UNIVERSIDADE DE SÃO PAULO
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

MATHEUS CARVALHO BÜRGER

**Análise Transcricional de RNAs Não Codificadores Longos em Pacientes com
Dengue**

São Paulo

2017

MATHEUS CARVALHO BÜRGER

**Análise Transcricional de RNAs Não Codificadores Longos em Pacientes
com Dengue**

Versão corrigida

Dissertação apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Área de concentração: Bioinformática

Orientador: Prof. Dr. Helder Takashi Imoto Nakaya

São Paulo

2017

FICHA CATALOGRÁFICA

B954a Burger, Matheus Carvalho
Análise transcricional de RNAs não codificadores longos em pacientes com dengue /
Matheus Carvalho Burger, [orient.] Helder Takashi Imoto Nakaya. São Paulo : 2017.
86 p.

Dissertação (Mestrado) - Universidade de São Paulo
Orientador: Prof. Dr. Helder Takashi Imoto Nakaya

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Bioinformática. 2. Transcrição gênica. 3. Dengue. 4. Sequenciamento genético.
5. Genômica. I. Nakaya, Helder Takashi Imoto, orient. II. Universidade de São Paulo.

CDD - 572.8

Elaborada pelo Serviço de Informação e Biblioteca “Carlos Benjamin de Lyra do IME-USP”

Aos meus pais, Adriana e Claudinei.

À minha irmã, Karoline.

À minha namorada, Patrícia.

Agradecimentos

Agradeço aos meus pais por me ajudarem no meu caminho até aqui, sem o suporte deles com certeza esse trabalho não seria possível.

Agradeço á minha namorada, Patrícia, por todo amor e carinho, por ficar ao meu lado durante esse trabalho.

Agradeço à todas as pessoas do laboratório CSBL. O ambiente colaborativo formado permitiu que todos me ajudassem de alguma forma.

Agradeço especialmente ao Thiago, além de ser o assistente de assuntos aleatórios do laboratório, me ajudou demais no final desse trabalho. Sem ele esse trabalho não seria possível.

Agradeço ao Lucas pelas discussões sobre novas tecnologias e pelo trabalho de reanotação realizado.

Agradeço ao Diógenes por me auxiliar com alguns *scripts*, revisão de alguns textos e discussões sobre meu trabalho.

Agradeço ao Pedro e Gustavo pelo desenvolvimento do pacote CEMiTool.

Agradeço ao Prof. Helder por permitir o desenvolvimento desse trabalho e por tudo que aprendi no CSBL.

Agradeço à CAPES pelo financiamento desse trabalho

Resumo

Bürger, Matheus Carvalho. **Análise Transcricional de RNAs Não Codificadores Longos em Pacientes com Dengue.** 2017. 86 f. Dissertação (Mestrado em Ciências) – Universidade de São Paulo, São Paulo, 2017.

A dengue é uma infecção viral sistêmica que pode se manifestar clinicamente de diversas formas, desde febres leves a hemorragia e síndrome do choque, condições potencialmente fatais. Diversos estudos já foram publicados investigando as mudanças globais de expressão que ocorrem durante a evolução da doença nesses diferentes quadros clínicos. Porém, nenhum desses estudos analisou o papel dos RNAs não codificadores longos (lncRNAs) na progressão da doença. Neste projeto, foi realizada uma metanálise dos dados de expressão provenientes desses estudos de dengue focando na expressão de lncRNAs e seus possíveis mecanismos de regulação gênica. Foram identificados dezenas de lncRNAs cuja expressão aumenta ou diminui em pacientes infectados com dengue comparado com pessoas saudáveis. Através de análise de *“guilty-by-association”*, procurou-se identificar genes codificadores de proteína possivelmente regulados por esses lncRNAs ou genes que os regulem. Nossos resultados fornecem evidência de novos mecanismos de regulação entre lncRNAs e mRNAs.

Palavras-chaves: biologia de sistemas, RNAs não codificadores longos, bioinformática, transcrito, Dengue, infecção viral.

Abstract

Bürger, Matheus Carvalho. **Transcriptional Analysis of Long Non-coding RNAs in Dengue Patients..** 2017. 86 p. Dissertation (Master of Science) – University of São Paulo, São Paulo, 2017.

Dengue fever is a systemic viral infection that can manifest clinically in a variety of ways, from mild fever to potentially fatal conditions such as hemorrhage and shock syndrome. Several studies have already been published investigating the global changes in expression that occur during the evolution of the disease in these different clinical settings. However, none of these studies analyzed the role of long non-coding RNAs (lncRNAs) in disease progression. In this project, we performed a meta-analysis of transcriptome data obtained from these dengue studies and focused on the expression of lncRNAs and their possible mechanisms of gene regulation. Dozens of lncRNAs have been identified whose expression increases or decreases in patients infected with dengue compared to healthy individuals. Through guilty-by-association analysis, we identified several lncRNAs that possibly regulate protein coding genes. Our results provide evidence of novel regulatory mechanisms between lncRNAs and mRNAs.

Keywords: systems biology, long non-coding RNA, bioinformatics, transcriptome, dengue, viral infection.

Lista de figuras

Figura 1 – Evidência de ocorrência de dengue no mundo. Extraído de Bhatt et al. (2013)	19
Figura 2 – Representação de mecanismos de regulação gênica por lncRNAs.	28
Figura 3 – Número de artigos sobre técnicas de larga escala publicados e indexados no PubMed.	31
Figura 4 – Número amostras utilizadas em cada estudo (sem amostras <i>outliers</i>).	36
Figura 5 – Fluxograma representando as etapas do processo de reanotação das sondas de <i>microarray</i>	37
Figura 6 – Fluxograma de obtenção e pré-processamento de dados transcritômicos do banco de dados GEO.	39
Figura 7 – Classes de transcritos existentes nas plataformas GPL570 e GPL2700.	48
Figura 8 – <i>Volcano plot</i> das sondas diferencialmente expressos nos estudos de transcriptoma.	49
Figura 9 – Diagrama de <i>Venn</i> dos genes com pelo menos uma sonda considerada diferencialmente expressa.	50
Figura 10 – Número de genes consistentemente diferencialmente expressos.	50
Figura 11 – Comparação entre transcritômica por <i>microarray</i> e RNAseq.	53
Figura 12 – Análise de enriquecimento das comparações com <i>Blood Transcriptional Modules</i>	54
Figura 13 – Análise de enriquecimento das comparações com <i>Reactome Pathways</i>	55
Figura 14 – <i>Gene Set Enrichment Analysis</i> para comparação entre <i>EarlyDengue</i> e <i>Control</i> com o banco de dados <i>Blood Transcriptional Modules</i>	56
Figura 15 – <i>Gene Set Enrichment Analysis</i> para comparação <i>LateDengue</i> e <i>Control</i> com o banco de dados <i>Blood Transcriptional Modules</i>	57
Figura 16 – Correlação dos lncRNAs com os respectivos genes que podem estar sendo regulados em <i>cis</i>	58
Figura 17 – Correlação entre EBNA1BP2 e o lncRNA NONHSAG001220 em quatro estudos de <i>microarray</i>	59
Figura 18 – Análise GSEA dos módulos de coexpressão.	60
Figura 19 – Análise de enriquecimento dos módulos de co-expressão	62

Figura 20 – Correlação entre os lncRNAs e módulos de co-expressão. 63

Lista de tabelas

Tabela 1 – Lista de LncRNAs com expressão alterada na comparação <i>EarlyDengue</i> <i>vs Control.</i>	51
Tabela 2 – Lista de LncRNAs com expressão alterada na comparação <i>LateDengue</i> <i>vs Control.</i>	52

Lista de abreviaturas e siglas

BTM	<i>Blood Transcriptional Modules</i>
cDNA	DNA complementar
CLI	<i>Command Line Interface</i>
DEG	<i>Differentially Expressed Genes</i>
DENV	Vírus da Dengue
DF	<i>Dengue Fever</i>
DHF	<i>Dengue Hemorrhagic Fever</i>
DNA	Ácido Desoxirribonucleico
DSS	<i>Dengue Shock Syndrome</i>
FTP	<i>File Transfer Protocol</i>
GSEA	<i>Gene Set Enrichment Analysis</i>
GTF	<i>Gene Transfer File</i>
H3K27	Vigésima sétima lisina da proteína H3 do octâmero da histona
H3K4	Quarta lisina da proteína H3 do octâmero da histona
HIV	Vírus da Imunodeficiência Humana
kb	Kilo bases
LCR	Região controladora do locus
lncRNA	RNA não codificador longo
miRNA	microRNA, RNA pequeno de aproximadamente 22 nucleotídeos
MM	<i>Mismatch</i>
mRNA	RNA mensageiro
NK	células <i>Natural Killers</i>

