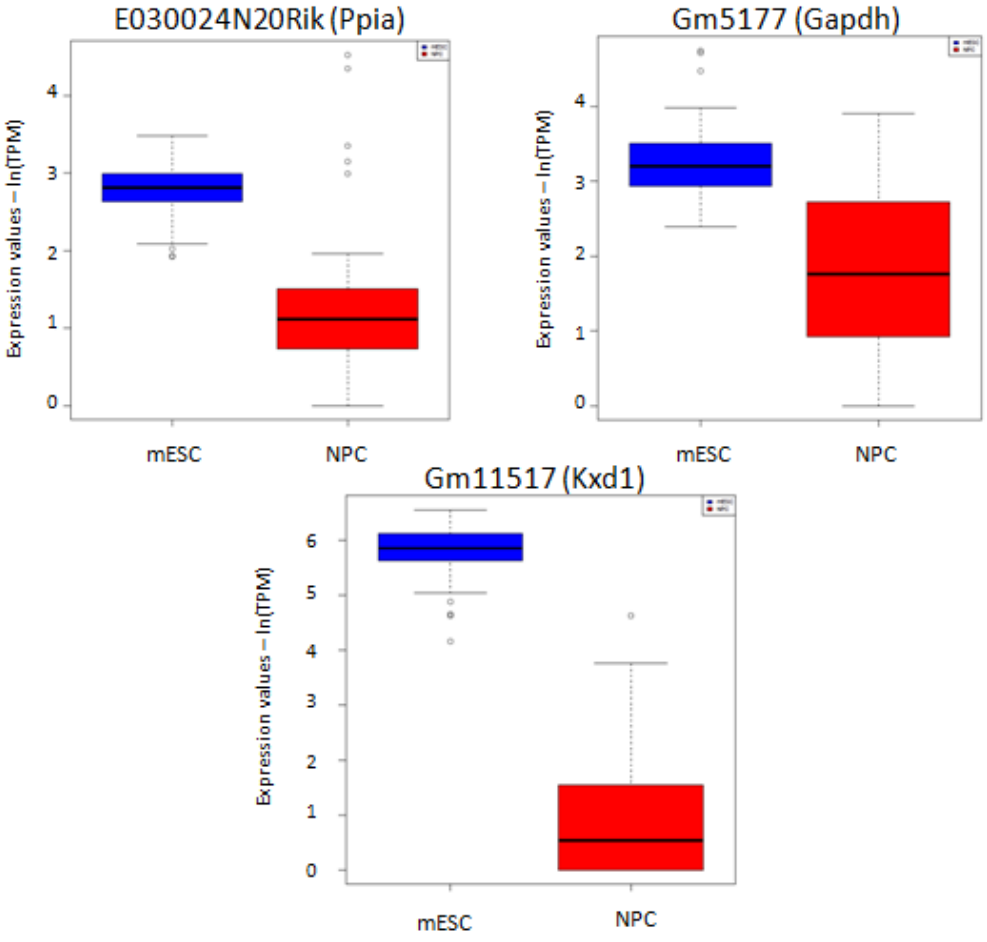

Figure 3.15.: Expression box plots for differentially expressed pseudogenes which parental genes are associated with relevant neural functions and pathways. Parental gene names between parenthesis.

For these three pseudogenes was analyzed their expression in our samples (**Figure 3.16**). Pseudogenes annotations are represented on the top of each set of tracks. The second track, counting from the top, represents the mappability and the other five the expression in RPMs for each time-point on mouse neural differentiation. *E030024N20Rik* pseudogene shows a really low uniqueness, because they have sequences very similar with other genomic regions. In the case of *Gm5177* pseudogene, there are no reads mapped despite its higher mappability, showing that it is not being expressed at all in this experiment. Lastly, *Gm11517* pseudogene showed similar expression levels for all time-points, thus differences observed in single-cell data may reflect sample heterogeneity (not observed in global transcriptomic data).

