UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

Implementação de abordagens computacionais para
identificação de RNAs longos não codificadores
envolvidos na diferenciação neural

Gabriel Francisco Zaniboni

SÃO PAULO

2015

Gabriel Francisco Zaniboni
Bacharel em Ciências Biológicas

Implementação de abordagens computacionais para
identificação de RNAs longos não codificadores
envolvidos na diferenciação neural

Dissertação apresentada ao Instituto de
Matemática e Estatística da Universidade de
São Paulo para obtenção do título de Mestre
em Bioinformática.

Área de Concentração: Bioinformática

Orientador:

Prof. Dr. Eduardo Moraes Rego Reis

Co-orientador:

Prof. Dr. Alan Mitchel Durham

São Paulo

2015

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Nome: ZANIBONI, Gabriel Francisco

Título: Implementação de abordagens computacionais para identificação de RNAs longos não codificadores envolvidos na diferenciação neural

Dissertação apresentada ao
Instituto de Matemática e
Estatística da Universidade de São
Paulo para obtenção do título de
Mestre em Bioinformática

Aprovado em:

Banca Examinadora

Prof. Dr. _____ Instituição:

Julgamento: _____
Assinatura: _____

Prof. Dr. _____ Instituição:

Julgamento: _____
Assinatura: _____

Prof. Dr. _____

Instituição:

Julgamento: _____

Assinatura: _____

Dedico este trabalho a meu pai Ângelo, minha mãe Isabel e meu irmão Lucas, pelo apoio ao longo desses anos, imprescindíveis para o término deste trabalho.

AGRADECIMENTOS

Ao Prof. Dr. Eduardo Moraes Rego Reis, pela confiança em mim depositada para a realização deste trabalho e sua disponibilidade e apoio durante sua orientação.

Ao Prof. Dr. Alan Mitchel Durham, pelos conhecimentos em Ciências da Computação ao longo das reuniões que participei, de grande importância para o desenvolvimento deste trabalho.

Ao Prof. Dr. Sergio Verjovski-Almeida, pela sua contribuição na forma de comentários e críticas ao longo das apresentações dos resultados dessa pesquisa.

À Dra. Ana Carolina Ayupe, pela ajuda constante e de grande valia no decorrer deste trabalho.

À Fundação de Amparo à Pesquisa do Estado de São Paulo, pela concessão da bolsa de mestrado e pelo apoio financeiro para a realização desta pesquisa.

À Ana Tahira, pela ajuda com as análises finais desse projeto.

Aos alunos do Prof. Dr. Eduardo (ReisLab) e do Prof. Dr. Sergio Verjovski-Almeida (VerjoLab), pelas críticas e dicas que recebi nas quartas-feiras em que apresentei meus seminários de resultados.

À Patrícia Martorelli, pela gentileza, ajuda indispensável e disponibilidade nesses anos de pós-graduação, nos momentos mais burocráticos.

À Mariana L. Peres Carvalho, pela sua amizade e companhia no início de minha vida em São Paulo.

Aos alunos do programa de bioinformática Lucas Ferreira e Diego Trindade de Souza, pelos bons momentos que passamos fora do laboratório, nos tempos em que ainda haviam festividades na USP e gummy bears eram permitidos.

Aos meus amigos Victor Augusti Negri, Nádia Maria Vieira Sampaio, Luiz Filipe Bartoleti, Rafael Kenji Murayama e Vinícius Carvalho, que apesar da distância que nos encontramos hoje, nunca deixaram de me apoiar e me ouvir chorar as pitangas, viverem momentos inesquecíveis no maior espetáculo da Terra, lembremos deles ou não, me acompanharem nos melhores concertos musicais que a raça humana já foi capaz de produzir e de sempre me fazerem sentir quando nos encontramos que nada em nossa amizade mudou, independentemente do tempo que passamos longe.

Aos meus amigos Fernando Issamu Suzuki e Laura Migliari Branco, pelos anos de convívio na Unicamp e agora em São Paulo, nos momentos em que há um horário disponível em suas agendas.

“We are just an advanced breed of monkeys on a minor planet of a very average star. But we can understand the Universe. That makes us something very special.”

– Stephen Hawking

RESUMO

Cada vez mais, RNAs longos não codificadores (lncRNAs) surgem como importantes reguladores da biologia celular, principalmente em processos de diferenciação durante o desenvolvimento. O interesse no estudo das funções e mecanismos de atuação dessa classe de transcritos durante esses processos é crescente, e mostra-se bastante relevante no processo de diferenciação neural, pelo qual são gerados neurônios e células da glia. A linhagem celular P19, uma célula pluripotente advinda de um tipo de carcinoma embrionário murino, é bem consolidada como modelo *in vitro* de diferenciação neural. Após tratamento com ácido retinóico, ela é capaz de se diferenciar em neurônios e células da glia (astrócitos e oligodendrócitos). Em busca de evidências que indiquem a atuação de lncRNAs durante o processo de diferenciação neural, nosso grupo realizou experimentos utilizando microarranjos para averiguar os níveis de expressão gênica de lncRNAs e genes codificadores de proteínas (mRNAs) durante a diferenciação de células P19 em neurônios (predominância após 10 dias de diferenciação) e glia (predominância em 14 dias de diferenciação). Em um primeiro momento foi realizada a reanotação das sondas referentes a esses lncRNAs da plataforma de microarranjo, visto que as informações presentes nos arquivos de anotação da mesma eram muito escassas e desatualizadas. Registros de lncRNAs e mRNAs foram obtidos a partir de bancos de dados públicos para esse fim, e ao final dessa etapa aproximadamente 25,0% das sondas que não tinham uma anotação foram reanotadas com identificadores advindos desses bancos de dados. A partir dos dados de expressão, foram identificados todos os lncRNAs e mRNAs que apresentaram expressão diferencial entre as diferentes condições estudadas. As informações dos mRNAs diferencialmente expressos foram então utilizadas para a realização de análises de enriquecimento de categorias gênicas do Gene Ontology, nas ontologias de processo biológico e função molecular. A partir das sondas reanotadas, foram realizadas análises de coexpressão entre lncRNAs e mRNAs. A partir do cruzamento das informações obtidas, foram selecionados lncRNAs que através dos princípios de *guilt by association* se mostraram propensos a desempenharem um papel regulatório na diferenciação neural. Assim, as informações geradas nesse trabalho servirão como base para estudos futuros de validação funcional desses lncRNAs.

Palavras-chave: microarranjo, RNA longo não codificador, diferenciação neural, ontologia gênica.

ABSTRACT

Increasingly, long noncoding RNAs (lncRNAs) emerge as important regulators of cell biology, especially in differentiation processes during development. The interest in the study of functions and mechanisms of action of this class of transcripts during these processes is growing, and shows quite relevant in the neural differentiation process by which neurons and glia are generated. The P19 cell line, pluripotent cells arising from a type of murine embryonal carcinoma, is well established as an *in vitro* model of neural differentiation. After treatment with retinoic acid, it is capable of differentiating into neurons and glial cells (astrocytes and oligodendrocytes). In search of evidence that indicate the action of lncRNAs during the neural differentiation process, our group conducted experiments using microarrays to assess gene expression levels of lncRNAs and protein coding genes (mRNAs) during differentiation of P19 cells into neurons (mainly after 10 days of differentiation) and glial cells (mainly after 14 days of differentiation). At first was performed the reannotation of the probes relating to these microarray's lncRNAs, as the information provided in the annotation files were very scarce or outdated. LncRNAs and mRNAs records were obtained from public databases for this purpose, and at the end of this stage approximately 25.0% of the probes without annotation were reannotated with identifiers arising from these databases. From the expression data, we identified all lncRNAs and mRNAs that showed differential expression between the different studied conditions. The information of differentially expressed mRNAs were then used to perform Gene Ontology enrichment, in the ontologies biological process and molecular function. From the reannotated probes, coexpression analyses were performed for lncRNAs and mRNAs. From the crosscheck of information obtained, we selected those lncRNAs that by the principles of guilt by association proved likely to play a regulatory role in neural differentiation. Thus, the information generated in this study will serve as a basis for future studies of functional validation of these lncRNAs.

Keywords: microarray, long noncoding RNA, neural differentiation, gene ontology.

SUMÁRIO

INTRODUÇÃO.....	1
1. RNAs LONGOS NÃO CODIFICADORES	1
2. IMPORTÂNCIA DOS LNCRNAs EM PROCESSOS DE DIFERENCIAÇÃO CELULAR	3
3. PADRÃO DE COEXPRESSION INFERIDO POR “GUILT-BY-ASSOCIATION” COM RNAs CODIFICADORES CARACTERIZADOS	5
OBJETIVOS	7
1. OBJETIVOS GERAIS	7
2. OBJETIVOS ESPECÍFICOS	7
MÉTODOS.....	8
1. REANOTAÇÃO DAS SONDAS DE LNCRNAs	8
2. ANÁLISE DE EXPRESSÃO	8
3. ANÁLISE DE ENRIQUECIMENTO DE <i>GENE ONTOLOGY</i>	8
4. CORRELAÇÃO DE PEARSON E TESTE DE <i>BOOTSTRAP</i>	9
5. IDENTIFICAÇÃO DE POSSÍVEIS LNCRNAs CANDIDATOS A VALIDAÇÃO FUNCIONAL	9
RESULTADOS E DISCUSSÃO	10
1. REANOTAÇÃO DAS SONDAS DE LNCRNAs	10
1.1 <i>BLAT</i>	12
1.2 <i>BEDTools</i>	13
1.3 Atualização das sondas de lncRNAs anotadas no microarranjo.....	18
1.4 Categorias de lncRNAs resultantes da reanotação e atualização.....	18
2. ANÁLISE DA EXPRESSÃO GÊNICA DURANTE A DIFERENCIAÇÃO DE CÉLULAS P19	19
2.1 Detecção de mRNAs e lncRNAs expressos	19
2.2 Identificação de mRNAs e lncRNAs diferencialmente expressos	20
3. ANÁLISE DE ENRIQUECIMENTO DE CATEGORIAS GO	27
3.1 <i>GSEA</i>	27
3.2 <i>GO-Function</i>	31
4. CORRELAÇÃO DE PEARSON E TESTE DE <i>BOOTSTRAP</i>	32
5. IDENTIFICAÇÃO DE POSSÍVEIS LNCRNAs CANDIDATOS A VALIDAÇÃO FUNCIONAL	33
5.1 <i>Cbln1</i>	38
5.2 <i>Igf2</i>	38
5.3 <i>Lsmp</i>	39
5.4 <i>Pou3f3</i>	39
5.4 Possíveis lncRNAs regulatórios em trans.....	40
CONCLUSÃO.....	43
ANEXOS	45
1. RESULTADOS PRELIMINARES	45

2.	IMUNOCITOQUÍMICA	49
3.	CITOMETRIA DE FLUXO.....	49
REFERÊNCIAS BIBLIOGRÁFICAS.....		53

INTRODUÇÃO

1. RNAs longos não codificadores

RNAs longos não codificadores (lncRNAs, *long non-coding RNAs*) são assim classificados por serem transcritos maiores do que 200 nucleotídeos e por possuírem um pequeno ou nenhum potencial de conter informação genética para síntese de proteínas. O conhecimento acerca dos lncRNAs tem crescido cada vez mais em anos recentes (Kung *et al.* 2013, Quinn *et al.* 2014, Iyer *et al.* 2015), transcritos estes que compõem uma parte significativa do transcrito de eucariotos superiores (Carninci *et al.* 2005, Kapranov *et al.* 2007). A comparação entre lncRNAs e outros dados do projeto ENCODE (*The Encyclopedia of DNA Elements*) (Encode Project Consortium 2004) indicou que os lncRNAs são gerados a partir de uma via similar àquela de genes codificadores de proteínas (Encode Project Consortium 2012). A partir da identificação das regiões onde foi detectada a atividade de RNAs, constatou-se que 62,0% dessas bases genômicas estão representadas em moléculas de RNA longas (maiores do que 200 nucleotídeos) sequenciadas ou em éxons do GENCODE; dessas bases, apenas 5,5% são explicadas pelos éxons do GENCODE. A maioria das bases transcritas estão entre ou sobrepostas a fronteiras de genes anotados (ex. intrônicos) e apenas 31,0% das bases nos transcritos sequenciados eram intergênicos (Djebali *et al.* 2012).

Uma grande quantidade desses transcritos também foi identificada no genoma de camundongo, no projeto FANTOM3 (*Functional Annotation of Mammalian cDNA*). Aproximadamente 35.000 transcritos não codificadores originados a partir de 10.000 *loci* distintos que apresentam sinais característicos de RNAs mensageiros (mRNAs) transcritos pela RNA polimerase II, como cap 5', *splicing* e poliadenilação, mas que na maior parte não possuem fases abertas de leitura (Carninci *et al.* 2005).

A quantidade de lncRNAs caracterizados funcionalmente vem crescendo continuamente graças aos avanços nesse campo de estudo da biologia molecular. As evidências recentes têm revelado que lncRNAs participam de numerosos processos biológicos que coordenam a expressão gênica, em todos os níveis de regulação (Bhan & Mandal 2014, Holoch & Moazed 2015), tais como modificação epigenética (Wang & Chang 2011, Nazer & Lei 2014, Spadaro *et al.* 2015), controle da transcrição (Orom *et al.* 2010, Pefanis *et al.* 2015) ou tradução (Gong & Maquat 2011, Bazin & Bailey-Serres 2015) e *imprinting* genômico (Pandey *et al.* 2008, Mohammad *et al.* 2010, Kanduri 2015). Os mecanismos moleculares pelos quais os lncRNAs exercem suas

funções regulatórias ainda não são completamente entendidos ou conhecidos, mas podem ser divididos em quatro principais arquétipos de acordo com as informações conhecidas acerca dos lncRNAs já caracterizados: sinalizadores, “esponjas” (sequestram proteínas limitando sua atuação), guias e arcabouços (Wang & Chang 2011).

A classificação por função não é a única forma de categorizar os lncRNAs. A classificação mais antiga e mais comumente utilizada é a classificação quanto ao *locus* genômico de onde se origina o transcrito em relação a genes codificadores de proteínas conhecidos. De acordo com essa classificação, um lncRNA pode ser intergênico (quando nenhuma parte se sobrepõe a um éxon de genes codificadores de proteína), intrônico (origina-se de um íntron de genes codificadores) ou antisenso (sobrepõem um gene codificador na fita oposta) (St Laurent *et al.* 2015). A disposição dos genes de acordo com essa classificação pode ser vista na figura 1.