NS	<i>Nonstructural Proteins</i>
PBMC	<i>Peripheral Mononuclear Blood Cells</i>
PME	<i>Posterior Mean Estimate</i>
PM	<i>Perfect Match</i>
RISC	<i>RNA-induced silencing complex</i>
RMA	<i>Robust Multi-array Average</i>
RNA	Ácido Ribonucleíco
RNA-seq	Sequenciamento de DNA empregado para o estudo do transcrito
siRNA	<i>small interfering RNA</i>
SRA	<i>Sequence Read Archive</i>
ssDNA	<i>single-stranded DNA</i>
TCR	<i>T cell receptor</i>
TH	linfócito T auxiliar
THP-1	Linhagem celular extraída de um paciente com Leucemia Mieloide Aguda
TPM	<i>Transcripts per Million</i>
TSV	<i>Table Separated Values</i>
UTR	Região não traduzida do RNA
WNV	Vírus da Febre do Oeste do Nilo
YFV	Vírus da Febre Amarela
ZIKV	Vírus Zika

Sumário

1	Introdução	19
1.1	<i>Infecção pelo vírus da dengue</i>	19
1.2	<i>Vírus da dengue</i>	20
1.3	<i>Manifestações Clínicas</i>	20
1.4	<i>Infecção pelo Vírus da Dengue</i>	21
1.5	<i>Resposta Imune Induzida pela Infecção por Dengue</i>	22
1.6	<i>Análise de Transcritoma de Pacientes Infectados por dengue</i>	23
1.7	<i>RNAs não codificadores longos</i>	25
1.8	<i>Análise de Transcritomas</i>	31
1.9	<i>Análise em larga escala de lncRNAs</i>	32
2	Objetivo geral	34
2.1	<i>Objetivos Específicos</i>	34
3	Materiais e Métodos	35
3.1	<i>Estudos Utilizados</i>	35
3.2	<i>Reanotação das Sondas nos Microarrays</i>	36
3.3	<i>Classificação das Amostras</i>	38
3.4	<i>Desenvolvimento do Pipeline para Pré-processamento e Análise</i>	38
3.5	<i>Obtenção e Pré-processamento dos dados</i>	39
3.6	<i>Análise de Expressão Diferencial</i>	41
3.7	<i>Análise de Enriquecimento</i>	42
3.8	<i>Análise de Correlação entre LncRNAs e mRNAs</i>	43
3.9	<i>Análise de Redes de Co-expressão</i>	44
3.10	<i>Análise de RNA-seq</i>	45
3.11	<i>Ferramentas de Análise de Transcritoma</i>	46
4	Resultados	48
4.1	<i>Reanotação das Sondas dos Microarrays</i>	48
4.2	<i>LncRNAs Diferencialmente Expressos em Infecção por Dengue</i>	49
4.3	<i>Análise Comparativa entre Microarray e RNA-seq</i>	52

4.4	<i>Análise de Enriquecimento</i>	53
4.5	<i>Gene Set Enrichment Analysis</i>	55
4.6	<i>Correlação em Cis</i>	57
4.7	<i>Análise de Redes de Co-expressão</i>	59
5	Discussão	64
6	Conclusão	70
	Referências	71
	Anexo A – Artigos Publicados	83

1 Introdução

1.1 Infecção pelo vírus da dengue

A dengue é uma infecção viral sistêmica transmitida por mosquitos do gênero *Aedes*. Estima-se que ocorram 390 milhões de casos de dengue anualmente, sendo 96 milhões sintomáticos (BHATT et al., 2013). Segundo Sreaton et al. (2015) a mortalidade pode chegar a 20% sendo que com a reposição de fluídos pode ser reduzida para 1%. A dengue ocorre em todas as regiões tropicais do globo (figura 1). No Brasil, a doença constitui um grave problema de saúde pública. Há relatos da presença da dengue no Brasil desde o século XIX, porém o primeiro caso confirmado ocorreu na década de 80 (OSANAI et al., 1983), sendo que, desde então, o número de casos tem aumentado demasiadamente, atingindo mais de 100 mil casos novos por semana em março de 2013 (MARQUES-TOLEDO et al., 2017). De acordo com Sreaton et al. (2015), os principais motivos apontados para esta pandemia da doença são a globalização, a propagação do mosquito *Aedes*, a urbanização não planejada e a ausência de uma terapia efetiva

Figura 1 – Evidência de ocorrência de dengue no mundo. Extraído de Bhatt et al. (2013)

1.2 Vírus da dengue

O vírus da dengue (DENV) pertence à família *Flaviviridae* e mais especificamente ao gênero *Flavivirus*, o mesmo gênero do vírus Zika (ZIKV), Febre Amarela (YFV) e Oeste do Nilo (WNV). Seu genoma possui material genético codificado em uma fita simples de RNA positiva. Existem 4 sorotipos conhecidos, nomeados DENV1, DENV2, DENV3 e DENV4 (SCREATON et al., 2015).

O vírus da dengue é envelopado e seu capsídeo possui 50 nanômetros de diâmetro, com estrutura icosaédrica e três proteínas estruturais: capsídeo (C), precursor de membrana (prM) e envelope (E). A proteína E é de vital importância para a ligação e entrada na célula hospedeira. Além das proteínas estruturais o genoma do vírus da dengue codifica sete proteínas não estruturais: *non-structural protein 1* (NS1), *Flavivirus nonstructural protein 2A* (NS2A), NS2B, NS3, NS4A, NS4B e NS5 (SCREATON et al., 2015).

1.3 Manifestações Clínicas

Acredita-se que existam muitos casos assintomáticos da doença. Entretanto, a maioria dos casos de dengue se caracteriza clinicamente por febre leve autolimitada e sem complicações graves. Os pacientes que possuem esses sintomas menos graves são classificados como *Dengue Fever*, descrita deste ponto em diante como DF. Entretanto, o fenótipo clínico pode variar e depende da idade, da predisposição genética e do *background* imunológico do indivíduo infectado. Outros sintomas comuns são dores de cabeça, dor retro-orbital, dores musculares, erupção cutânea, dores abdominais e trombocitopenia (SCREATON et al., 2015).

Em casos mais graves, chamados mais recentemente pela literatura de dengue grave, a infecção viral pode levar ao aumento de permeabilidade capilar e extravasamento plasmático o que pode levar a um quadro de hemorragia. Além disso, a infecção pode levar a danos em alguns órgãos, causando hepatite, miocardite e encefalite (SCREATON et al., 2015). A dengue grave em crianças causa um risco maior de choque enquanto que em adultos há um risco maior de hemorragias e danos a órgãos (SCREATON et al., 2015). Estes casos serão denominados aqui *Dengue Hemorrhagic Fever* ou DHF. Em 2012, a taxa de incidência de dengue no Brasil era de 301,47 casos por 100.000 habitantes ¹, chegando

¹ <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?idb2012/d0203.def>

a 1.100,65 por 100.000 habitantes no estado do Rio de Janeiro (Ministério da Saúde - Datasus). Neste mesmo ano, houveram 1.063 ² casos confirmados de febre hemorrágica da dengue (Ministério da Saúde - Datasus), com 159 casos no estado do Rio de Janeiro. No Brasil ocorreram 173 ³ óbitos comprovadamente causados por dengue em 2012, sendo que em apenas 3 anos esse número subiu para 570 em 2015.

Numa fração pequena de pacientes infectados por dengue pode ocorrer a síndrome de choque, potencialmente fatal, chamada assim devido ao choque hipovolêmico resultante do extravasamento plasmático. Esta condição será descrita aqui pela sigla DSS, do inglês *Dengue Shock Syndrome* (BHATT et al., 2013). Nestes casos, o monitoramento diário de plaquetas (LAM et al., 2013) e a terapia com fluido intravenoso para balancear o extravasamento plasmático (LAM et al., 2017) têm sido importantes para diminuir a taxa de mortalidade por DSS.

Recentemente foram realizados vários ensaios clínicos para drogas contra o vírus da dengue, porém nenhuma delas obteve o sucesso clínico desejado (SCREATON et al., 2015). Vacinas contra dengue estão sendo desenvolvidas e uma já foi registrada na National Regulatory Authorities, sendo inclusive recomendada para uso na população (World Health Organization, 2017). No entanto, na maioria dos casos, o tratamento da dengue ainda se resume a cuidados de suporte, como reposição de fluidos intravenosos e monitoramento.

1.4 Infecção pelo Vírus da Dengue

O DENV entra na célula por endocitose mediado por clatrina e outros fatores de ligação, que invagina a membrana celular, formando o endossomo. No ambiente acidificado do endossomo a Glicoproteína E sofre rearranjo conformacional irreversível, expondo o peptídeo de fusão e iniciando a fusão entre a bicamada lipídica do vírus e a membrana do endossomo. O capsídeo viral é então liberado para dentro do citoplasma e o RNA viral é separado do capsídeo por desnudamento. A maquinaria celular inicia o processo de tradução do material genético, o qual é transportado para o retículo endoplasmático (RE). No Complexo de *Golgi* ocorre a montagem de partículas virais (MUKHOPADHYAY; KUHN; ROSSMANN, 2005).