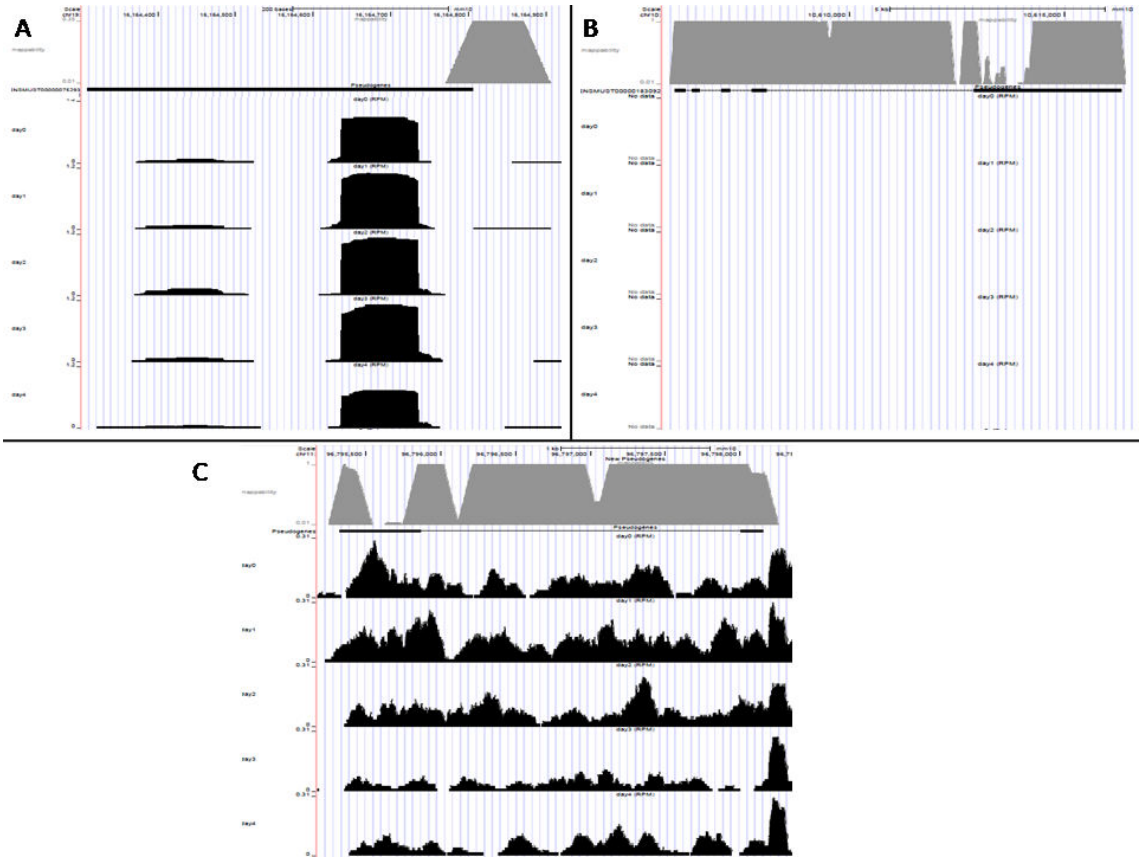

Figure 3.16.: UCSC Genome Browser tracks of mouse dataset for DEP in Single-Cell dataset. (A) *E030024N20Rik.* (B) *Gm5177*. (C) *Gm11517.*

This comparison was very important to highlight the idea that pseudogene expression studies depend largely on approaches and annotations used. To deeply assess pseudogene expression alterations we would have to re-do the analysis starting from the raw data and using more adequate methods for single-cell.

4. Conclusions

In this study was assessed and described the different steps and challenges involved in the analysis of pseudogene transcriptome. The most defiantly part of this work was to counterbalance low expression levels of pseudogenes and their similarity with respective parental genes, without a huge loss of information. For certain pairs of pseudogene and parental gene with high similarity, it is impossible to correctly distinguish the individual expression patterns. Thus, in this study we only considered reads that aligned uniquely, giving more confidence in the depicted expression levels. However, this approach may lead to a loss of sequencing reads for pseudogenes that normally already have overall low expression values.

In the future, with the development and increased accuracy of technologies like SMRT from Pacific Biosciences studies like this can be performed without loss of information. This technology allows sequencing longer reads (14,000 – 40,000) with an accuracy of 99% in a very short run time and without a previous amplification step (pacificbiosciences.com/). With this type of technology will be possible to completely map the transcripts and distinguish pseudogenes from cognate genes.

Even without the perfect technology for this study, we could identify 130 new possible pseudogenes, 41 in mouse genome and 89 in human genome. To depict their functional role in neural differentiation future experimental assays should be performed. One of these putative pseudogenes, located in chr4:42716171-42717210 of mouse genome possess a homologous pseudogene in human genome and the same ortholog parental gene (*FAM205A*).

After all analysis and filters applied we identified 513 differentially expressed pseudogenes along neural differentiation, from which 172 had the respective parental genes associated with neural-related functions.

Our analysis also revealed significant correlation between expression levels of pseudogenes and the respective parental genes in both organisms. This may reflect regulation of cognate genes expression by the pseudogene, however experimental validation will be necessary to confirm this hypothesis.