Por possuírem mecanismos tão amplos de atuação, os lncRNAs influenciam dramaticamente a biologia celular, e cada vez mais evidencia-se a importância deles em doenças como câncer (Prensner & Chinnaiyan 2011, Tahira *et al.* 2011, Reis & Verjovski-Almeida 2012, Spizzo *et al.* 2012) e no desenvolvimento (Mattick 2001, Amaral & Mattick 2008, Wang *et al.* 2011).

Figura 1: classificação utilizada nesse trabalho quanto a localização genômica dos lncRNAs. Genes codificadores de proteínas e seus éxons estão representados em azul, enquanto os lncRNAs e seus éxons estão representados em vermelho. **A)** lncRNA intergênico, transcrito entre genes de ambas as fitas. **B)** lncRNA intrônico, transcrito inteiramente de íntrons de genes codificadores de proteínas. **C)** lncRNA antisense, transcrito a partir da fita antisense de genes codificadores, sobrepondo regiões exônicas ou intrônicas ou cobrindo toda sua extensão dentro de um íntrons. Adaptado de (Ma *et al.* 2013).

2. Importância dos lncRNAs em processos de diferenciação celular

lncRNAs guias podem se associar fisicamente a proteínas regulatórias da cromatina promovendo seu recrutamento para regiões específicas do genoma e assim modular a expressão gênica (Khalil *et al.* 2009, Koziol & Rinn 2010, Beckedorff *et al.* 2013). Também foi documentado que lncRNAs podem ser regulados por fatores de transcrição cruciais em processos celulares tais como pluripotência e proliferação (Ponjavic *et al.* 2007, Khalil *et al.* 2009, Guttman *et al.* 2011).

Cada vez mais estudos têm demonstrado que os lncRNAs têm funções na embriogênese e em processos de desenvolvimento (Amaral & Mattick 2008, Dinger *et al.* 2008, Sigova *et al.* 2013). As diferenças encontradas (tamanho, formato, potencial de membrana, atividade metabólica, respostas a sinais, que em conjunto definem a

função da célula) entre as células menos especializadas (células tronco, CT) e as mais complexas (células diferenciadas) que são originadas pelo processo de diferenciação celular se devem principalmente a alterações altamente controladas na expressão gênica. No trabalho realizado por Guttman e colaboradores, foi demonstrado que o silenciamento transiente da maioria dos lncRNAs expressos em células tronco embrionárias (CTE) de camundongo resultou em um grande impacto no padrão de expressão gênica dessas células, comparável em magnitude ao de proteínas regulatórias bem conhecidas para esse tipo celular. Nesse estudo, foram identificados lncRNAs responsáveis pela manutenção do estado pluripotente da célula e outros que atuam reprimindo programas de expressão linhagem-específicos (Guttman *et al.* 2011).

Existe grande interesse em investigar os processos de diferenciação neural e o papel dos lncRNAs neste processo. A maior parte dos trabalhos visando o entendimento dos papéis regulatórios exercidos por lncRNAs na pluripotência e desenvolvimento neural utilizaram o camundongo como modelo (Dinger *et al.* 2008, Tochitani & Hayashizaki 2008, Mercer *et al.* 2010, Goff *et al.* 2015). Em um outro estudo utilizando uma população de células progenitoras neurais humanas evidenciaram-se lncRNAs indispensáveis para a diferenciação neuronal (Ng *et al.* 2012).

Durante o desenvolvimento neural, uma população relativamente pequena de CTs dá origem ao complexo sistema nervoso central (SNC). Essa população de CTs é composta por células multipotentes auto renovadoras capazes de se diferenciar em uma variedade de tipos celulares neuronais e gliais de maneira dependente de sua localização e tempo do desenvolvimento em que se originaram (Grabel 2012). A célula de carcinoma embrionário murino P19 é pluripotente e tem sido extensivamente validada como um modelo de diferenciação neural *in vitro* através da indução por ácido retinóico (Jones-Villeneuve *et al.* 1982, McBurney *et al.* 1982) (figura 2). Proteínas com expressão restrita a células P19 indiferenciadas (ex. receptor para endotelina B, receptor da substância P) (Ulrich & Majumder 2006), ou diferenciadas em tipos distintos de células de origem neural (ex. neurônio: enolase neuronal; glia: GFAP) podem ser utilizados como marcadores moleculares de diferenciação (Staines *et al.* 1994, Monge *et al.* 1995, Niemann & Schaller 1996).

Após 6 dias de diferenciação, 85,0% das células em cultura expressam marcadores neuronais, como proteínas do neurofilamento de 68 e 160 kD e o antígeno HNK-1. A proporção de células neuronais declina com o tempo porque os neurônios são pós-mitóticos e as células não neuronais continuam a proliferar. Em torno do dia 10,

células da astroglia surgem e podem ser reconhecidas por seus filamentos intermediários, compostos de GFAP (McBurney 1993).

As alterações no padrão de expressão gênica que ocorrem durante a diferenciação de P19 em neurônios ou em glia são ainda desconhecidas.

Figura 2 - Modelo de diferenciação neuronal in vitro da célula de carcinoma embrionário murino P19, até o dia 14. O tratamento com ácido retinóico leva à diferenciação em neurônios (predominam na cultura após 10 dias de tratamento; detectado em 80,0% das células) e células da glia (predominância em 14 dias; detectado em 55,0% das células). A curva em preto representa o total de células na cultura.

3. Padrão de coexpressão inferido por “*guilt-by-association*” com RNAs codificadores caracterizados

Com o conhecimento de milhares de *loci* de lncRNAs, o desafio que surge é a determinação das funções carreadas por esses lncRNAs. Um primeiro passo na formulação de hipóteses consiste em utilizar-se padrões de expressão desses lncRNAs para a identificação de tipos celulares ou processos biológicos associados com cada um deles. Os primeiros estudos a respeito desses transcritos identificaram lncRNAs que são altamente expressos em certas regiões do cérebro. Hibridizações *in situ* confirmaram

esses padrões de expressão rebuscados em subestruturas específicas do cérebro (Mercer *et al.* 2008). Um estudo similar do mesmo grupo identificou numerosos lncRNAs que são fortemente correlacionados a fatores de transcrição associados a pluripotência, sugerindo que muitos deles podem atuar em redes transcricionais de CTEs (Dinger *et al.* 2008).

Após esses estudos, um método informático denominado “culpa por associação” (*guilt by association*) permitiu um entendimento global de lncRNAs e genes codificadores de proteína que são fortemente coexpressos, logo, presumivelmente coregulados (Guttman *et al.* 2009). Esse método identifica genes codificadores de proteína e vias significativamente correlacionadas com um dado lncRNA utilizando análises de expressão gênica. Desse modo, baseado em funções conhecidas dos genes codificadores coexpressos, hipóteses podem ser formuladas a respeito das funções e potenciais reguladores desses lncRNAs candidatos. Essa abordagem predisse diversas funções de lncRNAs, desde pluripotência de CTs até câncer (Guttman *et al.* 2009). Por exemplo, numerosos lncRNAs que estavam fortemente correlacionados com a proteína p53 foram induzidos de uma maneira p53-dependente, muito mais do que seria esperado ao acaso (Khalil *et al.* 2009, Huarte *et al.* 2010). Esses lncRNAs também demonstraram enriquecimento para motivos de ligação a p53 em seus promotores. Do mesmo modo, vários lncRNAs com funções preditas relacionadas ao processo de adipogênese e pluripotência foram comprovados como essenciais na manutenção desses estados celulares (Sun *et al.* 2013).

Mesmo com o conhecimento de que os lncRNAs possuem papéis regulatórios na diferenciação neural, não se sabe quais os padrões de expressão desse tipo de molécula durante esse processo nas células P19 submetidas ao tratamento com ácido retinóico. Tendo essa questão em vista, o presente trabalho tem por objetivo principal a investigação das alterações nos níveis de expressão de lncRNAs observadas durante a diferenciação, na busca de possíveis lncRNAs regulatórios. Os trabalhos aqui mencionados fizeram uso de uma técnica ou abordagem específica a cada um deles, fator esse que pode limitar o potencial de descoberta e identificação de lncRNAs funcionais devido aos limites inerentes a cada uma delas. Nesse trabalho buscou-se mesclar mais de um tipo de abordagem (como expressão diferencial, padrão de coexpressão com RNAs codificadores centrais ao processo de diferenciação neural, enriquecimento de processos celulares) para a obtenção de resultados mais robustos.

OBJETIVOS

1. Objetivos gerais

- (i) Implementação de uma plataforma informática para a análise de dados de expressão gênica em larga escala com foco na identificação e anotação de possíveis lncRNAs regulatórios;
- (ii) Identificação e caracterização de mRNAs e lncRNAs diferencialmente expressos durante o processo de diferenciação neural em um modelo celular murinho.

2. Objetivos específicos

- (i) Desenvolvimento de rotinas informáticas para o processamento de dados de expressão gênica obtidos através de microarranjos de DNA contendo sondas para mRNAs e lncRNAs (Agilent SurePrint G3 Mouse GE 8x60k Microarray);
- (ii) Anotação genômica dos lncRNAs representados no microarranjo;
- (iii) Identificação de mRNAs e lncRNAs diferencialmente expressos durante a diferenciação de células P19 em neurônios ou glia.
- (iv) Identificação de possíveis lncRNAs regulatórios do processo de diferenciação neural candidatos a validação funcional.

MÉTODOS

1. Reanotação das sondas de lncRNAs

Para a reanotação das sondas de lncRNAs que continham pouca ou nenhuma informação nas especificações do microarranjo, primeiramente as sondas inespecíficas (múltiplos alinhamentos no genoma de camundongo) foram removidas com a utilização do BLAT (Kent 2002). Com as restantes, a ferramenta BEDTools (Quinlan & Hall 2010) foi usada para verificar a sobreposição dessas sondas com registros de lncRNAs já anotados depositados nos seguintes bancos de dados: GENCODE M4 (Harrow *et al.* 2006, Harrow *et al.* 2012), NONCODE v4.0 (Xie *et al.* 2014) e UCSC (Rosenbloom *et al.* 2015). As sondas que apresentaram sobreposição receberam os identificadores dos lncRNAs pertinentes. As sondas de lncRNAs que já possuíam identificadores do RefSeq (Pruitt *et al.* 2014) foram atualizadas em relação à versão mais recente do banco (Release 72), tendo como referência os registros mais recentes de seus identificadores.

2. Análise de expressão

Através dos dados obtidos com o *software* Agilent Feature Extraction, foram consideradas expressas as sondas que apresentaram positivo o filtro *well above background* em pelo menos 3 das 4 réplicas do experimento. Nos casos de sondas com réplicas técnicas dentro do microarranjo, uma média das intensidades foi obtida a partir das sondas expressas. Aos genes de mRNAs e lncRNAs representados por 2 ou mais sondas diferentes foi atribuída a média da intensidade dessas sondas.

Para a identificação de mRNAs e lncRNAs diferencialmente expressos, foi utilizado o pacote “samr” no ambiente R, que implementa o método SAM (*Significance Analysis of Microarrays*) (Tusher *et al.* 2001). Foram considerados diferencialmente expressos os genes que apresentaram variação significativa ($q\text{-value} < 0,01$).

3. Análise de enriquecimento de Gene Ontology

As análises de enriquecimento de GO (ontologia gênica, *gene ontology*) (Ashburner *et al.* 2000) dos genes diferencialmente expressos foram realizadas com a ferramenta GSEA (*Gene Set Enrichment Analysis*) (Subramanian *et al.* 2005) para as células P19 diferenciadas após 10 dias e diferenciadas após 14 dias.

Através do GO-Function, foram selecionados todos aqueles lncRNAs diferencialmente expressos e para cada um deles foi determinado um conjunto de

mRNAs expressos correlacionados ($r \geq 0,8$ ou $r \leq -0,8$, $p\text{-value} < 0,05$), e então foram determinadas as categorias GO enriquecidas nesses conjuntos.

4. Correlação de Pearson e teste de *bootstrap*

Para a identificação de lncRNAs de interesse, foram calculados coeficientes de correlação de Pearson entre cada lncRNA diferencialmente expresso e mRNA expresso. Para tal, foram utilizadas as intensidades obtidas das células P19 indiferenciadas, diferenciadas após 10 dias e diferenciadas após 14 dias. Para cada coeficiente calculado, foi realizado um teste de *bootstrap*, onde foram realizadas 1000 permutações dentre as intensidades observadas dos mRNAs em cada tempo (indiferenciada, 10 dias e 14 dias), gerando um $p\text{-value}$ para cada coeficiente de um par lncRNA – mRNA (mais detalhes na seção Resultados e Discussão).

5. Identificação de possíveis lncRNAs candidatos a validação funcional

A identificação de possíveis lncRNAs candidatos a validação funcional foi realizada a partir dos resultados da análise de enriquecimento de GO. A partir dos resultados obtidos com o GSEA, foram selecionados os mRNAs que contribuíram para o enriquecimento central (*core enrichment*) das categorias GO relacionadas à diferenciação neural cujo FDR (*false discovery rate*) foi menor do que 5,0%. Foram selecionados os lncRNAs diferencialmente expressos cujo coeficiente de correlação de expressão (r) com esses mRNAs era maior ou igual a 0,8 ou menor ou igual a -0,8, com $p\text{-value}$ no teste de *bootstrap* menor do que 0,05. Os lncRNAs obtidos dessa forma e cujo resultado de enriquecimento através do GO-Function compartilhava no mínimo 1 termo GO relacionado a diferenciação neural com os termos resultantes do GSEA foram selecionados. Através dos dados de variação nos níveis de expressão, proximidade a mRNAs de interesse e grupos de mRNAs correlacionados, foram determinados manualmente aqueles lncRNAs candidatos a validação funcional.