² <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?idb2012/d0110.def>

³ <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/obt10uf.def>

Em seguida, o complexo de replicação é montado na célula com o RNA viral e as proteínas NS (*Nonstructural Proteins*) (TEO; CHU, 2014). O RNA viral (+) molde é transcrito em RNA (-), que servirá como molde para nova síntese de RNA (+). No retículo endoplasmático ocorre o empacotamento e, através da protease furina, os vírions (partículas virais maduras) são liberados para fora da célula por exocitose (OSANAI et al., 1983).

O DENV é capaz de infectar diversos tipos celulares, especialmente monócitos, macrófagos e células dendríticas. Dentro das células infectadas, vias de sinalização como a do *interferon* tipo I, estresse do Retículo Endoplasmático, pró-inflamatórias e ativadas por receptores do tipo toll são alteradas (KWISSA et al., 2014b).

Novas vias moleculares que diferenciam a gravidade do acometimento por dengue estão sendo descobertas. Por exemplo, foi mostrado que a infecção por dengue pode modular vias relacionadas ao metabolismo energético (ALLONSO et al., 2015), organização da matriz extracelular e ubiquitinação (KANLAYA et al., 2010). De fato, foi recentemente mostrado que o processo da liberação do genoma do vírus (*uncoating*) requer a ubiquitinação (BYK et al., 2016).

1.5 Resposta Imune Induzida pela Infecção por Dengue

O sistema imune inato e adaptativo possuem um papel fundamental em conter a infecção. Estes sistemas irão atuar juntos para promover a morte de células infectadas e a neutralização dos vírus.

Inicialmente, há a ativação de uma cascata de sinalização por Receptores de Padrões Moleculares Associados a Patógenos (PAMP, sigla do inglês, *Pathogen Associated Molecular Pattern*), *toll-like receptors* (TLR), como TLR3 e TLR5, que são expressos em células mielóides, Lectinas do tipo-C, como *DC-Sign*, CLEC5, entre outras. Esta cascata culmina na ativação de diversos fatores de transcrição que induzem a produção de *interferon*. A resposta ao DENV mediada por *interferon* leva ao controle metabólico e limita a infecção viral. Porém o DENV é capaz de inibir a resposta inata antiviral mediada por *interferon-1* através das proteínas não estruturais NS2A, NS4A, NS4B e NS5, responsáveis por bloquear a sinalização por *interferon*, reduzindo a ativação de STAT (RODENHUIS-ZYBERT; WILSCHUT; SMIT, 2010).

A resposta antiviral mediada por *interferon* também promove a resposta imune adaptativa através da estimulação da maturação de células dendríticas e pela ativação das células B e T (RODENHUIS-ZYBERT; WILSCHUT; SMIT, 2010). Ainda é controversa a relevância das células T CD4+ e CD8+ no controle da patogênese do vírus, mas é sabido que há um aumento significativo destes tipos celulares principalmente em uma infecção secundária de dengue (SCREATON et al., 2015) .

Células B, no entanto, são claramente importantes em produzir anticorpos contra o vírus da dengue. Em geral, os anticorpos possuem como alvo as glicoproteínas M e E presentes na superfície do vírus. Foi também observado que anticorpos contra a proteína NS1 podem levar à lise das células infectadas (RODENHUIS-ZYBERT; WILSCHUT; SMIT, 2010). A neutralização do vírus por anticorpos depende principalmente do número de anticorpos ligados a ele e da afinidade do anticorpo por ele. No caso de um anticorpo com alta afinidade é necessário um número menor de anticorpos para que haja a neutralização do vírus. Este processo ocorre através do bloqueio da ligação do vírus ao receptor natural na superfície celular. A presença de anticorpos ligados ao vírus pode levar a sua internalização através de receptores para a fração cristalizável do anticorpo (FcR) em monócitos, macrófagos e células dendríticas, o que pode levar a maior carga viral, efeito conhecido pelo termo em inglês *Antibody-Dependent Enhancement* (ADE) (RODENHUIS-ZYBERT; WILSCHUT; SMIT, 2010). Acredita-se que o efeito ADE é um dos fatores para que infecções secundárias por sorotipos heterólogos tenham uma chance maior de desenvolvimento de dengue grave (RODENHUIS-ZYBERT; WILSCHUT; SMIT, 2010).

1.6 *Análise de Transcritoma de Pacientes Infectados por dengue*

O transcritoma é definido como o conjunto de todos os transcritos de uma célula ou tecido. Diversos estudos já analisaram os perfis de expressão gênica de pacientes infectados por dengue (DEVIGNOT et al., 2010), (GOMES et al., 2010), (HOANG et al., 2010), (KWISSA et al., 2014a), (LOKE et al., 2010), (LONG et al., 2009), (SIMMONS et al., 2007), (UBOL et al., 2008), (WARKE et al., 2008), (NASCIMENTO et al., 2009), (SUN et al., 2013).

O estudo de Simmons et al. (2007) avaliou a expressão gênica de amostras de células mononucleares do sangue periférico (PBMC) extraídas de quatorze pacientes adultos vietnamitas com infecção secundária através da tecnologia de *microarrays* de cDNA. As amostras foram extraídas um dia após a admissão do paciente no hospital e após um mês no período de convalescença. O estudo identificou a relação entre a infecção aguda pelo vírus da dengue e resposta ao stress do retículo endoplasmático e atividade mitótica. Além disso, um conjunto de 24 genes com expressão diminuída em pacientes que desenvolveram DSS foi identificado, sendo que muitos desses são genes estimulados por *interferon-1* e já possuem relação com resposta antiviral, como o gene MX1 e ISG15.

No estudo de Devignot et al. (2010), os autores compararam o transcrito sanguíneo de pacientes pediátricos com DF contra DHF e DSS utilizando a tecnologia de *microarray*. Eles propuseram que haveria um segundo ciclo de amplificação de inflamação que leva a uma síndrome inflamatória sistêmica levando a falência da homeostase vascular e da microcirculação, a alteração do balanço entre as respostas de linfócito T e resposta pró-inflamatória do sistema imunológico inato exacerbada.

Sun et al. (2013) avaliaram a expressão gênica de 64 indivíduos em diferentes tempos da infecção por dengue. Amostras de sangue foram coletadas diariamente do início dos sintomas até 10 dias, além de uma amostra obtida no período de convalescença. As análises realizadas revelaram duas fases distintas do processo da doença, sendo a fase inicial até 4 dias e a fase tardia da doença após este período. Considerando como referência as amostras convalescentes, os autores realizaram análises de expressão diferencial e anotação funcional. Estas análises revelaram que os genes diferencialmente expressos no estágio inicial da doença estão mais relacionados à resposta antiviral, o que coincide com o momento de maior viremia. Por outro lado, no estágio tardio os genes diferencialmente expressos estão mais relacionados à recuperação pós infecção. Muitos estudos mostram que quase todos os tipos celulares sofrem apoptose quando infectados por DENV, e há estudos mostrando a apoptose de células T após sua ativação durante a fase aguda da doença. Adicionalmente, há estudos mostrando a restauração da contagem de células T na fase tardia da dengue. No estudo de Sun et al. (2013) os genes diferencialmente expressos na fase tardia da doença estão relacionados com síntese de DNA e ciclo celular mitótico, sugerindo a recuperação da imunidade celular, o que está de acordo ao descrito na literatura. Sun et al. (2013) identificaram também a associação de IFN- γ e imunidade por célula T com menor risco de ocorrência de DHF.

Afroz et al. (2016) realizaram uma metanálise utilizando dados de expressão gênica de pacientes com dengue disponíveis na literatura a qual revelou a alteração durante a infecção de diversos genes relacionados a sinalização via *interferon* e desintoxicação de espécies reativas à oxigênio. Além disso, os genes PATJ e CRTAP, envolvidos na manutenção da integridade das junções celulares e montagem de colágeno, estavam com a expressão diminuída nos pacientes. Estes achados foram validados na linhagem celular THP-1 infectada com dengue e foi demonstrado que o aumento da expressão de CRTAP e PATJ podem restringir a infecção por dengue.

Banerjee et al. (2017) utilizaram a técnica de *RNA-seq* para avaliar o perfil transcripcional de amostras provenientes de células mononucleares do sangue periférico de pacientes com dengue e compararam com amostras extraídas de indivíduos normais e pacientes sofrendo de outras doenças. Eles identificaram que a assinatura transcripcional que diferencia os pacientes com dengue está relacionada a vias envolvendo metabolismo de glicina, serina e treonina, organização da matriz extracelular, ubiquitinação, citocinas e resposta inflamatória. Além disso, diversos genes com maior expressão em pacientes com dengue são conhecidos por facilitar a migração de leucócitos, ativação e degranulação de neutrófilos, entre eles, MPO, RNASE2, RNASE3, DEFA3, DEFA4 e ELANE.