Single-cell transcriptomic data revealed few differentially expressed pseudogenes, probably due to the different annotation used to preprocess the data in the original study. However, our analysis identified three pseudogenes with significant alterations for which the cognate genes are involved in neural differentiation and neurodegenerative diseases. Future and complete analysis of single-cell data will be necessary to deeply assess the differences from single and global transcriptomic approaches.

In conclusion, it was already proven that pseudogenes have important roles regulating their cognates, but there is a lack of understanding which of them are functionally important and how conserved these are between different organisms. Thus, our study provides insights to fulfill this scientific gap, but only with more experimental assays and more accurate sequencing technologies will be possible to extensively assess pseudogenes roles.

References

1. Abranches, E., Silva, M., Pradier, L., Schulz, H., Hummel, O., Henrique, D., Bekman, E., 2009. Neural differentiation of embryonic stem cells in vitro: a road map to neurogenesis in the embryo. PLoS One 4, e6286.

2. Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106.

3. Bhat, A., Andersen, P.L., Qin, Z., Xiao, W., 2013. Rev3, the catalytic subunit of Polζ, is required for maintaining fragile site stability in human cells. Nucleic Acids Res. 41, 2328–39.

4. Churchman, L.S., Weissman, J.S., 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469, 368–73.

5. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., Jacobsen, S.E., 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–9.

6. Core, L.J., Waterfall, J.J., Lis, J.T., 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–8.

7. Debatisse, M., Le Tallec, B., Letessier, A., Dutrillaux, B., Brison, O., 2012. Common fragile sites: mechanisms of instability revisited. Trends Genet. 28, 22–32.

8. Dhar, S.S., Johar, K., Wong-Riley, M.T.T., 2013. Bigenomic transcriptional regulation of all thirteen cytochrome c oxidase subunit genes by specificity protein 1. Open Biol. 3, 120176.

9. Felfly, H., Xue, J., Zambon, A.C., Muotri, A., Zhou, D., Haddad, G.G., 2011. Identification of a neuronal gene expression signature: role of cell cycle arrest in murine neuronal differentiation in vitro. Am. J. Physiol. Regul. Integr. Comp. Physiol. 301, R727–45.

10. Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., Snyder, M., 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res. 17, 669–81.

11. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., Lieb, J.D., 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 17, 877–885.

12. Gleyzer, N., Vercauteren, K., Scarpulla, R.C., 2005. Control of mitochondrial transcription specificity factors (TFB1M and TFB2M) by nuclear respiratory factors (NRF-1 and NRF-2) and PGC-1 family coactivators. Mol. Cell. Biol. 25, 1354–66.

13. Head, S.R., Kiyomi Komori, H., LaMere, S. a., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., Ordoukhanian, P., 2014. Library construction for next-generation sequencing: Overviews and challenges. Biotechniques 56, 61–77.

14. Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35, W169–75.

15. Jan, C.H., Friedman, R.C., Ruby, J.G., Bartel, D.P., 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 469, 97–101.

16. Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.-M., Cao, X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J., Iyer, M.K., Barrette, T., Shanmugam, A., Dhanasekaran, S.M., Palanisamy, N., Chinnaiyan, A.M., 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell 149, 1622–34.

17. Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome Res. 12, 656–64.

18. Kandouz, M., Bier, A., Carystinos, G.D., Alaoui-Jamali, M.A., Batist, G., 2004. Connexin43 pseudogene is expressed in tumor cells and inhibits growth. Oncogene 23, 4763–70.

19. Kondrashov, N., Pusic, A., Stumpf, C.R., Shimizu, K., Hsieh, A.C., Xue, S., Ishijima, J., Shiroishi, T., Barna, M., 2011. Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. Cell 145, 383–97.

20. Korneev, S.A., Park, J.-H., O'Shea, M., 1999. Neuronal Expression of Neural Nitric Oxide Synthase (nNOS) Protein Is Suppressed by an Antisense RNA Transcribed from an NOS Pseudogene. J. Neurosci. 19, 7711–7720.

21. Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., Jay DaleyKeyser, A., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J., Ferrante, T.C., Regev, A., Daley, G.Q., Collins, J.J., 2014. Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature 516, 56–61.

22. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., Kim, V.N., 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415–9.

23. Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V, Kolchanov, N.A., Ruan, Y., 2014. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics 15 Suppl 1, S11.

24. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–9.

25. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012.