RESULTADOS E DISCUSSÃO

1. Reanotação das sondas de lncRNAs

O conhecimento obtido acerca dos lncRNAs vem aumentando gradativamente graças às novas tecnologias de sequenciamento de alto rendimento e à maior atenção que vários grupos ao redor do mundo vêm dando a essa classe de molécula regulatória. Como resultado disso, as informações a respeito dos lncRNAs estão em constante expansão e atualização, tanto através da descoberta de novas moléculas funcionais quanto da caracterização daquelas já conhecidas. Devido a esse fato, a grande maioria das sondas de lncRNAs presentes no microarranjo utilizado nesse estudo continham pouca ou nenhuma informação a respeito de seus alvos, reflexo do conhecimento existente na época de sua idealização (2007). Apesar de as sondas que indagam mRNAs serem bem definidas e conterem bastante informação nas tabelas de anotação do microarranjo, aquelas que indagam lncRNAs são extremamente pobres.

Com esse cenário em vista, fez-se necessário implementar uma etapa no projeto voltada à reanotação destas sondas pobremente anotadas, com base na posição genômica fornecida nos arquivos de anotação do microarranjo para cada sonda.

A estratégia adotada nessa etapa foi comparar essas sondas, através de sua informação posicional, com informações atualizadas presentes em bancos de dados conhecidos e bem estabelecidos, como o Ensembl (Cunningham *et al.* 2015) e o NONCODE v4.0 (Xie *et al.* 2014), através da ferramenta para análises genômicas BEDTools (Quinlan & Hall 2010). A partir dessas análises, uma sonda de lncRNA poderia ser descartada por se sobrepor a transcritos codificadores de proteínas no sentido senso ou então por estar presente em regiões UTR desse mesmo tipo de transcrito (indicando um possível erro quando se classificou aquele transcrito como um lncRNA), ou então ser reanotada por se sobrepor a um registro anotado de lncRNA presente em algum desses bancos de dados.

Figura 3 (pág. anterior) – Fluxograma detalhando os passos executados para a reanotação das sondas de lncRNAs pobremente anotadas e atualização daquelas previamente anotadas, do microarranjo utilizado nesse trabalho. A circunferência marcada com a letra A indica o conjunto de sondas que permaneceram classificadas como pobremente anotadas devido à inexistência de registros nos bancos de dados utilizados para comparação ou à descontinuação dos lncRNAs previamente anotados.

1.1 BLAT

A partir das sequências das sondas de lncRNAs encontradas nos arquivos de anotação do microarranjo, o primeiro passo da reanotação foi a remoção daquelas sondas que apresentaram complementaridade com mais de uma região do genoma de camundongo (foi utilizado como referência a versão mm9/2007). Para tal, foi utilizado o algoritmo BLAT (*BLAST-Like Alignment Tool*) (Kent 2002), um alinhador de pares de sequências. A partir do resultado do BLAT, foi utilizado um *script* de análise a partir do qual foi possível separar aquelas sondas que resultaram em um único alinhamento (ou seja, apresenta complementaridade com apenas uma região do genoma) daquelas que resultaram em mais de um alinhamento. Além disso, o *script* calcula um valor denominado *score* para cada alinhamento, de acordo com sua cobertura e identidade:

$$Score = \frac{Identidade \times Cobertura}{10}$$

O valor de *score* teve a finalidade de reportar quais alinhamentos foram melhores nos casos onde uma sonda teve múltiplos alinhamentos.

Todas as sondas que se enquadraram na categoria de único alinhamento e que apresentaram uma cobertura de no mínimo 95,0% foram levadas adiante no processo de reanotação. Das que apresentaram múltiplos alinhamentos, apenas algumas sondas foram consideradas nas análises subsequentes:

- Sondas com menos de 5 alinhamentos (optou-se por considerar aquelas sondas que pudessem ser complementares a regiões com poucas repetições no genoma, indicando cópias de genes ou regiões pouco repetitivas);

- Sondas com uma diferença entre o *score* do primeiro e o do segundo alinhamento maior do que 100 (apenas os casos onde um dos alinhamentos era consideravelmente melhor do que os demais foram considerados).

Desse modo, das 16.133 sondas de lncRNAs sem anotação prévia do microarranjo, após a filtragem com a utilização do BLAT, foram removidas 1.519 delas das análises subsequentes (ver figura 3).

1.2 BEDTools

As análises subsequentes da reanotação foram realizadas com a ferramenta BEDTools, que possibilita uma vasta gama de análises tendo como entrada arquivos contendo coordenadas genômicas no formato BED. Mais especificamente, a função do BEDTools utilizada para a análise de sobreposição das sondas com registros dos bancos de dados conhecidos foi a *intersect* (figura 4A). As coordenadas das sondas foram utilizadas para verificar a existência de sobreposição com as coordenadas de conjuntos de comparação obtidos do GENCODE M4, Ensembl, NONCODE v4.0, e UCSC.

Assim como na etapa anterior (BLAT), é importante ressaltar que como as sondas foram desenhadas com base na montagem mm9 do genoma de camundongo, todas as coordenadas dos conjuntos de comparação utilizados na reanotação baseados na montagem mm10 do genoma foram convertidos para a montagem mm9 com o uso da ferramenta liftOver, disponível no website da UCSC (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>).

Foram obtidos arquivos BED de genes e transcritos dos seguintes bancos de dados disponíveis publicamente para que as sondas fossem comparadas:

- GENCODE M4 (6.951 registros BED);
- NONCODE v4.0 (114.103 registros BED);
- Éxons de RNAs codificadores do Ensembl (504.620 registros BED);
- RNAs de sequência completa do UCSC (1.592.416 registros BED);

A partir de outras duas funções do BEDTools, *subtract* e *flank* (figura 4B e 4C), foram obtidas as coordenadas dos íntrons dos RNAs codificadores de proteína do

Ensembl e das regiões flanqueadoras do início e fim desses genes em até 2 kb, levando à obtenção das regiões 5' e 3' UTR:

- Íntrons de RNAs codificadores do Ensembl (560.247 registros BED);
- Regiões 5' e 3' UTR até 2 kb (44.692 registros BED);

Com o intuito de se remover aquelas sondas que podem ter sido erroneamente anotadas como lncRNAs e que na verdade são mRNAs, tanto transcritos maduros quanto não processados, comparou-se as sondas remanescentes da filtragem do BLAT com os registros de éxons e de íntrons de genes codificadores do Ensembl com o BEDTools *intersect*. Em todas as análises de sobreposição desta etapa foi utilizado o parâmetro *force strand* do BEDTools, o que significa que só foi considerado como sobreposição aqueles casos em que tanto a sonda quanto o elemento do conjunto sendo comparado encontravam-se na mesma fita do DNA. Através dessa análise, um total de 2.159 sondas foram excluídas, sendo que 904 delas apresentaram sobreposição com éxons e 1.255 com íntrons (figura 3).

A) Intersect

B) Subtract

C) Flank

Figura 4 – Esquema gráfico da funcionalidade das ferramentas do BEDTools utilizadas na reanotação de sondas de lncRNA. **A)** Intersect – reporta quais elementos de Y apresentaram sobreposição com X; **B)** Subtract – subtrai de X tudo que se encontra em Y; **C)** Flank – reporta as regiões flangeadoras de X utilizando como referência um genoma. X: conjunto de partida, Y: conjunto de comparação; Z: conjunto resultante.

Do mesmo modo, as sondas remanescentes foram comparadas com o conjunto de regiões 5' e 3' UTR. Essas sondas podem ter sido anotadas como lncRNAs, mas na verdade podem representar fragmentos de mRNAs não processados. Um total de 731 sondas apresentaram sobreposição com essas regiões UTR. Entretanto, existem lncRNAs já conhecidos que se localizam em regiões flangeadoras de genes codificadores (Blume *et al.* 2003, Jia *et al.* 2010, Johnsson *et al.* 2013). Devido a esse fato, verificou-se se essas 731 sondas apresentavam sobreposição com o conjunto de lncRNAs anotados do GENCODE M4. Destas, um total de 26 representavam lncRNAs

anotados, e foram reanotadas com seus identificadores. As 705 sondas restantes foram removidas das etapas subsequentes da reanotação (figura 3).

Após as etapas de filtragem, das 16.133 sondas de lncRNAs sendo reanotadas, 11.698 foram mantidas para serem comparadas com registros de lncRNA já anotados. Nos casos onde uma mesma sonda apresentou sobreposição com mais de um transcrito dentre os diferentes conjuntos de anotação, todos os identificadores foram atribuídos àquela sonda.

O primeiro conjunto de lncRNAs anotados utilizado para anotação das sondas remanescentes foi o de lncRNAs do GENCODE M4. O GENCODE é parte do projeto científico ENCODE (*Encyclopedia of DNA Elements*), que tem por objetivo identificar todos os elementos funcionais do genoma humano e de camundongo, através de uma combinação de análises computacionais, anotação manual e validação experimental (Encode Project Consortium 2004, Harrow *et al.* 2006, Qu & Fang 2013). Dentre os conjuntos de comparação utilizados neste trabalho, o GENCODE M4 teve prioridade sobre os demais, devido ao fato de incluir etapas de anotação manual em seu *pipeline* (grupo HAVANA) (Kokocinski *et al.* 2010). Um total de 955 sondas foram reanotadas dessa maneira, e somadas às 26 que foram anotadas do grupo de possíveis regiões UTR, 981 sondas receberam identificadores do GENCODE, representando 288 genes de lncRNAs.

O mesmo conjunto de sondas que não foram removidas pelos filtros anteriores foram comparadas com os 114.103 registros de lncRNAs de camundongo presentes no banco de dados NONCODE, que em janeiro de 2014 teve sua base de dados atualizada para a versão 4.0 (Xie *et al.* 2014). Em sua última versão, o banco de dados especializado em RNAs não codificadores (não inclui tRNAs e rRNAs) agregou a seu acervo informações dos dois anos anteriores advindos de estudos de sequenciamento de alto rendimento (RNA-Seq), além de integrar aqueles lncRNAs presentes no RefSeq e Ensembl. Um total de 2.704 sondas apresentaram sobreposição com os lncRNAs do NONCODE, sendo os identificadores desse banco associados a essas sondas. Desse total de sondas, foram representados 816 genes de lncRNAs.

Dentre os arquivos disponibilizados no website do UCSC (*University of California, Santa Cruz*) estão dados de sequenciamento de RNAs de sequência completa, que foram produzidos como parte do Projeto Transcritoma ENCODE (Ruan & Ruan 2012). Essas sequências foram obtidas através de sequenciamento pareado (RNA-PET) de extratos de RNAs totais de alta qualidade que continham cauda poli-A.

1.3 Atualização das sondas de lncRNAs anotadas no microarranjo

Uma minoria das sondas de lncRNAs do microarranjo (2.012) possuía identificadores do RefSeq que se enquadravam em duas categorias: RNAs não codificadores (NR) e modelos preditos de RNAs não codificadores (XR). Devido à expansão no conhecimento a respeito dos lncRNAs, havia a possibilidade desses registros estarem anotados erroneamente, terem sido alterados ou descontinuados. Devido a esse fato, todas essas sondas foram atualizadas de acordo com os registros mais recentes do RefSeq. As sondas com registros descontinuados foram reclassificadas como pobremente anotadas, enquanto aquelas cujo registro ainda se encontra no RefSeq, atualizadas ou da mesma maneira como anotada no microarranjo, mantiveram-se como sondas anotadas.

De um total de 1.437 sondas NR (representando 1.256 lncRNAs), 1.096 se mantiveram como sondas anotadas (945 lncRNAs). No conjunto de sondas XR, de 575 sondas (representando 537 lncRNAs), 285 se mantiveram como sondas anotadas (260 lncRNAs).

1.4 Categorias de lncRNAs resultantes da reanotação e atualização

Nos bancos de dados GENCODE M4 e NONCODE v4.0, cada lncRNA está classificado de acordo com sua localização genômica. Apesar dessa classificação ainda estar em constante desenvolvimento e mudança (St Laurent *et al.* 2015), os lncRNAs reanotados e atualizados nesse estudo receberam uma de três possíveis classificações: (i) intergênico, (ii) intrônico ou (iii) antisense.

Após as etapas de reanotação de sondas pobremente anotadas e atualização das sondas previamente anotadas no microarranjo, os lncRNAs anotados de se dividiram nessas três categorias da seguinte maneira:

- 1.546 intergênicos;
- 27 intrônicos;
- 298 antisense.

Apesar de conhecidos, alguns desses lncRNAs ainda não possuem uma classificação clara, devido à complexidade inerente de alguns deles. Assim, um total de 87 permaneceram sem classificação.

2. Análise da expressão gênica durante a diferenciação de células P19

Os perfis globais de expressão gênica gerados a partir das células de carcinoma embrionário murino P19 indiferenciadas, após 10 dias de diferenciação neural e após 14 dias de diferenciação neural foram analisados em busca de alterações nos níveis de expressão de mRNAs e lncRNAs. Esses dados foram previamente obtidos no laboratório pela Dra. Ana Carolina Ayupe de Oliveira apresentados como resultados preliminares (ver Anexo I).

No âmbito de sua pesquisa, a Dr^a Ana Carolina Ayupe faz uso de análises de expressão de frações citoplasmáticas e de frações nucleares obtidas a partir de culturas de células P19. Como o foco deste trabalho não inclui analisar diferenças de expressão entre tais compartimentos, esses dados foram utilizados de maneira conjunta, ou seja, as razões de expressão das células diferenciadas em relação às indiferenciadas obtidas no citoplasma e/ou no núcleo foram consideradas como um conjunto único. Algumas análises subsequentes foram realizadas com base em subgrupos do total de genes detectados, os critérios levados em consideração são descritos nas seções pertinentes.

2.1 Detecção de mRNAs e lncRNAs expressos

Após a hibridização e lavagens seguindo o protocolo padrão do fabricante, as lâminas foram lidas em um scanner a laser e as imagens analisadas no programa Feature Extraction (Agilent), que converte os pixels em valores numéricos de intensidade, corrige vieses associados a diferenças de incorporação e emissão de luz dos fluoróforos Cy3 e Cy5, e calcula uma razão da expressão correspondente dos mRNAs e lncRNAs interrogados nas populações de células P19 indiferenciadas ou diferenciadas em 10 ou 14 dias. Para uma sonda ser considerada expressa, o ponto do microarranjo onde ela está localizada deve ter apresentado as seguintes condições em pelo menos 3 das 4 réplicas do experimento:

- (i) Sinal da sonda após subtração da intensidade de fundo é positiva e significativa;
- (ii) Sinal da sonda é 2,6 maior do que o desvio padrão da intensidade de fundo (filtro denominado pelo *software* de extração como *well above background*).