Apesar de todos estes estudos e outros similares já publicados, ainda não foi sistematicamente avaliado na literatura o papel dos RNAs não codificadores longos (lncRNAs) na progressão da dengue.

1.7 RNAs não codificadores longos

Recentemente ficou claro que grande parte do transcrito é composto por RNAs que não são traduzidos em proteínas. RNAs não codificadores longos (lncRNAs) representam uma classe de transcritos de RNA com mais de 200 nucleotídeos que não são traduzidos em proteína mas que podem ter função regulatória. Grande parte destas moléculas são similares a RNAs mensageiros (mRNAs), ou seja, são transcritos pela RNA polimerase II, possuem adição de *cap 5'*, poliadenilação e podem sofrer *splicing* (NAKAYA et al., 2007).

Os lncRNAs possuem diversos mecanismos de atuação, podendo participar do remodelamento da cromatina (WANG et al., 2011), da formação de complexos proteicos (FENG et al., 2006), (TRIPATHI et al., 2010), (WONG et al., 2007) ou impedindo a ligação

de microRNAs a mRNAs (MORAN; PERERA; KHALIL, 2012). Além disso, sabe-se que os lncRNAs antissenso (localizados na fita oposta de DNA ao gene codificador de proteína) podem atuar na regulação dos mRNAs transcritos no mesmo locus genômico (MORAN; PERERA; KHALIL, 2012).

A biogênese, regulação e processamento dos lncRNAs se assemelha aos genes codificadores de proteínas. Ou seja, a maioria são transcritos pela RNA polimerase II, são regulados por fatores de transcrição e modificação da cromatina, podem sofrer *splicing*, possuem *cap 5'* e sofrem poliadenilação (LEEUEWEN; MIKKERS, 2010). Após transcrito, o lncRNA pode agir tanto no núcleo quanto no citoplasma e até mesmo mediar o *cross-talk* entre o núcleo e certas organelas (DONG et al., 2017). Além disso, há evidências que lncRNAs são secretados em vesículas (GEZER et al., 2014), agem em outras células (HEWSON et al., 2016), podem percorrer a corrente sanguínea (WANG et al., 2016) e ser potenciais biomarcadores de doenças (SUN et al., 2017) (TIAN et al., 2017).

Os lncRNAs podem atuar em diferentes níveis da regulação da expressão gênica, regulando mecanismos epigenéticos, processamento de mRNA, e eventos pós-transcricionais (MERCER; DINGER; MATTICK, 2009) (GEISLER; COLLIER, 2013) (KUNG; COLOGNORI; LEE, 2013) (MAASS; LUFT; BÄHRING, 2014). A regulação epigenética por lncRNAs pode ocorrer pelo deslocamento de proteínas da cromatina (O'LEARY et al., 2015), recrutamento de complexo mediador para uma região *enhancer* onde ocorre a formação e estabilização de *loop* (YANG et al., 2013), interferência na interação entre a região do *enhancer* com o promotor inibindo a transcrição do gene alvo, e recrutamento de proteínas que modificam a cromatina em regiões específicas no genoma (GROTE; HERRMANN, 2013).

A regulação transcricional se dá principalmente pela modulação do processamento do mRNA. Os lncRNAs podem modular assinaturas da cromatina e alterar os padrões de *splicing* (WOHLKONIG et al., 2010). Além disso, o *splicing* alternativo pode ser regulado pela modulação das proteínas responsáveis pelo *splicing* (TRIPATHI et al., 2010)(figura 2 G). Os lncRNAs também podem interagir diretamente com a RNA polimerase II e inibir a transcrição de alguns mRNAs (MARINER et al., 2008).

A regulação pós-transcricional dos lncRNAs pode ocorrer por várias formas. Elementos repetitivos Alu em lncRNAs, quando associados à região 3' UTR do mRNA alvo, direcionam o mRNA à degradação (GONG; MAQUAT, 2011).

Além disso, lncRNAs podem recrutar proteínas como STAU1, estabilizar o mRNA e consequentemente previne a sua degradação (KRETZ et al., 2012) (figura 2). Outro mecanismo pelo qual lncRNAs diminuem a degradação dos mRNAs alvos é o mascaramento de sítios de ligação com miRNAs bloqueando o silenciamento do mRNA pelo complexo RISC (FAGHIHI et al., 2010).

Além disso, alguns lncRNAs funcionam como sequestradores de proteínas (LEE et al., 2016), que regulam a estabilidade genômica. LncRNAs também podem se ligar a miRNAs (HANSEN et al., 2013), inibindo ou diminuindo a disponibilidade dessas moléculas e consequente redução de interação com os mRNAs alvos.

A presença de elementos repetitivos na sequência do lncRNA pode aumentar a tradução do mRNA alvo pela associação com a região 5' linha do mesmo (CARRIERI et al., 2012). Por fim, já foi descrito que os lncRNAs podem atuar como proteínas de sustentação em complexos de proteínas (CARMO-FONSECA; RINO, 2011) (figura 2 E).

Figura 2 – Representação de mecanismos de regulação gênica por lncRNAs. A) Os transcritos de lncRNA deslocam proteínas da cromatina. B) lncRNAs recrutam o complexo Mediador para uma região de *enhancer*, estabilizando a formação de loop e transcrição do gene associado. C) lncRNAs interferem na interação entre a região do *enhancer* com o promotor, inibindo assim a transcrição do gene alvo. D) lncRNAs recrutam proteínas, tais como complexos que modificam a cromatina para locais alvo específicos no genoma. E) lncRNAs atuam como proteínas de sustentação que liga diferentes proteínas que necessitam de uma ação conjunta. F) lncRNAs ligam e captam proteínas para inibir ou diminuir a disponibilidade do alvo, por exemplo na ligação aos mRNAs (esquerda); circRNA sequestrando miRNAs para evitar sua ligação aos mRNAs (direita). G) lncRNAs afetam *splicing* alternativo, se ligando ao transcrito primário e alterando o padrão de *splicing*. H) lncRNAs recrutando proteínas, estabilizam e previnem a degradação de mRNAs. Adaptado de Schmitz, Grote e Herrmann (2016).

Na infecção por vírus, já foi mostrado que lncRNAs podem desempenhar papéis importantes na regulação gênica (FORTES; MORRIS, 2016). Células infectadas podem expressar lncRNAs virais, celulares e quiméricos (LAU et al., 2014). Há evidências que os lncRNAs virais, como o HSUR1 (CAZALLA; YARIO; STEITZ, 2010) e o EBER2 (LEE et al., 2015) podem diminuir a expressão de miRNA da célula hospedeira. Além disso, lncRNAs podem regular a epigenética da transcrição viral em infecções como HIV (SAAYMAN et al., 2014).

Já foi estabelecida a relação entre lncRNAs e o sistema imunológico. Diversas evidências sugerem que os lncRNAs são expressos de maneira específica em cada tipo celular e, desta forma, possuem papel importante nos programas de desenvolvimento que dão origem às diferentes células do organismo (AUNE; Spurlock Iii, 2016). Hu et al. (2013) mostraram isso para as células T em todos os estágios de desenvolvimento e diferenciação. Por outro lado, a maior parte dos mRNAs são expressos de maneira não específica. O lncRNA IFNG-AS1 (NeST) foi um dos primeiros lncRNAs descobertos com expressão em células do sistema imunológico, ele é expresso em células T CD4+, T CD8+ e *Natural Killers* (NK). Como o nome sugere, este gene está presente na fita oposta do gene IFNG e é expresso em resposta a sinais de diferenciação TH1. O IFNG-AS1 atua aparentemente recrutando enzimas H3K4 metil-transferases e criando um ambiente favorável a transcrição nas regiões promotoras e *enhancers* do gene IFNG. Em particular, o lncRNA IFNG-AS1 confere resistência a infecções letais por bactérias (AUNE; Spurlock Iii, 2016).

Spurlock et al. (2015) avaliaram através da técnica de RNA-seq lncRNAs expressos de maneira específica em linhagens TH1, TH2 ou TH17. A maioria dos mRNAs expressos de maneira específica estão adjacentes a lncRNAs, sendo que os mRNAs são co-expressos com os lncRNAs. Desta forma, os lncRNAs podem estar regulando o perfil de expressão específico de cada linhagem que é essencial para que as células efectoras TH1, TH2 e TH17 executem suas funções imunológicas únicas.