26. Michel, A.M., Fox, G., M Kiran, A., De Bo, C., O'Connor, P.B.F., Heaphy, S.M., Mullan, J.P.A., Donohue, C.A., Higgins, D.G., Baranov, P. V, 2014. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Res. 42, D859–64.

27. Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F., 2000. Vertebrate pseudogenes. FEBS Lett. 468, 109–114.

28. Murigneux, V., Saulière, J., Roest Crollius, H., Le Hir, H., 2013. Transcriptome-wide identification of RNA binding sites by CLIP-seq. Methods 63, 32–40.

29. Muro, E.M., Mah, N., Andrade-Navarro, M.A., 2011. Functional evidence of post-transcriptional regulation by pseudogenes. Biochimie 93, 1916–21.

30. Niimura, Y., 2009. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Hum. Genomics 4, 107–18.

31. Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), pp.5463–7.

32. Olender, T., Lancet, D., Nebert, D.W., 2008. Update on the olfactory receptor (OR) gene superfamily. Hum. Genomics 3, 87–97.

33. Oshlack, A., Robinson, M.D., Young, M.D., 2010. From RNA-seq reads to differential expression results. Genome Biol. 11, 220.

34. Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X.J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T.J., Harrow, J., Gerstein, M.B., 2012. The GENCODE pseudogene resource. Genome Biol. 13, R51.

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

36. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 14, R95.

37. Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–40.

38. Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S.C., Mallard, W., Morse, M., Swerdel, M.R., D'Ecclessis, M.F., Moore, J.C., Lai, V., Gong, G., Yancopoulos, G.D., Frendewey, D., Kellis, M., Hart, R.P., Valenzuela, D.M., Arlotta, P., Rinn, J.L., 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife 2, e01749.

39. Sequence Read Archive Submissions Staff. Using the SRA Toolkit to convert .sra files into other formats. In: SRA Knowledge Base [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.

40. Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., Dym, M., de Massy, B., Mikkelsen, T.S., Kaessmann, H., 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 3, 2179–90.

41. Song, L., Crawford, G.E., 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. 2010, pdb.prot5384.

42. Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.-K., Sheffield, N.C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R.M., Shibata, Y., Showers, K.A., Simon, J.M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N.D., Birney, E., Iyer, V.R., Crawford, G.E., Lieb, J.D., Furey, T.S., 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21, 1757–67.

43. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P. V, Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A.,

Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–31.

44. Treangen, T.J., Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46.
45. Torrents, D., Suyama, M., Zdobnov, E., Bork, P., 2003. A genome-wide survey of human pseudogenes. Genome Res. 13, 2559–67.

46. van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426.

47. Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

48. Welch, J.D., Baran-Gale, J., Perou, C.M., Sethupathy, P., Prins, J.F., 2015. Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. BMC Genomics 16, 113.

49. Wong-Riley, M.T.T., 2012. Bigenomic regulation of cytochrome c oxidase in neurons and the tight coupling between neuronal activity and energy metabolism. Adv. Exp. Med. Biol. 748, 283–304.

50. Yang, C.-C., Buck, M.J., Chen, M.-H., Chen, Y.-F., Lan, H.-C., Chen, J.J.W., Cheng, C., Liu, C.-C., 2013. Discovering chromatin motifs using FAIRE sequencing and the human diploid genome. BMC Genomics 14, 310.

51. Yeung, K.Y., Ruzzo, W.L., 2001. Principal component analysis for clustering gene expression data. Bioinformatics 17, 763–74.

52. Zhang, P., Cong, B., Yuan, H., Chen, L., Lv, Y., Bai, C., Nan, X., Shi, S., Yue, W., Pei, X., 2008. Overexpression of spindlin1 induces metaphase arrest and chromosomal instability. J. Cell. Physiol. 217, 400–8.

53. Zhang, Z., Gerstein, M., 2003. Identification and characterization of over 100 mitochondrial ribosomal protein pseudogenes in the human genome. Genomics 81, 468–480.

54. Zhou, L., Lim, Q.-E., Wan, G., Too, H.-P., 2010. Normalization with genes encoding ribosomal proteins but not GAPDH provides an accurate quantification of gene expressions in neuronal differentiation of PC12 cells. BMC Genomics 11, 75.

55. Zimmer, B., Kuegler, P.B., Baudis, B., Genewsky, A., Tanavde, V., Koh, W., Tan, B., Waldmann, T., Kadereit, S., Leist, M., 2011. Coordinated waves of gene expression during neuronal differentiation of embryonic stem cells as basis for novel approaches to developmental neurotoxicity testing. Cell Death Differ. 18, 383–95.