A quantidade de sondas expressas durante o processo de diferenciação neural está evidenciada na tabela 1. Os dados mostrados dizem respeito à união dos dados provenientes das frações citoplasmática e nuclear. Os lncRNAs se encontram divididos entre “pobremamente anotados” (todas as sondas de lncRNAs que não foram reanotadas ou aquelas previamente anotadas pelo microarranjo, mas que foram descontinuadas quando atualizadas) e “anotados” (todas as sondas que receberam um identificador dos bancos de dados mencionados anteriormente ou as sondas previamente anotadas pelo microarranjo que tiveram seus identificadores verificados e atualizados). Por sua vez, os lncRNAs anotados estão subdivididos de acordo com sua categoria gênica de acordo com sua anotação: intergênico, intrônico ou antisense.

2.2 Identificação de mRNAs e lncRNAs diferencialmente expressos

Para a obtenção dos genes diferencialmente expressos, foi utilizado o teste estatístico SAM (*Significance Analysis of Microarrays*) através do pacote “samr” (Tusher *et al.* 2001), implementado no ambiente R. O SAM determina se a variação de expressão de um dado gene é estatisticamente significativa através da realização de um teste t para cada gene. Ele calcula um valor estatístico que mede a força da relação existente entre a variável independente (nesse caso, a variação observada de cada gene) e a variável dependente (nesse caso, os diferentes estados celulares, indiferenciada ou diferenciada). Nesse método, permutações dos dados são usadas para se determinar se a expressão de algum gene é significativa em relação à resposta. A utilização de uma análise baseada em permutações evita a suposição de que as distribuições de genes individuais sejam paramétricas, o que representa uma vantagem desse método em relação a outros que assumem variância igual e/ou independente dos genes (Tusher *et al.* 2001, Zang *et al.* 2007).

A partir dos valores de intensidade de fluorescência obtidos no experimento de microarranjo, foram calculadas as razões na base $\log_2$ entre as duas condições de células diferenciadas (tratamento após 10 e 14 dias) em relação às células indiferenciadas. Esse cálculo é obtido através da divisão da intensidade de uma sonda em uma condição (10 ou 14 dias) com a intensidade na condição indiferenciada, em escala logarítmica (um valor para cada uma das réplicas, no total de 4):

$$Raz\tilde{a}o = \log_2 \left(\frac{intensidade\ na\ c\acute{e}lula\ diferenciada}{intensidade\ na\ c\acute{e}lula\ indiferenciada} \right)$$

A raz\~ao \xe9 utilizada pelo SAM no teste "One class", onde \xe9 testado se o valor de varia\~ao m\~edia de cada gene difere de zero. A raz\~ao tamb\~em \xe9 utilizada para se calcular o valor de varia\~ao dos genes entre as condi\~oes diferenciada e indiferenciada, atrav\~es da potencia\~ao da base 2 com o valor da m\~edia das raz\~oes de determinado gene:

$$Varia\~ao = 2^{\bar{x}\ Raz\tilde{o}es}$$

Foram considerados diferencialmente expressos os mRNAs e lncRNAs que apresentaram um valor de FDR (*false discovery rate*) menor que 1,0%. Assim, foram identificados os conjuntos diferencialmente expressos em cada tempo de diferencia\~ao (10 e 14 dias). Esses resultados s\~ao apresentados na tabela 1.

Tabela 1 – Quantidade de mRNAs e lncRNAs detectados e diferencialmente expressos. Foram considerados expressos os mRNAs e lncRNAs cujos sinais dos *spots* no microarranjo foram positivos, significantes e cuja intensidade foi maior do que 2,6 vezes o desvio padrão da intensidade de fundo em determinada condição (célula P19 indiferenciada, célula diferenciada após 10 dias ou após 14 dias). Os genes diferencialmente expressos foram determinados utilizando-se o SAM, e correspondem às hibridizações das células diferenciadas após 10 dias em relação às indiferenciadas e após 14 dias em relação às indiferenciadas. Os lncRNAs anotados dividem-se de acordo com sua classificação: intergênicos, antisense, intrônicos e sem classificação.

Tipo	Número de genes no microarranjo	Genes detectados			Genes diferencialmente expressos	
		Indiferenciada	10 dias	14 dias	10d / indiferenciada	14d / indiferenciada
mRNAs codificadores de proteína	25620	20117	20352	20335	5646	14683
lncRNAs anotados	1958	1335	1368	1380	316	338
○ <i>Intergênico</i>	1546	1050	1075	1081	247	260
○ <i>Antisense</i>	298	197	202	209	49	29
○ <i>Intrônico</i>	27	23	24	24	7	7
○ <i>Sem classificação</i>	87	65	67	66	13	42
ncRNAs pobremente anotados	4578	3826	3889	3878	838	1856
Total	32156	25278	25609	25593	6800	16877

Nas análises de coexpressão subsequentes (ver seção 4 e 6), foram utilizados subconjuntos de mRNAs e lncRNAs diferencialmente expressos ao invés do conjunto total (que inclui a divisão entre fração citoplasmática e nuclear). Devido à grande semelhança nos padrões de expressão observados entre as duas frações celulares, optou-se por utilizar o conjunto de mRNAs diferencialmente expressos em pelo menos um dos tempos sendo analisados (10 ou 14 dias), em qualquer uma das frações (núcleo e/ou citoplasma). Já os lncRNAs utilizados nas análises de coexpressão e seleção de

possíveis candidatos a validação funcional foram aqueles diferencialmente expressos em pelo menos um dos tempos de diferenciação (10 e/ou 14 dias) na fração nuclear, visto que lncRNAs que regulam a expressão gênica ao nível da transcrição são mais comumente encontrados no núcleo (Bernard *et al.* 2010). O painel mostrado na figura 6 corresponde a esses subconjuntos.

Figura 6 – Paineis contendo as informações sobre os mRNAs e lncRNAs diferencialmente expressos. **A)** Quantidades de mRNAs que tiveram variação de expressão positiva (*upregulated*) ou negativa (*downregulated*) nas células P19 diferenciadas após 10 dias em relação às indiferenciadas e 14 dias em relação às indiferenciadas, considerando as frações citoplasmática e nuclear. **B)** Como em A, mas para os lncRNAs anotados e os lncRNAs pobremente anotados, considerando apenas a fração nuclear. **C)** Porcentagem dos mRNAs detectados na condição indiferenciada ou diferenciada (10 ou 14 dias) que foram diferencialmente expressos, fração

citoplasmática ou nuclear. **D)** Como em C, mas para os lncRNAs anotados e lncRNAs pobremente

(cont.) anotados, fração nuclear. **E)** Gráfico de Venn mostrando a sobreposição entre mRNAs diferencialmente expressos após 10 dias e 14 dias em relação às células indiferenciadas, fração citoplasmática ou nuclear. **F)** Como em E, mas para os lncRNAs anotados e lncRNAs pobremente anotados, fração nuclear. **G)** *Heatmap* das razões das intensidades (célula diferenciada / indiferenciada) dos mRNAs e lncRNAs detectados em ao menos uma das condições estudadas (são mostradas 4 réplicas, indicada pelo número no final do identificador). As razões encontram-se na escala $\log_2$. É possível observar uma clara distinção entre os padrões de expressão em 10 dias/indiferenciada e 14 dias/indiferenciada. *Ratio*: razão, C: fração citoplasmática, N: fração nuclear, vermelho: variação positiva, azul: variação negativa.

3. Análise de enriquecimento de categorias GO

Os valores de variação de expressão obtidos na análise de genes diferencialmente expressos foram também utilizados na análise de enriquecimento de categorias gênicas de GO. Essa análise foi realizada através do uso de duas ferramentas:

(i) O método computacional Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.* 2005);

(ii) A ferramenta empregada no ambiente R denominada GO-Function (Wang *et al.* 2012).

Os usos separados dessas duas ferramentas tiveram objetivos específicos, como discutido nas seções a seguir.

3.1 GSEA

O GSEA é um método computacional que determina se um conjunto de genes definidos *a priori* demonstra significância estatística de acordo com diferenças observadas entre dois estados biológicos distintos (como por exemplo, células diferenciadas e células indiferenciadas) (Mootha *et al.* 2003, Subramanian *et al.* 2005).

O objetivo ao se utilizar o GSEA foi o de caracterizar quais processos biológicos estavam mais representados tendo como referência os genes diferencialmente expressos em cada um dos experimentos de microarranjo (células diferenciadas após 10 dias em relação às células indiferenciadas e após 14 dias em relação às indiferenciadas). Os conjuntos de genes codificadores de proteínas representativos das categorias GO relacionadas à diferenciação neural resultantes dessa análise serviram como referência na busca de lncRNAs relevantes durante esse processo (ver seção 5), além de corroborarem o fato de que as células P19 foram de fato diferenciadas em células do sistema nervoso.

Os conjuntos de genes definidos *a priori* utilizados nas análises de GSEA foram obtidos no MSigDB (*Molecular Signature Database*) (Subramanian *et al.* 2005):

- (i) C5.bp (que abrange a categoria de processos biológicos, *GO biological process*) e;
- (ii) C5.mf (categoria de funções moleculares, *GO molecular function*).

O tipo de teste utilizado foi o GseaPreranked, através do qual uma lista ordenada pelos valores de variação de expressão de cada gene codificador nas diferentes etapas do processo de diferenciação neural foi usada como critério para se determinar as categorias enriquecidas. Os lncRNAs não fizeram parte dessa lista porque os conjuntos gênicos do MSigDB fazem uso da anotação disponíveis dos genes, que em sua grande maioria corresponde a genes codificadores (Subramanian *et al.* 2005).

Para cada conjunto de genes analisados pelo GSEA, é calculado um valor de enriquecimento (ES, *enrichment score*) que reflete o quanto esse conjunto está sobre-representado em seus extremos (genes que mais variaram positivamente e aqueles que mais variaram negativamente) de toda a lista de genes ordenados. Para tal, o ES é calculado descendo a lista, aumentando uma estatística somatória quando um gene do conjunto sendo analisado é encontrado. O ES é o máximo desvio de zero encontrado dessa forma, que corresponde a uma estatística assemelhada ao teste de Kolgomorov-Smirnov. O valor de ES é então normalizado de acordo com o tamanho do conjunto de genes, o que gera resulta em um NES (*normalized enrichment score*) (Hollander & Wolfe 1999, Subramanian *et al.* 2005). Por exemplo, conjuntos de genes que representam o termo GO “diferenciação neuronal” tiveram uma expressão maior em relação à observada nas células indiferenciadas; desse modo, a categoria “diferenciação neuronal” mostrou-se enriquecida nas células diferenciadas.

A tabela 2 mostra os termos GO (processo biológico) mais enriquecidos na população de células P19 diferenciadas após 10 dias, utilizando-se uma lista dos genes comuns diferencialmente expressos tanto na fração nuclear quanto na citoplasmática. Os termos estão ordenados por NES, e são mostrados apenas aqueles com FDR < 0,05.

Tabela 2 - Enriquecimento de termos GO (processo biológico) – 10 dias.

Termo GO	NES	FDR
NERVOUS SYSTEM DEVELOPMENT	2,24	0,008
CENTRAL NERVOUS SYSTEM DEVELOPMENT	2,23	0,005
MULTICELLULAR ORGANISMAL DEVELOPMENT	2,16	0,011
SYSTEM DEVELOPMENT	2,04	0,041
TRANSCRIPTION FROM RNA POLYMERASE II PROMOTER	2,0	0,046
ANATOMICAL STRUCTURE DEVELOPMENT	1,98	0,049

A mesma análise foi realizada com os genes diferencialmente expressos após 14 dias de diferenciação, e as categorias GO enriquecidas nas células diferenciadas são mostradas na tabela 3.

Tabela 3 - Enriquecimento de termos GO (processo biológico) – 14 dias.

Termo GO	NES	FDR
NERVOUS SYSTEM DEVELOPMENT	2,91	0
SYNAPTIC TRANSMISSION	2,48	0,001
TRANSMISSION OF NERVE IMPULSE	2,48	0,001
SYSTEM DEVELOPMENT	2,47	0
CENTRAL NERVOUS SYSTEM DEVELOPMENT	2,47	0
MULTICELLULAR ORGANISMAL DEVELOPMENT	2,46	0
SYNAPTOGENESIS	2,37	0,003
EXTRACELLULAR STRUCTURE ORGANIZATION AND BIOGENESIS	2,35	0,004
SYNAPSE ORGANIZATION AND BIOGENESIS	2,34	0,005
ANATOMICAL STRUCTURE DEVELOPMENT	2,33	0,005
NEUROLOGICAL SYSTEM PROCESS	2,31	0,006
BRAIN DEVELOPMENT	2,18	0,036

Visualizando as tabelas com os resultados de enriquecimento das células diferenciadas após 10 e 14 dias, pode-se notar que a maioria dos termos enriquecidos em 10 dias também foram enriquecidos em 14 dias (com a exceção dos termos “*central nervous system development*” e “*transcription from RNA polymerase II promoter*”). Os termos que aparecem apenas na análise de 14 dias representam um grau maior de especialização celular do que aqueles vistos em 10 dias, como “*synaptogenesis*” (sinaptogênese, processo de formação de sinapses entre neurônios), “*synapse organization and biogenesis*”, “*brain development*” e “*transmission of nerve impulse*”. Esses resultados podem ser indicativos do maior grau de especialização que as células adquirem durante o processo de diferenciação celular; enquanto que em 10 dias os genes com expressão aumentada em relação às células indiferenciadas dão início ou fazem parte de processos mais amplos relacionados a determinação de linhagens celulares específicas (como “*nervous system development*” e “*multicellular organismal development*”, onde células pluri ou multipotentes se comprometem geneticamente a

tornarem-se células do sistema nervoso ou outros tecidos, indicando um maior grau de especialização, mas que ainda podem originar células ainda mais especializadas), em 14 dias os genes com expressão aumentada refletem processos mais específicos, como a formação de sinapses entre neurônios já formados, formação do cérebro e transmissão de impulsos nervosos.