Um dos lncRNAs identificados por Spurlock et al. (2015), foi o TH2-LCR, nomeado assim por possuir expressão específica em células TH2 e estar localizado na região controladora de locus (LCR) em camundongos. A depleção do aglomerado TH-lncRNA através de siRNAs suprimiu a expressão de citocinas TH2, sendo elas, IL4, IL-13 e IL-5. O lncRNA TH2-LCR se associa a WDR5 que é um componente do complexo H3K4 metil-transferase. Sendo assim, a diminuição dos níveis do TH2-LCR suprime o recrutamento de WDR5

nas regiões genômicas dos promotores e *enhancers* dos genes IL-4 e IL-13, bloqueando a trimetilação nesses locais.

Linc-MAF-4 é um lncRNA que está a aproximadamente 150 kb de distância do gene codificador de proteína MAF e leva à sua inibição através da trimetilação de H3K27 no loci do gene MAF o que é realizado em conjunto com os genes LSD1 e EZH2. A inibição do gene MAF pelo Linc-MAF-4 leva à diferenciação de TH1 e a expressão de IFNG, enquanto que a expressão de MAF está envolvida na diferenciação de linhagens TH2 (RANZANI et al., 2015).

Outro lncRNA interessante é o NRON (*non-coding RNA repressor of NFAT*), o qual está relacionado à regulação da atividade do fator de transcrição NFAT, um dos responsáveis pela regulação da expressão da IL-2. NRON, NFAT e algumas quinases formam uma ribonucleoproteína presente no citoplasma. Quando há a desfosforilação da NFAT, o complexo ribonucleoproteico entra no núcleo e leva à ativação da expressão de seus genes alvos. A depleção de NRON aparentemente leva à desfosforilação de NFAT. É importante notar que o NRON, diferentemente dos lncRNAs citados anteriormente, não está presente no núcleo, não está localizado próximo do gene NFAT e sua atuação não depende de mecanismos epigenéticos (AUNE; SPURLOCK, 2016).

FAS-AS1 é um lncRNA que controla o *splicing* alternativo do gene FAS ao se ligar a um RBP (*RNA binding protein*) nomeado RBM5 e inibindo o *exon skipping* do mRNA da FAS. A ausência desse *exon* leva à solubilização da proteína FAS e faz com que ela se encontre no citoplasma ao invés da membrana celular. O FAS-AS1 então se liga à FASL, impedindo que a FASL se ligue à FAS presente na membrana, o que leva a resistência à apoptose mediada por FASL. O lncRNA FAS-AS1 pode estar relacionado a linfomas primários de células B, já que apresenta níveis de expressão menores nestes linfomas (AUNE; SPURLOCK, 2016).

lnc-DC promove a diferenciação de monócitos em células dendríticas convencionais se ligando ao fator de transcrição STAT3 e inibindo a sua desfosforilação pela tirosina fosfatase SHP1, levando desta forma à translocação do citoplasma para o núcleo, permitindo a atividade da STAT3 (AUNE; SPURLOCK, 2016).

1.8 Análise de Transcritomas

Uma verdadeira revolução aconteceu nas últimas décadas com a crescente utilização de tecnologias de larga escala (*high-throughput technologies*) em estudos de biologia molecular. Tecnologias como espectroscopia de massa, *microarrays* e sequenciamento de última geração permitem medir, em paralelo, os níveis de quase todos os componentes biológicos em diferentes condições experimentais.

Com a análise do transcrito é possível inferir as vias de sinalização ativadas ou inibidas por infecção (WESTERMANN et al., 2016), doenças inflamatórias (KELSEN et al., 2015), câncer (ROBINSON et al., 2017) ou tratamento com drogas (GRAVITZ, 2014). Também é possível estudar mecanismos regulatórios relacionados com processamento alternativo de RNA (SESSIONS et al., 2013), miRNAs (ETEBARI et al., 2015) e RNAs não codificadores (BAYTAK et al., 2017). Finalmente, as análises de transcrito podem ser usadas para identificar biomarcadores que permitam prever respostas biológicas a tratamento (JULIA et al., 2009), progressão de doença (LEE et al., 2017) ou até eficácia de vacinas (THOREAU et al., 2015). Desde 1998, os estudos do transcrito publicados e indexados no PubMed ultrapassaram o número de 100.000 artigos (figura 3).

Figura 3 – Número de artigos sobre técnicas de larga escala publicados e indexados no PubMed. Foram usados os termos [MeSh] “*Gene Expression Profiling*” (Transcritoma), “*Oligonucleotide Array Sequence Analysis*” (Microarrays) e “*High-Throughput Nucleotide Sequencing*” (RNA-seq).

Das técnicas capazes de medir o transcrito, a tecnologia de *microarray* foi a mais utilizada com mais de 80.000 artigos (figura 3, linha marrom). Essa tecnologia consiste em imobilizar milhares de oligonucleotídeos de DNA simples fita (ssDNA, do inglês *single-stranded DNA*) em um slide de vidro organizada em forma de grade. Também denominados sondas, os ssDNAs possuem sequências complementares aos mRNA dos genes. Ao isolar o mRNA de uma amostra, converter para cDNA e submeter ao processamento de medição por *microarray* é possível visualizar grande parte do transcrito em um único procedimento.

Entretanto, a medição do transcrito por *microarray* possui algumas limitações como: artefatos de hibridização cruzada, quantificação imprecisa para genes com expressão muito baixa ou muito alta, e a necessidade de se conhecer a sequência *a priori* (GROEN, 2001). A existência de oligonucleotídeos controles que auxiliam na correção de *mismatch* e *background effect* minimizam alguns problemas técnicos.

A redução do custo das tecnologias de sequenciamento de larga escala ou de alto desempenho, tornou a técnica de RNA-seq o melhor método para realizar a quantificação do transcrito (figura 3, linha cinza). A tecnologia consiste em quantificar a expressão gênica através do sequenciamento massivo das moléculas de RNA.

A publicação de artigos científicos com resultados de transcrito exigem que os autores sigam as normas do *Minimum Information about a Microarray Experiment* (MIAME) para a coleta de dados e análise dos experimentos. Além disso, frequentemente exige dos autores a submissão dos dados de expressão brutos e/ou processados, dados das amostras e o desenho experimental a bancos de dados públicos (BALL et al., 2004).

Os principais bancos de dados públicos com dados de transcrito são o *Gene Expression Omnibus* (GEO) e ArrayExpress. Até o mês de julho de 2017, o banco de dados GEO já possuía 85.447 estudos de diversas plataformas, totalizando 2.098.626 de amostras analisadas, enquanto o *Array Express* incluía 70.124 experimentos de genômica funcional *high-throughput*.

1.9 Análise em larga escala de lncRNAs

A técnica de *microarray* já foi utilizada para analisar o perfil transcricional de lncRNAs (FENG; ZHANG,) (ZHANG et al., 2015) (LIAO et al., 2011). Nesses estudos,

os autores tiveram que identificar as sondas dos *microarrays* que representavam lncRNAs. Esta re-anotação consiste em basicamente comparar a posição genômica ou sequência da sonda com regiões ou sequências conhecidas de lncRNAs. Após a correta re-anotação das sondas é possível aplicar as mesmas técnicas usadas para analisar genes codificadores de proteína. Com o advento do método de RNA-seq, as análises de expressão de lncRNAs não precisam mais ser limitadas a sondas disponíveis no *microarray*. Essa tecnologia foi utilizada para analisar os lncRNAs em diversos contextos incluindo câncer (IYER et al., 2015), infecção (BRUEGGE; EINSPANIER; SHARBATI, 2017) e doenças inflamatórias (IIOTT et al., 2014). No entanto, nenhum estudo analisou lncRNAs em amostras infectadas com dengue. Encontrar tais lncRNAs poderiam revelar aspectos importantes da regulação gênica durante a infecção pelo vírus da dengue. Para isso, é possível reanalisar os dados de transcriptoma de pacientes com dengue com maior ênfase nos lncRNAs.

2 Objetivo geral

Utilização de dados de *microarrays* para identificar possíveis mecanismos de regulação gênica envolvendo RNAs Não Codificadores Longos em momentos distintos da infecção por dengue em seres humanos.

2.1 *Objetivos Específicos*

1. Identificar genes diferencialmente expressos de forma consistente entre estudos independentes no momento agudo inicial e agudo tardio da infecção;
2. Dentre os transcritos diferencialmente expressos, identificar lncRNAs;
3. Inferir o impacto dos lncRNAs na expressão de genes codificadores de proteína.

3 Materiais e Métodos

3.1 Estudos Utilizados

Selecionamos estudos de transcriptoma realizados com amostras de sangue ou células mononucleares do sangue periférico (PBMC) extraídas de pacientes infectados com dengue do banco de dados do GEO. Foram incluídos apenas estudos provenientes de pacientes convalescentes e que disponibilizaram dados brutos para análise. Foram excluídos estudos que não eram passíveis de realizar o processo de re-anotação de sondas. O número de amostras utilizadas está representado na figura 4.