Além de indicar quais processos biológicos tiveram maior ocorrência na população de células em diferenciação, os mRNAs representantes desses processos serviram como referência para a busca de lncRNAs importantes para a diferenciação neural. Para isso, foram selecionados os mRNAs tanto da fração citoplasmática quanto da nuclear que tenham sido diferencialmente expressos em 10 dias e/ou 14 dias. Com esse conjunto, fez-se uma nova análise de enriquecimento de GO com as classificações processo biológico e função molecular. A análise dos processos biológicos mostrou um resultado esperado e muito similar aos já anteriormente obtidos, sem nenhum termo GO novo em relação a eles. A metodologia empregada para a obtenção de lncRNAs de interesse a partir dessa análise de enriquecimento de GO é discutida na seção 5.

Na classificação GO de função molecular, os termos enriquecidos estão apresentados na tabela 4. Aqueles relacionados ao sistema nervoso que mais chamaram a atenção foram “*glutamate receptor activity*” e “*ionotropic glutamate receptor activity*”. Glutamato é um neurotransmissor das sinapses que ocorrem no sistema nervoso, sendo importante nas comunicações neurais, formação da memória, aprendizado e regulação (Meldrum 2000). Outros termos que indicam processos moleculares que ocorrem com maior frequência em células em diferenciação e proliferação (células sob fino controle transcricional) também apresentaram enriquecimento, como “*transcription factor activity*”, “*transcription cofactor activity*”, “*transcription corepressor activity*” e “*sequence specific DNA binding*” (Copland *et al.* 2009).

Tabela 4 - Enriquecimento de termos GO (função molecular) – 10 e/ou 14 dias.

Termo GO	NES	FDR
TRANSCRIPTION COFACTOR ACTIVITY	2,28	0,002
TRANSCRIPTION COACTIVATOR ACTIVITY	2,22	0,003
GLYCOSAMINOGLYCAN BINDING	2,22	0,008
SEQUENCE SPECIFIC DNA BINDING	2,17	0,011
HEPARIN BINDING	2,17	0,011
POLYSACCHARIDE BINDING	2,17	0,012
TRANSCRIPTION COREPRESSOR ACTIVITY	2,16	0,015
GLUTAMATE RECEPTOR ACTIVITY	2,15	0,021
PATTERN BINDING	2,06	0,041
IONOTROPIC GLUTAMATE RECEPTOR ACTIVITY	2,05	0,041

3.2 GO-Function

O pacote GO-Function é uma ferramenta de análise de enriquecimento de GO desenvolvido para tratar da redundância resultante da estrutura da ontologia gênica ou de genes com anotações múltiplas. A ferramenta foi desenvolvida tendo como objetivo encontrar termos significativos tanto estatística quanto biologicamente (Wang *et al.* 2012).

Ao se utilizar o GO-Function, buscou-se caracterizar quais processos biológicos estão representados pelos mRNAs correlacionados a um determinado lncRNA. Para isso, para cada lncRNA diferencialmente expresso na fração nuclear, foram calculados coeficientes de correlação de Pearson com todos os mRNAs detectados em no mínimo um dos três períodos de diferenciação estudados (células indiferenciadas, 10 dias após diferenciação e 14 dias após diferenciação). A cada lncRNA foi atribuído o conjunto dos mRNAs com coeficiente de correlação maior ou igual a 0,8 ou menor ou igual a -0,8. A análise de enriquecimento realizada pelo GO-Function fez uso de cada um desses conjuntos, atribuindo a cada lncRNA os termos GO enriquecidos através dos conjuntos de mRNAs correlacionados. Os resultados, por serem individuais a cada lncRNA, foram utilizados em conjunto com as análises do GSEA na busca por lncRNAs de interesse (seção 6).

4. Correlação de Pearson e teste de *bootstrap*

A função de genes codificadores durante a diferenciação neural é bem compreendida (Ying & Smith 2003, Kawaguchi *et al.* 2008), e através da comparação desses padrões de expressão com aqueles exibidos pelos lncRNAs, pode-se identificar padrões de coexpressão durante a diferenciação celular. A evidência de coexpressão pode implicar que esses lncRNAs desempenhem suas funções nas mesmas vias desses mRNAs, já que podem estar sendo regulados por um mesmo elemento (Zhao *et al.* 2014).

Para a identificação daqueles lncRNAs coexpressos com mRNAs durante a diferenciação neural, foi calculado o coeficiente de correlação de Pearson dos lncRNAs diferencialmente expressos com mRNAs pertencentes às categorias GO enriquecidas.

Para se verificar estatisticamente a significância desses coeficientes de correlação, implementou-se um teste estatístico utilizando permutações no cálculo desses coeficientes, denominado de *bootstrap*. Para cada lncRNA, após cada correlação calculada com um mRNA, foram calculadas 1.000 outras correlações utilizando-se valores permutados dentre todas as razões observadas dos mRNAs. O valor de *p-value* atribuído por esse teste a uma determinada correlação lncRNA-mRNA correspondeu à porcentagem das correlações permutadas onde o módulo do coeficiente calculado após a permutação foi menor do que o módulo do coeficiente calculado através dos valores observados para aquele lncRNA e mRNA (figura 7).

Figura 7 – Passos para o cálculo da correlação de Pearson (r) e obtenção de seu p -value para um dado par lncRNA-mRNA. O cálculo do p -value envolve 1.000 repetições do cálculo de correlação utilizando as medidas de expressão gênica nas 3 condições estudadas (célula indiferenciada, diferenciada após 10 dias e diferenciada após 14 dias) obtidas de um determinado lncRNA com valores permutados dentre o total de medidas observadas, para as mesmas condições, dos mRNAs (r_p). O p -value corresponde à porcentagem de correlações utilizando valores permutados onde $|r_p| < |r|$.

5. Identificação de possíveis lncRNAs candidatos a validação funcional

Para a identificação de possíveis lncRNAs que possam vir a ter uma função regulatória no processo de diferenciação neural de células P19 e, portanto, candidatos a validação funcional, foi empregado o conceito de *guilt by association* (Pauli et al. 2012), onde as possíveis funções desses lncRNAs são hipotetizadas a partir dos padrões de coexpressão com mRNAs conhecidos.

Esses lncRNAs foram comparados com os mRNAs pertencentes ao *core enrichment* das categorias GO enriquecidas relacionadas a diferenciação neural.

Pertencer ao *core enrichment* significa ter uma maior contribuição para o enriquecimento daquele termo devido à sua variação entre os fenótipos (indiferenciada e diferenciada). Logo, esses mRNAs são responsáveis pelos processos celulares cruciais que levam às diferenças observadas entre as células indiferenciadas e diferenciadas. O fato de um lncRNA ser coexpresso com esses mRNAs reforça a ideia de que esse lncRNA também pode ter uma função e ser regulado por elementos inerentes aos processos representados por esses mRNAs.

Os mRNAs foram selecionados a partir das seguintes categorias GO enriquecidas (figura 8):

- Desenvolvimento cerebral (*Brain development*);
- Desenvolvimento do sistema nervoso central (*Central nervous system development*);
- Desenvolvimento do sistema nervoso (*Nervous system development*);
- Processo do sistema neurológico (*Neurological system process*);
- Organização e biogênese sináptica (*Synapse organization and biogenesis*);
- Transmissão sináptica (*Synaptic transmission*);
- Sinaptogênese (*Synaptogenesis*);
- Desenvolvimento de sistema (*System development*);
- Transmissão do impulse nervoso (*Transmission of nerve impulse*).

Foram selecionados os lncRNAs que apresentaram um coeficiente de correlação maior que 0,8 (para relações diretas) ou menor do que -0,8 (para relações inversas) aos mRNAs pertencentes ao *core enrichment*. A cada um deles foi atribuído um conjunto de mRNAs correlacionados e categorias GO representadas por esse conjunto. Os termos GO associados a esses lncRNAs foram comparados com aqueles obtidos através da ferramenta GO-Function. Apenas foram considerados os lncRNAs que, entre as duas análises, compartilharam ao menos 1 termo GO enriquecido relacionado a processos relacionados à diferenciação neural.

Caso algum dos mRNAs associados dessa maneira fosse o gene codificador mais próximo (com ou sem sobreposição) do lncRNA em questão, esse lncRNA foi

considerado como potencial regulador em cis. Os demais foram considerados como possíveis reguladores em trans. Foram obtidos dessa forma 9 lncRNAs em cis e 229 em trans. O processo decisório está evidenciado na figura 9.

Figura 8 – Representação gráfica das categorias GO enriquecidas (processo biológico) no conjunto de mRNAs diferencialmente expressos nas frações citoplasmática e/ou nuclear, após 10 e/ou 14 dias. As categorias aqui mostradas apresentaram $FDR < 0,05$, e têm relação direta com o processo de diferenciação neural. Os picos (pontos mais altos da linha em verde) representam o valor de ES (*enrichment score*) do termo em questão. As barras pretas abaixo do gráfico representam em que posição da lista de variação de expressão utilizada na análise encontram-se os genes presentes no conjunto definido *a priori* dessa categoria (vermelho = variação positiva, *upregulated*; azul = variação negativa, *downregulated*).

Figura 9 – Fluxograma representando o processo decisório para a seleção de lncRNAs que mostraram potencial de possuírem algum papel regulatório ou serem regulados durante o processo de diferenciação neural. Foram considerados os lncRNAs anotados, diferencialmente expressos e que apresentaram um coeficiente de correlação maior que 0,8 ou menor que -0,8, com $p\text{-value} < 0,05$. Todos os mRNAs utilizados foram diferencialmente expressos em 10 e/ou 14 dias e participaram do *core enrichment* das categorias GO em questão. No caso do gene mais próximo de um lncRNA ser um mRNA desse conjunto, esse lncRNA foi classificado como “cis”; caso contrário como “trans”.

Apesar da maioria dos lncRNAs obtidos através dessa análise corresponder a possíveis reguladores em trans, a evidência de expressão diferencial de lncRNAs vizinhos a mRNAs importantes para a diferenciação neural é mais fácil de ser analisada, pois se tem um forte indício de que há alguma relação entre esse par lncRNA-mRNA graças à informação posicional. A tabela 5 mostra os 9 lncRNAs com possível regulação em cis, seus mRNAs vizinhos e variação em 10 e 14 dias. A figura 10 mostra a localização desses lncRNAs referentes a seus mRNAs vizinhos.

Tabela 5 – Possíveis lncRNAs reguladores da diferenciação neural em cis. A quantidade de mRNAs correlacionados representa a quantidade de mRNAs cujo $|r| > 0,8$.

lncRNA	mRNAs correlacionados	mRNA mais próximo	Correlação (<i>r</i>)	Variação 10d / Indiferenciada	Variação 14d / Indiferenciada
Gm2694	78	Cbln1	0,95	5,35	11,55
Igf2as	61	Igf2	0,98	3,0	7,6
AV512926	33	Lsamp	0,84	9,23	36,5
NONMMUG016481	34	Lsamp	0,84	6,0	32,37
NONMMUG016484	33	Lsamp	0,83	5,5	46,8
NONMMUG016486	33	Lsamp	0,84	5,0	15,0
NONMMUG016497	34	Lsamp	0,87	3,81	8,5
Pantr1	62	Pou3f3	0,94	6,15	28,6
NONMMUG000506	69	Pou3f3	0,91	3,17	9,05

5.1 *Cbln1*

O gene *cbln1* codifica uma proteína precursora específica do cerebelo, preceberelina. Essa proteína é processada para dar origem a diversos derivados, incluindo a ceberelina, que é altamente enriquecida em estruturas pós-sinápticas das células de Purkinje (uma classe de neurônios GABAérgicos localizados no cerebelo). Sua presença é essencial para a integridade e plasticidade sináptica, especialmente nesse tipo de neurônio (Urade *et al.* 1991). O lncRNA intergênico Gm2694 localiza-se a uma distância de 221 pb do início do gene *cbln1*, com orientação oposta a ele e já foi validado em estudos anteriores (Diez-Roux *et al.* 2011). Não existem estudos mais aprofundados a respeito desse lncRNA, mas há evidências de alterações em seus níveis de expressão quando outros lncRNAs que atuam na diferenciação neural são nocauteados (Guttman *et al.* 2011), tornando esse lncRNA um candidato a validação funcional.

5.2 *Igf2*

O gene *igf2* codifica um membro da família da insulina, composta de fatores de crescimento envolvidos do desenvolvimento e crescimento. A proteína IGF-II atua como um potente agente mitogênico para células cultivadas *in vitro*, e *in vivo* pode ter um papel no desenvolvimento fetal (Hill *et al.* 1998). Apesar de possuir um papel na

proteção de neurônios através do controle homeostático do cálcio (Mattson & Cheng 1993), o gene *igf2* é regulado pelo ácido retinóico (Matsumoto *et al.* 1992), podendo seus níveis de expressão observados nesse trabalho terem sido alterados devido ao tratamento com a substância no protocolo de diferenciação das células P19. O lncRNA mais próximo a esse gene é o antisense *Igf2as*, e não há relação entre ele e a diferenciação neural na literatura existente, sendo esse transcrito associado ao câncer de pulmão (Hosgood *et al.* 2008) e ao tumor de Wilm, um tipo raro de câncer renal (Okutsu *et al.* 2000). Esse lncRNA apresenta indícios de que esteja relacionado a processos de divisão celular, não sendo sua atuação restrita ao desenvolvimento neural.

5.3 *Lsamp*

A proteína codificada pelo gene *Lsamp* é uma glicoproteína de superfície neuronal encontrada nas regiões corticais e subcorticais do sistema límbico, região do cérebro responsável pelas emoções e comportamentos sociais. Durante o desenvolvimento do sistema límbico essa proteína pode ser encontrada na superfície de membranas axônicas e cones em crescimento, onde ela atua como uma molécula de adesão seletiva, guiando o desenvolvimento de padrões específicos de conexões neurais (Pimenta & Levitt 2004). Polimorfismos desse gene estão associados a esquizofrenia, distúrbios depressivos maiores e distúrbios do pânico (Koido *et al.* 2012, Koido *et al.* 2014). Cinco lncRNAs se localizam nas imediações desse gene, cada um deles com um tamanho médio de 2.000 pb. Todos eles sobrepõem um mesmo transcrito sem classificação do RefSeq, AK082117, podendo representar éxons mal classificados desse transcrito. A importância desse gene no sistema nervoso e a variação observada desses lncRNAs na diferenciação (tabela 5) sugerem que eles possam ter alguma relação com esse processo, sejam eles transcritos individuais ou o transcrito AK082117.