O conjunto de sondas a ser incluído em uma lâmina comercial de *microarray* é definido pelos critérios de desenho de cada empresa. Além disso, as sondas (*probes*) são organizadas de forma diferente de acordo com as tecnologias empregadas na fabricação do *microarray*. Chamamos aqui a tecnologia de *microarray* utilizada e o desenho específico do *microarray* como plataforma de *microarray*, identificada por GPL e um número no banco de dados GEO. Os quatros estudos selecionados utilizam diferentes plataformas de *microarray*. O estudo GSE43777 utiliza a plataforma GPL570 (*Affymetrix HG-U133-plus2*), a qual possui 54.675 conjuntos de sondas (em inglês, *probe set*). Cada conjunto de sonda, classificado como “Perfect Match” (PM), representa um mesmo transcrito e possui em média 11 sondas de 22 nucleotídeos presentes na região 3’ do transcrito a ser medido. Além disso, para fins de controle a plataforma GPL570 inclui conjuntos de sondas classificados como “Mismatch” (MM) que não hibridizam com o transcrito sendo medido. O estudo GSE5180 utiliza a plataforma GPL13158 que é similar à plataforma GPL570 porém não apresenta as sondas “Mismatch” (MM). Apesar das plataformas GPL570 e GPL13158 utilizarem um conjunto de sondas como representante de um transcrito os chamamos, por simplicidade, de sondas. Já os estudos GSE13052 e GSE28405 utilizam a plataforma GPL2700 (*Sentrix HumanRef-8 Expression BeadChip*) de *microarray* comercializada pela empresa Illumina.

Figura 4 – Número amostras utilizadas em cada estudo (sem amostras *outliers*). A) Considerando somente as classes *Control* e *Dengue*. B) Considerando as classes *DSS* (síndrome do choque da dengue), *DHF* (dengue hemorrágica), *DF* (dengue sem complicações), *Dengue* (contendo *DF*, *DHF* e *DSS*), *Convalescent* (amostras convalescentes) e *Healthy* (contendo amostras saudáveis)

3.2 Reanotação das Sondas nos Microarrays

Realizamos a reanotação das plataformas GPL2700 e GPL570 para melhor identificar as sondas que representam os lncRNAs utilizando as versões mais atuais dos bancos de dados disponíveis (figura 5). De maneira sucinta, a reanotação consiste em confirmar ou designar novos transcritos (*e.g.* genes codificadores, lncRNAs, etc) para as sondas desta plataforma, por meio da intersecção entre suas coordenadas genômicas.

Nós utilizamos quatro bancos de dados contendo lncRNAs, sendo eles o *Gencode* (HARROW et al., 2012), *Noncode* (XIE et al., 2014), *LNCipedia* (VOLDERS et al., 2013) (VOLDERS et al., 2015) e *MiTranscriptome* (IYER et al., 2015).

Figura 5 – Fluxograma representando as etapas do processo de reanotação das sondas de *microarray*. Arquivos estão representados pelas formas em azul claro e programas e processos por formas em azul escuro.

A figura 5 representa esquematicamente a etapa de reanotação das sondas do *microarray*. O primeiro passo da reanotação foi obter as coordenadas genômicas dos transcritos dos quatro bancos, utilizando a montagem HG38 do genoma humano. Os banco de dados *Gencode* e *Noncode* já utilizam esta montagem, permitindo partir diretamente dos arquivos GTF (*Gene Transfer File*) disponibilizados. As coordenadas genômicas do *LNCipedia* foram obtidas alinhando com a ferramenta *BLAT* (KENT, 2002) as sequências dos transcritos em formato *fasta* no genoma HG38. Para o banco de dados *MiTranscriptome* não pudemos realizar o alinhamento das sequências contra o HG38, já que as sequências não são disponibilizadas. Neste caso, utilizamos a ferramenta *liftOver* (HINRICHS et al., 2006) para fazer a conversão das coordenadas da montagem HG19 para HG38.

A seguir, as coordenadas genômicas das sondas de cada plataforma foram obtidas através do alinhamento com a ferramenta *BLAT* (KENT, 2002), no genoma hg38. Para a plataforma GPL570 utilizamos as sequências alvos (arquivo *fasta*: *Target Sequences*)

disponibilizadas no *site* da empresa *Affymetrix*, que são as sequências do qual as sondas são desenhadas. Para a plataforma GPL2700 utilizamos as sequências das sondas disponibilizadas no FTP da empresa *Illumina*.

Finalmente, realizamos a intersecção das coordenadas obtidas de cada banco de dados com as coordenadas das sondas utilizando a ferramenta *intersectBed* do *BedTools* (QUINLAN; HALL, 2010). Para a intersecção ser considerada válida, nós exigimos que houvesse pelo menos 60% de sobreposição da sonda com um transcrito do banco. Designamos um transcrito para cada sonda utilizando a seguinte ordem de preferência: *Gencode*, *Noncode*, *LNCipedia* e *MiTranscriptome*. Ou seja, caso haja a intersecção entre um transcrito do *Gencode* e uma sonda então esta sonda é “anotada” como o transcrito do *Gencode*. Caso contrário, procuramos por um transcrito no *Noncode* e assim por diante. Para mostrar o resultado de maneira esquemática, carregamos as coordenadas dos lncRNAs e das sondas no *GenomeBrowser*¹ e avaliamos de forma manual dezenas de sobreposições.

3.3 Classificação das Amostras

As amostras provenientes de indivíduos que não possuem qualquer tipo de infecção ou estão em período de convalescença, pelo menos 14 dias após o início dos sintomas, foram classificadas como Controle. Já o grupo dengue possui amostras provenientes de indivíduos infectados pelo vírus da dengue e estão na fase aguda da doença. O grupo dengue foi classificado em Agudo Inicial ou *EarlyDengue* para amostras que foram extraídas até 3 dias após o início dos sintomas e as outras amostras foram classificadas como Agudo Tardio ou *LateDengue*.

3.4 Desenvolvimento do Pipeline para Pré-processamento e Análise

O *pipeline* para o pré-processamento e análise das amostras de *microarray* foi desenvolvido usando as linguagens de programação Bash, Python e R. Os *scripts* escritos em linguagem Bash, são responsáveis por gerenciar a execução do *pipeline* e executam os *scripts* desenvolvidos em R e Python. Os *scripts* escritos em R e Python realizam passos específicos do *pipeline* e podem ser aplicados em qualquer análise de dados de *microarray*.

¹ <https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38>

A fim de facilitar o uso destes *scripts* em contextos diferentes, o pacote *docopt* foi utilizado, permitindo desenvolver interfaces de linha de comando (CLI) padronizadas.

Todo o código desenvolvido para a realização das análises está presente no repositório DengueAnalysis do usuário matheuschburger presente no GitHub ². O *script run.sh* é responsável por executar todos os passos necessários para a análise.

3.5 Obtenção e Pré-processamento dos dados

A etapa de obtenção e pré-processamento dos dados está esquematizada na figura 6.

Figura 6 – Fluxograma de obtenção e pré-processamento de dados transcritômicos do banco de dados GEO. As setas tracejadas indicam o retorno para um passo anterior.

Inicialmente as amostras são obtidas pelo *script getData.sh* que irá executar o *script get-geo.R* para cada estudo analisado. O *script get-geo.R* faz uso do pacote *GEOquery* disponibilizado pelo projeto Bioconductor (GENTLEMAN et al., 2004), que é uma interface simples para a obtenção dos dados do GEO (SEAN; MELTZER, 2007). Os seguintes dados são obtidos e salvos:

- anotação das sondas utilizadas no experimento;
- dados de expressão gênica fornecidos pelo autor do estudo;
- anotação das amostras;

² <https://github.com/matheuschburger/DengueAnalysis/>

- arquivos suplementares e
- dados brutos.

Para as análises posteriores utilizamos os dados brutos, ou seja, dados provenientes da leitura do *scanner* sem nenhum tipo de pré-processamento. Os estudos que utilizam a plataforma GPL570 de *microarray*, fornecida pela empresa *Affymetrix*, oferecem arquivos do tipo CEL contendo os dados brutos. Os estudos, GSE28405 e GSE13052, que utilizam plataformas de *microarray* fornecidas pela empresa *Illumina*, oferecem os dados brutos em arquivos de texto. Para obter dados em formato tabular desenvolvemos os *scripts* em *Python*, *getDataGSE28405.py* e *getDataGSE13052.py*.

A etapa de pré-processamento dos dados é inicializada pelo *script* *preprocess.sh*. Este *script* executa os *scripts* *preprocess_Affymetrix.sh*, *preprocess_Illumina.sh* e *final_preprocess.sh*.

Os *scripts* *preprocess_Illumina.sh* e *preprocess_Affymetrix.sh* são responsáveis por realizar a avaliação da qualidade, detecção de *outliers* e realizar a remoção de *outliers*. Estes passos são realizados antes e depois do processo de normalização dos dados de maneira iterativa, ou seja, caso uma amostra *outlier* seja removida os passos anteriores serão repetidos. Nos estudos que utilizam *microarrays* provenientes da empresa *Affymetrix* o método de normalização RMA (*Robust Multi-array Average*) (IRIZARRY et al., 2003) disponibilizado pelo pacote *affy* (GAUTIER et al., 2004) foi utilizado. Nos estudos que utilizam *microarrays* provenientes da empresa *Illumina*, o método de normalização *Quantile* foi empregado (BOLSTAD et al., 2003). O produto final desta etapa é uma tabela de expressão gênica (extensão TSV) contendo as sondas nas linhas e amostras nas colunas.

Na etapa final de pré-processamento realizada pelo *script* *final_preprocess.sh*, será incluída na tabela de expressão gênica o identificador relativo a cada sonda proveniente da re-anotação das sondas. Para evitar problemas devido a leituras abaixo do limite de detecção do *scanner*, 20% das sondas com menor expressão são removidas. Como produto desta etapa duas tabelas são obtidas, uma contendo a expressão de cada sonda e outra contendo a expressão de cada gene. Já que existem múltiplas sondas para um mesmo gene, a sonda com maior expressão média em todas as amostras é escolhida como representante de um gene, pois a medida das sondas com menor expressão pode ser menos confiável devido ao limite de detecção do *scanner* empregado.

3.6 Análise de Expressão Diferencial

A detecção de genes diferencialmente expressos entre pacientes infectados e saudáveis fornece pistas importantes sobre o que acontece durante a infecção por dengue. O Teste T de *Student* permite avaliar se há diferença entre a média de expressão de um gene entre duas classes. Porém, pode ocorrer que a estimativa da variância seja próxima de zero puramente ao acaso levando erroneamente a altos valores absolutos de T e consequentemente a falsos positivos (TARCA; ROMERO; DRAGHICI, 2006). Sendo assim, utilizamos o pacote em R *LIMMA* (SMYTH, 2005) para encontrar sondas diferencialmente expressos (DE). Este teste pode ser considerado de forma simplificada como um teste T moderado, o qual utiliza a informação de outras sondas para uma estimativa mais robusta da variância (SMYTH, 2005).

Este passo foi realizado utilizando o *script DEG.sh*, que executa três outros *scripts*, *do_comparisons.R*, *join_stats.R* e *cutDEGs.R*. A partir de um arquivo de definição das classes, das comparações a serem realizadas e de expressão gênica, o *script do_comparisons.R* calcula as estatísticas relativas a comparação como o $\log_2$ do *fold-change* e o p-valor do teste estatístico realizado pelo pacote *LIMMA* (RITCHIE et al., 2015). O *script join_stats.R* recebe a saída do *script do_comparisons.R* para múltiplos estudos que utilizam a mesma plataforma de *microarray* e retorna uma única tabela, adicionando também valores que agregam os resultados da comparação para vários estudos, como p-valor máximo e valor mínimo do $\log_2$ do *fold-change* absoluto. Por fim, o *script cutDEGs.R* associa os resultados provenientes do *script join_stats.R* obtidos para as plataformas GPL570 e GPL2700 utilizando como identificador o “*Ensembl Gene ID*”, identifica genes diferencialmente expressos, ou seja, com p-valor abaixo de um limiar e $\log_2(\text{foldchange})$ absoluto acima de um limiar, calcula a taxa de estudos no qual o gene está diferencialmente expresso e a direção em que o gene é diferencialmente expresso. Um gene será considerado hipo-expresso, caso possua maior expressão no grupo teste, ou hiper-expresso, caso tenha maior expressão no grupo controle ou discordante, caso o sinal do valor de $\log_2(\text{foldchange})$ seja diferente em todos os estudos.

3.7 Análise de Enriquecimento

Os genes diferencialmente expressos (DEG), hiper-expressos ou hipo-expressos encontrados na etapa anterior foram utilizados na análise de enriquecimento funcional. Nesta análise utilizamos bancos de dados contendo listas de genes predefinidas, que definem vias de sinalização ou metabólicas (The Gene Ontology Consortium, 2000) (BLAKE et al., 2015) (CROFT et al., 2014) e também módulos gênicos (LI et al., 2014). A significância do enriquecimento é obtida pelo teste exato de *Fisher* ou pelo teste hipergeométrico. Utilizamos o Reactome que consiste em um banco de dados de vias gênicas e que fornece detalhes a nível molecular da transdução de sinais, replicação do DNA metabolismo e outros processos celulares (CROFT et al., 2014).

Além disso, usamos para esta análise o conjunto de módulos de genes co-expressos chamado *Blood Transcriptional Modules* (BTM) (LI et al., 2014). Li et al. (2014) realizaram a integração em larga escala de redes de transcritomas de sangue públicos e bancos de dados em contextos biológicos específicos e foram capazes de deduzir um conjunto de módulos transcricionais em sangue relacionados ao estudo de vacinas.

A análise de enriquecimento é iniciada pelo *script enrichment.sh*, o qual utiliza somente os genes diferencialmente expressos de maneira concordante em pelo menos 60% dos estudos, ou seja, nas comparações definidas nos quatro estudos exigimos que o gene esteja diferencialmente expresso em pelo menos três estudos. Os genes selecionados são utilizados como entrada para o *script do_ORA.R* resultando em uma tabela contendo o nome do módulo, o tamanho do módulo, o tamanho intersecção entre a lista de genes de entrada e o módulo, o p-valor resultante do teste exato de *Fisher* e o p-valor ajustado pelo método de *Benjamini-Hochberg* (BENJAMINI; HOCHBERG, 1995). O pacote *clusterProfiler* do projeto *Bioconductor* é utilizado nesta etapa (YU et al., 2012).

Outra maneira de avaliar o enriquecimento de módulos em dados de expressão gênica é através do método conhecido como *Gene Set Enrichment Analysis* (GSEA) (SUBRAMANIAN et al., 2005). Neste método não é necessário definir os genes diferencialmente expressos previamente. Os genes são ordenados de acordo com um *score* como $\log_2$ do *fold-change*, valor de T ou p-valor. Posteriormente, verificamos se os genes membros de um determinado módulo estão distribuídos aleatoriamente na lista ordenada. Caso os genes do módulo estejam no início ou final da lista o módulo é considerado enriquecido. Uma

outra vantagem em usar este tipo de análise é que permite revelar maior similaridade entre estudos independentes, já que uma mesma via gênica alterada pode envolver genes (DE) diferentes (KHATRI; SIROTA; BUTTE, 2012) (SUBRAMANIAN et al., 2005). Igualmente a análise de enriquecimento utilizamos só o banco de dados *Blood Transcriptional Modules* e *Reactome* para realizar o GSEA.

O *script gsea.sh* utiliza os resultados gerados pela etapa de análise de expressão diferencial e extrai os valores de *Log2* do *Fold-Change* de cada comparação utilizando o *script comparisons2gseainput.R*. Caso haja mais de um valor para um mesmo gene, o maior valor é selecionado. Além disso, os identificadores dos genes são convertidos de “*ensembl_gene_id*” para “*external_gene_name*” através do pacote *biomaRt* (DURINCK et al., 2009) (DURINCK et al., 2005). A tabela gerada pelo *script comparisons2gseainput.R* contendo todas as comparações realizadas nas colunas e os genes nas linhas é utilizada como entrada pelo *script fgsea.R* que utiliza o pacote *fgsea* para realizar o enriquecimento (SERGUSHICHEV, 2016).

3.8 Análise de Correlação entre LncRNAs e mRNAs

Um lncRNA pode agir em *cis* e regular a expressão do gene (também chamado aqui de mRNA) localizado no mesmo locus (DIMITROVA et al., 2014). O aumento da expressão de um lncRNA pode levar tanto ao aumento (LI et al., 2013) quanto à diminuição da expressão do mRNA alvo (YU et al., 2008). No primeiro caso, observaria-se uma correlação positiva ou direta entre a expressão do lncRNA e o mRNA e no segundo caso, uma correlação negativa ou inversa.

Para identificar possíveis casos de regulação em *cis*, foi calculada o coeficiente de correlação de *Spearman* da expressão entre pares de lncRNAs e mRNAs de um mesmo locus. Esta etapa é realizada pelo *script corr.sh* que executa o *cor_lnc.R* para todos os estudos separadamente. Este, por sua vez, retorna uma tabela contendo as informações sobre cada par como distância, cromossomo, início e fim dos genes, valor de correlação e p-valor.