5.4 *Pou3f3*

O gene *pou3f3* codifica uma proteína que contém um domínio POU e que atua como um fator de transcrição. O acrônimo POU é derivado dos nomes de três fatores de transcrição: **Pit-1** (*Pituitary-specific*), **Oct1/Oct2** (*Octamer transcription fator*) e **Unc-86**. Esses fatores de transcrição são altamente conservados entre organismos como *Caenorhabditis elegans*, *Drosophila*, *Xenopus*, peixe zebra e humano (Ruvkun & Finney 1991). O fator de transcrição específico POU3F3 atua em conjunto com os fatores SOX11 e SOX4 durante o desenvolvimento neuronal (Dominguez *et al.* 2013).

O lncRNA intergênico Pantr1 é um transcrito adjacente ao gene *pou3f3*, no sentido oposto. Sabe-se que a deleção desse lncRNA leva a perturbações nos níveis de expressão de genes ligados ao desenvolvimento neural em camundongos em desenvolvimento e pós-natais (Sauvageau *et al.* 2013, Goff *et al.* 2015). Também nessa região genômica está presente o lncRNA NONMMUG000506, que se sobrepõe a éxons do gene *pou3f3* no sentido senso. Nenhum trabalho foi encontrado detalhando sua origem na literatura.

5.4 Possíveis lncRNAs regulatórios em trans

A partir dos critérios de variação de expressão após 10 e 14 dias de diferenciação e coeficientes de correlação com mRNAs relacionados a processos neurais, foram selecionados manualmente 9 dentre os 229 lncRNAs com possível regulação em trans que mostraram mais predisposição a terem alguma função regulatória na diferenciação neural. Destes, 5 possuem variação positiva, significando que são mais expressos nas células diferenciadas do que nas indiferenciadas. Todos eles possuem correlação $|r| > 0,8$ em comum com mRNAs neurais, sendo os mais destacados:

- *cacna1h* (subunidade de um canal de cálcio voltagem dependente expresso no cérebro) (Chemin *et al.* 2002);
- *cdk5* (uma quinase dependente de ciclina essencial para o ciclo celular neuronal) (Zhu *et al.* 2011);
- *gria2* (subunidade do receptor de glutamato relacionados a canais de íons no sistema nervoso central) (Dziedzic *et al.* 2003);
- *zic1* (um membro da família de proteína *zinc finger*, importantes durante o desenvolvimento, principalmente do cérebro) (Hong & Saint-Jeannet 2007).

Os 4 restantes possuem variação negativa, sendo mais expressos nas células indiferenciadas. Estes possuem apenas correlações negativas com mRNAs mais expressos nas células diferenciadas. A ausência de correlações positivas com mRNAs mais expressos nas células indiferenciadas se deve ao fato de que as categorias GO enriquecidas nessas células resultaram em altos índices de FDR. Portanto, foram

investigados apenas lncRNAs que tivessem correlação com os mRNAs enriquecidos nas células diferenciadas, devido aos maiores graus de confiabilidade ($FDR < 5,0\%$) do enriquecimento dessas categorias.

Os lncRNAs com possível regulação em trans e seus valores de variação são evidenciados na tabela 6.

Tabela 6 - Possíveis lncRNAs reguladores da diferenciação neural em trans e seus valores de variação durante a diferenciação neural.

lncRNA	Variação 10d / Indiferenciada	Variação 14d / Indiferenciada
C330013F16Rik	-14.18	-18.04
Gm16880	-7.66	-19.09
NONMMUG033336	-5.23	-5.84
AV099323	-2.43	-5.74
NEAT2	3.03	5.61
NONMMUG013142	4.92	5.30
NONMMUG044909	5.18	7.52
RP24-89G21.1	5.08	6.07
RP24-144C10.1	6.21	1.68

Figura 10 – posição genômica dos lncRNAs com possível atuação em cis, evidenciando os genes codificadores mais próximos. Em azul escuro encontram-se os genes codificadores presentes na base de dados do UCSC ou RefSeq, e em azul claro os lncRNAs também presentes nessas bases. Em vermelho estão indicados os lncRNAs determinados por este trabalho. A) lncRNA Gm2694/gene codificador *cbln1*. B) lncRNA IGF2as/gene codificador *igf2*. C) lncRNAs AV512926, NONMMUG016481, NONMMUG016484, NONMMUG016486 e NONMMUG016497/gene codificador *lsamp*. D) lncRNAs Pantr1 e NONMMUG000506/gene codificador *pou3f3*.

CONCLUSÃO

O primeiro passo deste trabalho, a reanotação das sondas de lncRNAs pobremente anotadas, se mostrou muito útil na identificação de lncRNAs diferencialmente expressos durante a diferenciação em neurônios e glia. As informações a respeito de cada lncRNA contidas nos bancos de dados de onde se originaram essas anotações permitem uma análise muito mais rápida, além de referenciar os trabalhos prévios onde tais lncRNAs foram estudados. Por se tratar de uma plataforma comercial, a publicação desta reanotação pode ser de grande valia para pesquisadores que a utilizam. Com vista às constantes adições de novas moléculas ao rol de lncRNAs anotados, a inclusão de atualizações constantes a esse *pipeline* é uma ideia para o seu melhoramento. Além disso, os procedimentos adotados podem ser aplicados a outras plataformas que tenham como foco o estudo de lncRNAs, e não apenas de camundongo, mas qualquer espécie.

A quantidade de lncRNAs encontrados através das análises empregadas nesse trabalho é grande, e trabalhos futuros poderão explorar outros aspectos da biologia dessas moléculas para a determinação de lncRNAs que possam vir a ser reguladores de processos biológicos durante a diferenciação neural. Tais análises envolvem estimativa de estrutura secundária, procura por motivos de ligação de proteínas ou outras moléculas regulatórias em suas sequências, estimativa de conservação evolutiva, entre outros. Conforme novos aspectos da biologia dos lncRNAs venham a ser descobertas, poderão ser empregados na busca de novos lncRNAs funcionais. Além disso, a metodologia aqui empregada não deixa de ser um importante primeiro passo em estudos que busquem investigar grandes volumes de dados biológicos na procura dessas moléculas, tendo como guia os conceitos de coexpressão e coregulação, que ajudam a relacionar informações obtidas sobre lncRNAs desconhecidos e mRNAs já anotados. Entretanto, a análise bioinformática não deve ser a metodologia única e definitiva empregada em tais estudos. Ela deve ser utilizada como forma de identificar, dentre várias possibilidades de supostos lncRNAs funcionais, aqueles que apresentam evidências que justifiquem sua caracterização, direcionando os esforços a candidatos cujas chances de realmente possuírem alguma função sejam maiores e diminuindo o tempo e recursos gastos nessas empreitadas.

Em suma, este trabalho contribuiu para o conhecimento acerca da variação que ocorre nos níveis de expressão de lncRNAs e mRNAs durante o processo de

diferenciação *in vitro* de células P19 após 10 dias, onde existe uma maior quantidade de neurônios na cultura, e após 14 dias, onde há uma mistura de neurônios e células da glia. A maioria dos mRNAs avaliados nesse trabalho, que serviram de referência na busca por lncRNAs, foram mais estudados ou possuem funções mais reconhecidas em neurônios, o que inadvertidamente levou ao estabelecimento de relações mais próximas entre esse tipo celular e as possíveis funções exercidas pelos lncRNAs em detrimento de relações com células da glia. Outro fato que dificulta a relação direta de lncRNAs com apenas um dos tipos celulares estudados, neurônios ou células da glia, é a heterogeneidade das culturas celulares a partir das quais os dados de expressão foram gerados, principalmente em 14 dias, onde existem frações elevadas de ambos os tipos celulares. As análises implementadas através de rotinas informáticas ajudaram na identificação de lncRNAs, alguns destes ainda pouco caracterizados, que aparentam possuir funções regulatórias durante a diferenciação neural. Esses lncRNAs selecionados poderão ser objetos de estudos futuros que visem caracterizá-los funcionalmente, onde será possível estabelecer com mais certeza os tipos celulares específicos onde esses lncRNAs atuam e seu grau de importância no processo de diferenciação como um todo.

ANEXOS

1. Resultados preliminares

Os dados de expressão gênica da linhagem P19 indiferenciada e após a diferenciação em neurônios ou glia foram gerados recentemente pela Dra. Ana Carolina Ayupe, que realiza pós-doutorado em nosso grupo, em colaboração com o grupo do Prof. Dr. Alexander Henning Ulrich do IQ-USP. O projeto busca investigar mecanismos de retenção de RNA no núcleo e irá utilizar células P19 como modelo para avaliar mudanças na distribuição subcelular de RNAs durante a diferenciação celular.

Foram utilizadas células P19 tratadas com ácido retinóico e mantidas em cultura por 10 e 14 dias para obtenção de perfis de expressão gênica de frações nucleares e citoplasmáticas de P19 durante a diferenciação neural. Segundo já observado pelo grupo do professor Dr. Ulrich, que tem vasta experiência no cultivo e diferenciação de células P19, no 10º dia de diferenciação ocorre predominância de neurônios na cultura, enquanto no 14º tem-se predominantemente células da glia na cultura, verificado através da detecção por imunofluorescência de marcadores específicos de neurônio ou glia.

Foram feitos o cultivo e o fracionamento em núcleo e citoplasma de células P19 indiferenciadas e diferenciadas de 10 e 14 dias. Após extração do RNA das frações, foram obtidos RNAs enriquecidos em núcleo e citoplasma. A observação da presença do precursor de RNA ribossômico (*ribossomal RNA*, rRNA) 45S apenas na fração nuclear, que não deixa o núcleo até ser processado nas subunidades 28S e 18S, demonstra a boa qualidade do fracionamento das amostras (figura A1).

Figura A1 - Eletroforese microfluídica dos RNAs enriquecidos nas frações nucleares e citoplasmáticas de células P19 indiferenciadas e diferenciadas de 10 (predominância de neurônio) e 14 dias (predominância de células da glia). É relevante notar a presença do precursor de rRNA 45S apenas na fração nuclear das preparações, o que é indicativo de boa qualidade do fracionamento. (N) fração nuclear, (C) fração citoplasmática, (IND) indiferenciada, (DIF) diferenciada.

Além deste controle, foram extraídas proteínas das frações nucleares e citoplasmáticas das células em estudo e realizado um ensaio de *western blot*, utilizando anticorpos para GAPDH (enriquecida no citoplasma) e histona H3 (enriquecida no núcleo) como controles da qualidade do fracionamento. Através destes controles, foi confirmado o enriquecimento das frações obtidas em núcleo e citoplasma (figura A2).

Figura A2 – *Western blot* para avaliação da qualidade do fracionamento núcleo/citoplasma de células P19 indiferenciadas e diferenciadas por 10 dias (predominância de neurônio) ou 14 dias (predominância de células da glia). GAPDH foi utilizado como controle citoplasmático e histona H3 foi utilizada como controle nuclear. (N) fração nuclear, (C) fração citoplasmática. Os números 1, 2, 3 e 4 referem-se a 4 fracionamentos independentes.

Após a verificação da qualidade do fracionamento, os RNAs enriquecidos em núcleo e citoplasma de células P19 indiferenciadas, diferenciadas em neurônio e em células da glia foram marcados com fluoróforos distintos (Cy3 ou Cy5) e hibridizados em microarranjos de 8x60k fabricados pela Agilent (Agilent SurePrint G3 Mouse GE 8x60K Microarray, catálogo #G4852A), que contém sondas com 60-mer que interrogam mRNAs de todos os genes anotados no genoma do camundongo (39.430 sondas), além de 16.251 sondas para lncRNAs (figura A3).

Após a hibridização, as lâminas foram lidas em um scanner a laser e as imagens foram analisadas no programa Feature Extraction (Agilent) para obtenção dos valores numéricos correspondentes a expressão dos mRNAs e lncRNAs nos três diferentes tipos celulares (indiferenciada, neurônio e glia), gerando uma planilha de dados em formato texto tabulado que foi utilizada no desenvolvimento deste projeto.

Figura A3 - Esquema das hibridizações de células P19 indiferenciadas e diferenciadas predominantemente em neurônios e células da glia em microarranjos de 8x60k. Cada lâmina contém 8 *arrays* independentes. No painel A, temos a lâmina que foi hibridizada com RNAs enriquecidos na fração nuclear (4 *arrays* superiores) e na fração citoplasmática (4 *arrays* inferiores) de células P19 diferenciadas em neurônio e de P19 indiferenciada. No painel B, temos a lâmina que foi hibridizada com RNAs enriquecidos na fração nuclear (4 *arrays* superiores) e na fração citoplasmática (4 *arrays* inferiores) de células P19 diferenciadas em células da glia e de P19 indiferenciada. (N) fração nuclear, (C) fração citoplasmática, (DN) P19 diferenciada em neurônio, (IN) P19 indiferenciada, (DG) P19 diferenciada em células da glia. Os números 1,2, 3 e 4 referem-se as 4 réplicas biológicas provenientes de fracionamentos celulares independentes.

2. Imunocitoquímica

No décimo e no décimo quarto dia de diferenciação, as células foram fixadas com 4,0 % paraformaldeído (PFA) por 20 minutos. Em seguida, as mesmas foram incubadas em PBS com 0,2% de Triton X-100 e 3,0% SFB por 20 min. As amostras foram, então, incubadas com os anticorpos primários anti-GFAP (1:500; DAKO Systems) e anti- β 3-Tubulina (1:1000; Sigma-Aldrich) em 0,2% de Triton X-100, 3,0% SFB, PBS por 2 horas em temperatura ambiente. Após, foram realizadas duas lavagens com PBS. As amostras foram incubadas por 1h com anticorpos secundários conjugados a fluoróforos (Alexa Fluor 488 e 555; ambos na diluição de 1:1000; Sigma-Aldrich). Após duas lavagens com PBS, foi realizada uma incubação durante 5 min 4',6-diamidino-2-phenylindole (DAPI; 1:10.000 em PBS) para marcar os núcleos. As lamínas foram montadas utilizando-se o meio de montagem DPX (Sigma-Aldrich). Os controles negativos foram feitos sem a adição dos anticorpos primários. As imagens foram obtidas com câmera digital Nikon DXM1200F e microscópio Axiovert 200 (Zeiss). As análises das imagens foram realizadas utilizando-se o programa ImageJ (National Institutes of Health). Os experimentos foram realizados em triplicatas experimentais.

3. Citometria de fluxo

No décimo e no décimo quarto dia de diferenciação, as amostras foram dissociadas pela incubação com tripsina (Invitrogen) por 3 min a 37°C. Em seguida, adicionou-se SFB e as células foram dissociadas mecanicamente e filtradas em poros de 40 μ m para obtenção de células individualizadas. Posteriormente, as células foram fixadas em 4,0% de PFA por 20 minutos. Após duas lavagens com PBS, as células foram incubadas em PBS contendo 3,0% de SFB, 0,2% de Triton X-100 e os anticorpos primários anti-GFAP e anti- β 3-Tubulina, ambos na diluição de 1:500. Após 30 minutos de incubação a temperatura ambiente, as células foram lavadas com PBS e incubadas novamente por 30 minutos com os anticorpos secundários conjugados a fluoróforos (Alexa Fluor 488 e 555; ambos na diluição de 1:1.000). Por fim, as células foram lavadas e ressuspensas em 500 mL de PBS para aquisição no citômetro de fluxo Attune (Life Technology). No mínimo 30.000 eventos foram adquiridos por amostra.

Gates no *dot-plot* do canal de dispersão frontal (FS – *forward scatter*) versus o canal de dispersão lateral (SS – *side scatter*) foram realizados para excluir *debris* celulares. A amostra negativa consistiu de células não expostas ao anticorpo primário. A

partir dessa amostra, traçou-se a região que delimita as células negativas, onde os eventos observados a direita foram considerados positivos. As análises foram realizadas no *software* do equipamento citômetro de fluxo Attune.

P19 dia 10

β -III tubulina
(neurônio)

GFAP
(glia)

DAPI
(núcleo)

MERGE

Negativo

β III tubulina

Negativo

GFAP

P19 dia 14

β -III tubulina
(neurônio)

GFAP
(glia)

DAPI
(núcleo)

MERGE

Negativo

β III tubulina

Negativo

GFAP

Figura A4 (pág. anterior) – (Microfotografias) Imunocitoquímica de células P19 diferenciadas durante 10 dias (P19 dia 10) e 14 dias (P19 dia 14) utilizando anticorpos para detecção da expressão de β 3-tubulina (marcador de neurônio) e GFAP (marcador de astrócitos, célula da glia). **(Dot-plots)** Análise por citometria de fluxo da expressão de marcadores específicos para neurônios (β 3-tubulina) e para astrócitos (GFAP) em células P19 diferenciadas durante 10 dias (P19 dia 10) e 14 dias (P19 dia 14). O controle negativo foi realizado com a marcação somente pelo anticorpo secundário. Notar o maior enriquecimento do marcador de neurônio (β 3-tubulina) em relação ao marcador de células da glia (GFAP) em 10 dias e o maior enriquecimento de GFAP em relação à β 3-tubulina em 14 dias.

REFERÊNCIAS BIBLIOGRÁFICAS

- Amaral, P. P. and J. S. Mattick (2008). "Noncoding RNA in development." Mamm Genome **19**(7-8): 454-492. <<http://www.ncbi.nlm.nih.gov/pubmed/18839252>>
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, . . . G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29. <<http://www.ncbi.nlm.nih.gov/pubmed/10802651>>
- Bazin, J. and J. Bailey-Serres (2015). "Emerging roles of long non-coding RNA in root developmental plasticity and regulation of phosphate homeostasis." Front Plant Sci **6**: 400. <<http://www.ncbi.nlm.nih.gov/pubmed/26106399>>
- Beckedorff, F. C., A. C. Ayupe, R. Crocci-Souza, M. S. Amaral, H. I. Nakaya, D. T. Soltys, . . . S. Verjovski-Almeida (2013). "The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation." PLoS Genet **9**(8): e1003705. <<http://www.ncbi.nlm.nih.gov/pubmed/23990798>>
- Bernard, D., K. V. Prasanth, V. Tripathi, S. Colasse, T. Nakamura, Z. Xuan, . . . A. Bessis (2010). "A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression." EMBO J **29**(18): 3082-3093. <<http://www.ncbi.nlm.nih.gov/pubmed/20729808>>
- Bhan, A. and S. S. Mandal (2014). "Long noncoding RNAs: emerging stars in gene regulation, epigenetics and human disease." ChemMedChem **9**(9): 1932-1956. <<http://www.ncbi.nlm.nih.gov/pubmed/24677606>>
- Blume, S. W., Z. Meng, K. Shrestha, R. C. Snyder and P. D. Emanuel (2003). "The 5'-untranslated RNA of the human dhfr minor transcript alters transcription pre-initiation complex assembly at the major (core) promoter." J Cell Biochem **88**(1): 165-180. <<http://www.ncbi.nlm.nih.gov/pubmed/12461786>>
- Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, . . . G. Genome Science (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563. <<http://www.ncbi.nlm.nih.gov/pubmed/16141072>>
- Chemin, J., A. Monteil, E. Perez-Reyes, E. Bourinet, J. Nargeot and P. Lory (2002). "Specific contribution of human T-type calcium channel isoforms (alpha(1G), alpha(1H) and alpha(1I)) to neuronal excitability." J Physiol **540**(Pt 1): 3-14. <<http://www.ncbi.nlm.nih.gov/pubmed/11927664>>
- Copland, J. A., M. Sheffield-Moore, N. Koldzic-Zivanovic, S. Gentry, G. Lamprou, F. Tzortzatou-Stathopoulou, . . . S. A. Vlahopoulos (2009). "Sex steroid receptors in skeletal differentiation and epithelial neoplasia: is tissue-specific intervention possible?" Bioessays **31**(6): 629-641. <<http://www.ncbi.nlm.nih.gov/pubmed/19382224>>
- Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, . . . P. Flicek (2015). "Ensembl 2015." Nucleic Acids Res **43**(Database issue): D662-669. <<http://www.ncbi.nlm.nih.gov/pubmed/25352552>>

Diez-Roux, G., S. Banfi, M. Sultan, L. Geffers, S. Anand, D. Rozado, . . . A. Ballabio (2011). "A high-resolution anatomical atlas of the transcriptome in the mouse embryo." *PLoS Biol* **9**(1): e1000582. <<http://www.ncbi.nlm.nih.gov/pubmed/21267068>>

Dinger, M. E., P. P. Amaral, T. R. Mercer, K. C. Pang, S. J. Bruce, B. B. Gardiner, . . . J. S. Mattick (2008). "Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation." *Genome Res* **18**(9): 1433-1445. <<http://www.ncbi.nlm.nih.gov/pubmed/18562676>>

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, . . . T. R. Gingeras (2012). "Landscape of transcription in human cells." *Nature* **489**(7414): 101-108. <<http://www.ncbi.nlm.nih.gov/pubmed/22955620>>

Dominguez, M. H., A. E. Ayoub and P. Rakic (2013). "POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex." *Cereb Cortex* **23**(11): 2632-2643. <<http://www.ncbi.nlm.nih.gov/pubmed/22892427>>

Dziedzic, B., V. Prevot, A. Lomniczi, H. Jung, A. Cornea and S. R. Ojeda (2003). "Neuron-to-glia signaling mediated by excitatory amino acid receptors regulates ErbB receptor function in astroglial cells of the neuroendocrine brain." *J Neurosci* **23**(3): 915-926. <<http://www.ncbi.nlm.nih.gov/pubmed/12574420>>

Encode Project Consortium, N. (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science* **306**(5696): 636-640. <<http://www.ncbi.nlm.nih.gov/pubmed/15499007>>

Encode Project Consortium, N. (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57-74. <<http://www.ncbi.nlm.nih.gov/pubmed/22955616>>

Goff, L. A., A. F. Groff, M. Sauvageau, Z. Traves-Gibson, D. B. Sanchez-Gomez, M. Morse, . . . J. L. Rinn (2015). "Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain." *Proc Natl Acad Sci U S A* **112**(22): 6855-6862. <<http://www.ncbi.nlm.nih.gov/pubmed/26034286>>

Gong, C. and L. E. Maquat (2011). "lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements." *Nature* **470**(7333): 284-288. <<http://www.ncbi.nlm.nih.gov/pubmed/21307942>>

Grabel, L. (2012). "Developmental origin of neural stem cells: the glial cell that could." *Stem Cell Rev* **8**(2): 577-585. <<http://www.ncbi.nlm.nih.gov/pubmed/22350457>>

Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, . . . E. S. Lander (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." *Nature* **458**(7235): 223-227. <<http://www.ncbi.nlm.nih.gov/pubmed/19182780>>

Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, . . . E. S. Lander (2011). "lncRNAs act in the circuitry controlling pluripotency and differentiation." *Nature* **477**(7364): 295-300. <<http://www.ncbi.nlm.nih.gov/pubmed/21874018>>

Harrow, J., F. Denoeud, A. Frankish, A. Reymond, C. K. Chen, J. Chrast, . . . R. Guigo (2006). "GENCODE: producing a reference annotation for ENCODE." Genome Biol **7 Suppl 1**: S4 1-9. <<http://www.ncbi.nlm.nih.gov/pubmed/16925838>>

Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, . . . T. J. Hubbard (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Res **22**(9): 1760-1774. <<http://www.ncbi.nlm.nih.gov/pubmed/22955987>>

Hill, D. J., J. Petrik and E. Arany (1998). "Growth factors and the regulation of fetal growth." Diabetes Care **21 Suppl 2**: B60-69. <<http://www.ncbi.nlm.nih.gov/pubmed/9704229>>

Hollander, M. and D. A. Wolfe (1999). Nonparametric statistical methods. New York, Wiley.

Holoch, D. and D. Moazed (2015). "RNA-mediated epigenetic regulation of gene expression." Nat Rev Genet **16**(2): 71-84. <<http://www.ncbi.nlm.nih.gov/pubmed/25554358>>

Hong, C. S. and J. P. Saint-Jeannet (2007). "The activity of Pax3 and Zic1 regulates three distinct cell fates at the neural plate border." Mol Biol Cell **18**(6): 2192-2202. <<http://www.ncbi.nlm.nih.gov/pubmed/17409353>>

Hosgood, H. D., 3rd, I. Menashe, M. Shen, M. Yeager, J. Yuenger, P. Rajaraman, . . . Q. Lan (2008). "Pathway-based evaluation of 380 candidate genes and lung cancer susceptibility suggests the importance of the cell cycle pathway." Carcinogenesis **29**(10): 1938-1943. <<http://www.ncbi.nlm.nih.gov/pubmed/18676680>>

Huarte, M., M. Guttman, D. Feldser, M. Garber, M. J. Koziol, D. Kenzelmann-Broz, . . . J. L. Rinn (2010). "A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response." Cell **142**(3): 409-419. <<http://www.ncbi.nlm.nih.gov/pubmed/20673990>>

Iyer, M. K., Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, . . . A. M. Chinnaiyan (2015). "The landscape of long noncoding RNAs in the human transcriptome." Nat Genet **47**(3): 199-208. <<http://www.ncbi.nlm.nih.gov/pubmed/25599403>>

Jia, H., M. Osak, G. K. Bogu, L. W. Stanton, R. Johnson and L. Lipovich (2010). "Genome-wide computational identification and manual annotation of human long noncoding RNA genes." RNA **16**(8): 1478-1487. <<http://www.ncbi.nlm.nih.gov/pubmed/20587619>>

Johnsson, P., A. Ackley, L. Vidarsdottir, W. O. Lui, M. Corcoran, D. Grandér and K. V. Morris (2013). "A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells." Nat Struct Mol Biol **20**(4): 440-446. <<http://www.ncbi.nlm.nih.gov/pubmed/23435381>>

Jones-Villeneuve, E. M., M. W. McBurney, K. A. Rogers and V. I. Kalnins (1982). "Retinoic acid induces embryonal carcinoma cells to differentiate into neurons and glial cells." J Cell Biol **94**(2): 253-262. <<http://www.ncbi.nlm.nih.gov/pubmed/7107698>>

Kanduri, C. (2015). "Long noncoding RNAs: Lessons from genomic imprinting." Biochim Biophys Acta. <<http://www.ncbi.nlm.nih.gov/pubmed/26004516>>

Kapranov, P., J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, . . . T. R. Gingeras (2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription." Science **316**(5830): 1484-1488. <<http://www.ncbi.nlm.nih.gov/pubmed/17510325>>

Kawaguchi, A., T. Ikawa, T. Kasukawa, H. R. Ueda, K. Kurimoto, M. Saitou and F. Matsuzaki (2008). "Single-cell gene profiling defines differential progenitor subclasses in mammalian neurogenesis." Development **135**(18): 3113-3124. <<http://www.ncbi.nlm.nih.gov/pubmed/18725516>>

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664. <<http://www.ncbi.nlm.nih.gov/pubmed/11932250>>

Khalil, A. M., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, . . . J. L. Rinn (2009). "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression." Proc Natl Acad Sci U S A **106**(28): 11667-11672. <<http://www.ncbi.nlm.nih.gov/pubmed/19571010>>

Koido, K., S. Janno, T. Traks, M. Parksepp, U. Ljubajev, P. Veiksaar, . . . E. Vasar (2014). "Associations between polymorphisms of LSAMP gene and schizophrenia." Psychiatry Res **215**(3): 797-798. <<http://www.ncbi.nlm.nih.gov/pubmed/24491686>>

Koido, K., T. Traks, R. Balotsev, T. Eller, A. Must, S. Koks, . . . E. Vasar (2012). "Associations between LSAMP gene polymorphisms and major depressive disorder and panic disorder." Transl Psychiatry **2**: e152. <<http://www.ncbi.nlm.nih.gov/pubmed/22892717>>

Kokocinski, F., J. Harrow and T. Hubbard (2010). "AnnoTrack--a tracking system for genome annotation." BMC Genomics **11**: 538. <<http://www.ncbi.nlm.nih.gov/pubmed/20923551>>

Kozioł, M. J. and J. L. Rinn (2010). "RNA traffic control of chromatin complexes." Curr Opin Genet Dev **20**(2): 142-148. <<http://www.ncbi.nlm.nih.gov/pubmed/20362426>>

Kung, J. T., D. Colognori and J. T. Lee (2013). "Long noncoding RNAs: past, present, and future." Genetics **193**(3): 651-669. <<http://www.ncbi.nlm.nih.gov/pubmed/23463798>>

Ma, L., V. B. Bajic and Z. Zhang (2013). "On the classification of long non-coding RNAs." RNA Biol **10**(6): 925-933. <<http://www.ncbi.nlm.nih.gov/pubmed/23696037>>

Matsumoto, K., C. Gaetano, W. H. Daughaday and C. J. Thiele (1992). "Retinoic acid regulates insulin-like growth factor II expression in a neuroblastoma cell line." Endocrinology **130**(6): 3669-3676. <<http://www.ncbi.nlm.nih.gov/pubmed/1375906>>

Mattick, J. S. (2001). "Non-coding RNAs: the architects of eukaryotic complexity." EMBO Rep **2**(11): 986-991. <<http://www.ncbi.nlm.nih.gov/pubmed/11713189>>

Mattson, M. P. and B. Cheng (1993). "Growth factors protect neurons against excitotoxic/ischemic damage by stabilizing calcium homeostasis." Stroke **24**(12 Suppl): I136-140; discussion I144-135. <<http://www.ncbi.nlm.nih.gov/pubmed/8249011>>

McBurney, M. W. (1993). "P19 embryonal carcinoma cells." Int J Dev Biol **37**(1): 135-140. <<http://www.ncbi.nlm.nih.gov/pubmed/8507558>>

McBurney, M. W., E. M. Jones-Villeneuve, M. K. Edwards and P. J. Anderson (1982). "Control of muscle and neuronal differentiation in a cultured embryonal carcinoma cell line." *Nature* **299**(5879): 165-167. <<http://www.ncbi.nlm.nih.gov/pubmed/7110336>>

Meldrum, B. S. (2000). "Glutamate as a neurotransmitter in the brain: review of physiology and pathology." *J Nutr* **130**(4S Suppl): 1007S-1015S. <<http://www.ncbi.nlm.nih.gov/pubmed/10736372>>

Mercer, T. R., M. E. Dinger, S. M. Sunkin, M. F. Mehler and J. S. Mattick (2008). "Specific expression of long noncoding RNAs in the mouse brain." *Proc Natl Acad Sci U S A* **105**(2): 716-721. <<http://www.ncbi.nlm.nih.gov/pubmed/18184812>>

Mercer, T. R., I. A. Qureshi, S. Gokhan, M. E. Dinger, G. Li, J. S. Mattick and M. F. Mehler (2010). "Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation." *BMC Neurosci* **11**: 14. <<http://www.ncbi.nlm.nih.gov/pubmed/20137068>>

Mohammad, F., T. Mondal, N. Guseva, G. K. Pandey and C. Kanduri (2010). "Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1." *Development* **137**(15): 2493-2499. <<http://www.ncbi.nlm.nih.gov/pubmed/20573698>>

Monge, J. C., D. J. Stewart and P. Cernacek (1995). "Differentiation of embryonal carcinoma cells to a neural or cardiomyocyte lineage is associated with selective expression of endothelin receptors." *J Biol Chem* **270**(25): 15385-15390. <<http://www.ncbi.nlm.nih.gov/pubmed/7797528>>

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, . . . L. C. Groop (2003). "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." *Nat Genet* **34**(3): 267-273. <<http://www.ncbi.nlm.nih.gov/pubmed/12808457>>

Nazer, E. and E. P. Lei (2014). "Modulation of chromatin modifying complexes by noncoding RNAs in trans." *Curr Opin Genet Dev* **25**: 68-73. <<http://www.ncbi.nlm.nih.gov/pubmed/24534715>>

Ng, S. Y., R. Johnson and L. W. Stanton (2012). "Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors." *EMBO J* **31**(3): 522-533. <<http://www.ncbi.nlm.nih.gov/pubmed/22193719>>

Niemann, S. and H. C. Schaller (1996). "Head-activator and the neuroectodermal differentiation of P19 mouse embryonal carcinoma cells." *Neurosci Lett* **207**(1): 49-52. <<http://www.ncbi.nlm.nih.gov/pubmed/8710207>>

Okutsu, T., Y. Kuroiwa, F. Kagitani, M. Kai, K. Aisaka, O. Tsutsumi, . . . F. Ishino (2000). "Expression and imprinting status of human PEG8/IGF2AS, a paternally expressed antisense transcript from the IGF2 locus, in Wilms' tumors." *J Biochem* **127**(3): 475-483. <<http://www.ncbi.nlm.nih.gov/pubmed/10731720>>

Orom, U. A., T. Derrien, R. Guigo and R. Shiekhattar (2010). "Long noncoding RNAs as enhancers of gene expression." *Cold Spring Harb Symp Quant Biol* **75**: 325-331. <<http://www.ncbi.nlm.nih.gov/pubmed/21502407>>

Pandey, R. R., T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, . . . C. Kanduri (2008). "Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation." Mol Cell **32**(2): 232-246. <<http://www.ncbi.nlm.nih.gov/pubmed/18951091>>

Pauli, A., E. Valen, M. F. Lin, M. Garber, N. L. Vastenhouw, J. Z. Levin, . . . A. F. Schier (2012). "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis." Genome Res **22**(3): 577-591. <<http://www.ncbi.nlm.nih.gov/pubmed/22110045>>

Pefanis, E., J. Wang, G. Rothschild, J. Lim, D. Kazadi, J. Sun, . . . U. Basu (2015). "RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity." Cell **161**(4): 774-789. <<http://www.ncbi.nlm.nih.gov/pubmed/25957685>>

Pimenta, A. F. and P. Levitt (2004). "Characterization of the genomic structure of the mouse limbic system-associated membrane protein (Lsamp) gene." Genomics **83**(5): 790-801. <<http://www.ncbi.nlm.nih.gov/pubmed/15081109>>

Ponjavic, J., C. P. Ponting and G. Lunter (2007). "Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs." Genome Res **17**(5): 556-565. <<http://www.ncbi.nlm.nih.gov/pubmed/17387145>>

Prensner, J. R. and A. M. Chinnaiyan (2011). "The emergence of lncRNAs in cancer biology." Cancer Discov **1**(5): 391-407. <<http://www.ncbi.nlm.nih.gov/pubmed/22096659>>

Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, . . . J. M. Ostell (2014). "RefSeq: an update on mammalian reference sequences." Nucleic Acids Res **42**(Database issue): D756-763. <<http://www.ncbi.nlm.nih.gov/pubmed/24259432>>

Qu, H. and X. Fang (2013). "A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project." Genomics Proteomics Bioinformatics **11**(3): 135-141. <<http://www.ncbi.nlm.nih.gov/pubmed/23722115>>

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842. <<http://www.ncbi.nlm.nih.gov/pubmed/20110278>>

Quinn, J. J., I. A. Ilik, K. Qu, P. Georgiev, C. Chu, A. Akhtar and H. Y. Chang (2014). "Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification." Nat Biotechnol **32**(9): 933-940. <<http://www.ncbi.nlm.nih.gov/pubmed/24997788>>

Reis, E. M. and S. Verjovski-Almeida (2012). "Perspectives of Long Non-Coding RNAs in Cancer Diagnostics." Front Genet **3**: 32. <<http://www.ncbi.nlm.nih.gov/pubmed/22408643>>

Rosenbloom, K. R., J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, . . . W. J. Kent (2015). "The UCSC Genome Browser database: 2015 update." Nucleic Acids Res **43**(Database issue): D670-681. <<http://www.ncbi.nlm.nih.gov/pubmed/25428374>>

Ruan, X. and Y. Ruan (2012). "Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET)." Methods Mol Biol **809**: 535-562. <<http://www.ncbi.nlm.nih.gov/pubmed/22113299>>

Ruvkun, G. and M. Finney (1991). "Regulation of transcription and cell identity by POU domain proteins." *Cell* **64**(3): 475-478. <<http://www.ncbi.nlm.nih.gov/pubmed/1991317>>

Sauvageau, M., L. A. Goff, S. Lodato, B. Bonev, A. F. Groff, C. Gerhardinger, . . . J. L. Rinn (2013). "Multiple knockout mouse models reveal lincRNAs are required for life and brain development." *Elife* **2**: e01749. <<http://www.ncbi.nlm.nih.gov/pubmed/24381249>>

Sigova, A. A., A. C. Mullen, B. Molinie, S. Gupta, D. A. Orlando, M. G. Guenther, . . . R. A. Young (2013). "Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells." *Proc Natl Acad Sci U S A* **110**(8): 2876-2881. <<http://www.ncbi.nlm.nih.gov/pubmed/23382218>>

Spadaro, P. A., C. R. Flavell, J. Widagdo, V. S. Ratnu, M. Troup, C. Ragan, . . . T. W. Bredy (2015). "Long Noncoding RNA-Directed Epigenetic Regulation of Gene Expression Is Associated with Anxiety-like Behavior in Mice." *Biol Psychiatry*. <<http://www.ncbi.nlm.nih.gov/pubmed/25792222>>

Spizzo, R., M. I. Almeida, A. Colombatti and G. A. Calin (2012). "Long non-coding RNAs and cancer: a new frontier of translational research?" *Oncogene* **31**(43): 4577-4587. <<http://www.ncbi.nlm.nih.gov/pubmed/22266873>>

St Laurent, G., C. Wahlestedt and P. Kapranov (2015). "The Landscape of long noncoding RNA classification." *Trends Genet* **31**(5): 239-251. <<http://www.ncbi.nlm.nih.gov/pubmed/25869999>>

Staines, W. A., D. J. Morassutti, K. R. Reuhl, A. I. Ally and M. W. McBurney (1994). "Neurons derived from P19 embryonal carcinoma cells have varied morphologies and neurotransmitters." *Neuroscience* **58**(4): 735-751. <<http://www.ncbi.nlm.nih.gov/pubmed/7910670>>

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, . . . J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* **102**(43): 15545-15550. <<http://www.ncbi.nlm.nih.gov/pubmed/16199517>>

Sun, L., L. A. Goff, C. Trapnell, R. Alexander, K. A. Lo, E. Hacisuleyman, . . . J. L. Rinn (2013). "Long noncoding RNAs regulate adipogenesis." *Proc Natl Acad Sci U S A* **110**(9): 3387-3392. <<http://www.ncbi.nlm.nih.gov/pubmed/23401553>>

Tahira, A. C., M. S. Kubrusly, M. F. Faria, B. Dazzani, R. S. Fonseca, V. Maracaja-Coutinho, . . . E. M. Reis (2011). "Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer." *Mol Cancer* **10**: 141. <<http://www.ncbi.nlm.nih.gov/pubmed/22078386>>

Tochitani, S. and Y. Hayashizaki (2008). "Nkx2.2 antisense RNA overexpression enhanced oligodendrocytic differentiation." *Biochem Biophys Res Commun* **372**(4): 691-696. <<http://www.ncbi.nlm.nih.gov/pubmed/18538132>>

Tusher, V. G., R. Tibshirani and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response." *Proc Natl Acad Sci U S A* **98**(9): 5116-5121. <<http://www.ncbi.nlm.nih.gov/pubmed/11309499>>

Ulrich, H. and P. Majumder (2006). "Neurotransmitter receptor expression and activity during neuronal differentiation of embryonal carcinoma and stem cells: from basic research towards clinical applications." Cell Prolif **39**(4): 281-300. <<http://www.ncbi.nlm.nih.gov/pubmed/16872363>>

Urade, Y., J. Oberdick, R. Molinar-Rode and J. I. Morgan (1991). "Precerebellin is a cerebellum-specific protein with similarity to the globular domain of complement C1q B chain." Proc Natl Acad Sci U S A **88**(3): 1069-1073. <<http://www.ncbi.nlm.nih.gov/pubmed/1704129>>

Wang, J., X. Zhou, J. Zhu, Y. Gu, W. Zhao, J. Zou and Z. Guo (2012). "GO-function: deriving biologically relevant functions from statistically significant functions." Brief Bioinform **13**(2): 216-227. <<http://www.ncbi.nlm.nih.gov/pubmed/21705405>>

Wang, K. C. and H. Y. Chang (2011). "Molecular mechanisms of long noncoding RNAs." Mol Cell **43**(6): 904-914. <<http://www.ncbi.nlm.nih.gov/pubmed/21925379>>

Wang, K. C., Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, . . . H. Y. Chang (2011). "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression." Nature **472**(7341): 120-124. <<http://www.ncbi.nlm.nih.gov/pubmed/21423168>>

Xie, C., J. Yuan, H. Li, M. Li, G. Zhao, D. Bu, . . . Y. Zhao (2014). "NONCODEv4: exploring the world of long non-coding RNA genes." Nucleic Acids Res **42**(Database issue): D98-103. <<http://www.ncbi.nlm.nih.gov/pubmed/24285305>>

Ying, Q. L. and A. G. Smith (2003). "Defined conditions for neural commitment and differentiation." Methods Enzymol **365**: 327-341. <<http://www.ncbi.nlm.nih.gov/pubmed/14696356>>

Zang, S., R. Guo, L. Zhang and Y. Lu (2007). "Integration of statistical inference methods and a novel control measure to improve sensitivity and specificity of data analysis in expression profiling studies." J Biomed Inform **40**(5): 552-560. <<http://www.ncbi.nlm.nih.gov/pubmed/17317331>>

Zhao, Y., H. Luo, X. Chen, Y. Xiao and R. Chen (2014). "Computational methods to predict long noncoding RNA functions based on co-expression network." Methods Mol Biol **1182**: 209-218. <<http://www.ncbi.nlm.nih.gov/pubmed/25055914>>

Zhu, J., W. Li and Z. Mao (2011). "Cdk5: mediator of neuronal development, death and the response to DNA damage." Mech Ageing Dev **132**(8-9): 389-394. <<http://www.ncbi.nlm.nih.gov/pubmed/21600237>>