Os resultados obtidos para cada estudo são combinados utilizando o *script join_corr.R*. Este *script* obtém o maior valor de correlação para cada gene caso haja mais de uma sonda para um mesmo gene. Posteriormente, ele combina as tabelas de todos os estudos

utilizando como chave os identificadores do mRNA e do lncRNA. Uma coluna de p-valor máximo entre todos os estudos é adicionada para facilitar a busca na tabela final.

3.9 Análise de Redes de Co-expressão

A análise de co-expressão foi realizada através de uma ferramenta chamada *CEMiTool* o qual ajudei a desenvolver (RUSSO et al., 2017).

Em nosso pipeline, esta etapa é realizada pelo *script CEMiTool.sh* que realiza a execução do *script* de mesmo nome distribuído com o pacote *CEMiTool*. Os parâmetros fornecidos para o pacote *CEMiTool* são o arquivo de anotação das amostras, "*Class*" que é o nome da coluna em que as classes estão definidas neste arquivo, "*Sample_geo_accession*" que é o nome da coluna que possui o nome das amostras, "*pearson*" como tipo de correlação, o arquivo de expressão gênica e "*Symbol*" como o nome da coluna contendo o nome dos genes. O pacote *CEMiTool* seleciona somente os genes mais relevantes para análise utilizando seu próprio filtro com o valor padrão para parâmetro p-valor igual a 0.01. Além disso, o pacote seleciona um valor adequado para o parâmetro Beta do pacote *WGCNA*; caso isso não seja possível, a análise é interrompida.

O *script CEMiTool.sh* detecta quais estudos finalizaram a execução da análise com sucesso e associa os resultados da definição dos módulos presentes no arquivo "*modules.tsv*" através do *script join_cemitool.R*. Considerando que cada módulo compõe um grafo não direcionado completo, ou seja, onde todos os genes estão conectados, uma lista de adjacência é gerada para cada estudo. A lista de adjacência é combinada sendo que o número de estudos em que a aresta está presente é considerado o peso e o nome dos estudos é considerado outro atributo da aresta.

O *script get_modules_CEMiTool_joined.R* recebe a lista de adjacência definida anteriormente e utiliza os algoritmos de detecção de comunidade do pacote *igraph* (CSÁRDI; NEPUSZ, 2006) para detectar novos módulos, porém agora considerando todos os estudos. O peso de cada aresta é definido como o número de estudos em que a aresta está presente e o número total de estudos. O algoritmo de detecção de comunidades escolhido foi o Walktrap (PONS; LATAPY, 2006) devido ao tempo de execução menor comparado aos outros métodos.

A fim de entender melhor a função de cada módulo definido a análise de enriquecimento foi realizada utilizando o *script do_ORA.R*, considerando como módulos o banco de dados *Blood Transcriptional Modules*. Para definir a atividade de cada módulo em cada grupo de amostras, a análise de enriquecimento foi realizada também através do método GSEA usando o *script fgsea.R*.

Por fim, calculamos o coeficiente de correlação de *Pearson* de todos os lncRNAs com os genes definidos em cada módulo gênico utilizando o *script lnc_cor_modules.R* para cada estudo. O script salva uma matriz dos valores do coeficiente de correlação e p-valor. Além disso, é gerada uma tabela contendo os lncRNAs, módulos, informações como número de genes correlacionados, proporção de genes correlacionados dentro do módulo, correlação com a média de todos os genes de um módulo, média de correlação e identificador dos genes correlacionados. Para isso, os limiares considerados foram 0,6 para a correlação absoluta, 0,1 para o p-valor ajustado e 0,05 para o p-valor. Os resultados gerados pelo *script lnc_cor_modules.R* para cada estudo são associados utilizando o *script join_lnc_cor_modules.R*.

3.10 Análise de RNA-seq

A fim de avaliar a expressão dos genes identificados nos dados que utilizam *microarrays* em outras tecnologias, analisamos dados de RNA-seq disponíveis publicamente no banco de dados *Gene Expression Omnibus* (GEO). O estudo GSE52166 analisou 117 amostras de sangue provenientes de indivíduos saudáveis utilizando a técnica de RNA-seq. A plataforma utilizada foi a GPL11154 que emprega o sequenciador *Illumina HiSeq 2000* e leituras do tipo *paired-end* de aproximadamente 100 pares de bases.

Os dados no formato SRA foram obtidos do banco de dados *Sequence Read Archive* e convertidos para o formato FASTQ através do comando *fastq-dump*. A qualidade do sequenciamento foi avaliada através do software FastQC (ANDREWS, 2010). Os adaptadores foram removidos usando o programa Trim Galore (KRUEGER, 2012). Os arquivos gerados pelos programas FastQC e TrimGalore foram agregados em um único arquivo HTML usando o software MultiQC (EWELS et al., 2016). As seguintes amostras foram removidas por problemas técnicos: SRR1177737, SRR1026875, SRR1026876, SRR1026910,

SRR1026911, SRR1177671, SRR1177693, SRR1177713. Sendo assim no total 109 amostras foram utilizadas.

Os valores de expressão gênica foram obtidos através do software *rsem* (LI; DEWEY, 2011), pelo comando *rsem-calculate-expression* usando os seguintes parâmetros “*-time -star-gzipped-read-file -star -calc-pme -p 15*”. A montagem HG38 do genoma humano foi utilizada como referência. O arquivo em formato GTF contendo a definição dos genes utilizado foi o mesmo definido na seção de reanotação deste trabalho contendo os bancos de dados: *Gencode*, *Noncode*, *LNCipedia* e *MiTranscriptome*. Os arquivos de expressão de todas as amostras foram combinados utilizando a linguagem R e estão presentes no repositório *DengueAnalysis*. As estimativas a posteriori de transcritos por milhão (PME TPM) foram utilizados para todas as análises e gráficos deste trabalho.

3.11 Ferramentas de Análise de Transcritoma

Centenas de ferramentas foram desenvolvidas com o objetivo de manipular e processar os dados de transcritoma, assim como de realizar os mais variados tipos de análises. Diante deste grande número de opções, fica difícil para um pesquisador com pouca experiência na área realizar uma análise com segurança. Esta seção irá descrever algumas das ferramentas utilizadas em nosso *pipeline*.

GEOquery: pacote do *Bioconductor* em R (SEAN; MELTZER, 2007). Parte de um GEO ID e pode extrair os dados de expressão de um estudo do banco de dados GEO, assim como as anotações das amostras.

arrayQualityMetrics: pacote do *Bioconductor* em R (KAUFFMANN; GENTLEMAN; HUBER, 2009) utilizado para verificar a qualidade de experimentos de expressão gênica realizados em *microarray*. Este pacote gera um relatório com gráficos que identificam o perfil dos dados de expressão gênica, amostras *outliers*, efeitos de lote.

LIMMA: pacote *Linear Models for Microarray Data* do *Bioconductor* em R (RITCHIE et al., 2015) utilizado para análise de expressão gênica diferencial de dados obtidos de experimentos de *microarray*. A identificação de genes que possuem expressão diferente entre grupos experimentais pode evidenciar potenciais genes relacionados à vias à diferença genotípica e consequentemente de fenótipo entre dois grupos experimentais.

ggplot2: pacote do *The Comprehensive R Archive Network* (CRAN) em R (WICKHAM, 2009) utilizado para realização de gráficos para a visualização dos resultados e dados dos vários tipos de análises.

corrplot: pacote do repositório CRAN em R (WEI; SIMKO, 2016) utilizado para realização de gráficos específico para dados de correlação. Também utilizado para visualizar dados de enriquecimento.

WGCNA: pacote do R (LANGFELDER; HORVATH, 2008) (LANGFELDER; HORVATH, 2004) utilizado para a obtenção de grupos (módulos) de genes co-expressos a partir de dados de expressão gênica. A obtenção de módulos de co-expressão é importante para encontrar genes que possuem o mesmo comportamento de variação de expressão e consequentemente estar sofrendo a mesma regulação gênica.

4 Resultados

4.1 Reanotação das Sondas dos Microarrays

Após a reanotação das duas plataformas, identificamos as sondas que representam vários tipos (ou classes) de transcritos diferentes e também provenientes de banco de dados diferentes (figura 7). A maior parte dos transcritos de ambas plataformas foram identificadas com os dados do *Gencode*. Encontramos uma maior proporção de sondas não anotadas na plataforma GPL570 (19%) quando comparada à plataforma GPL2700 (3%).

Como esperado, a maioria das sondas representam genes codificadores de proteína (figura 7). No entanto, foi possível identificar milhares de sondas que representam lncRNAs. Interessantemente, a plataforma GPL570 possui muito mais transcritos classificados como *noncoding* do que a plataforma GPL2700 (figura 7).

Figura 7 – Classes de transcritos existentes nas plataformas GPL570 e GPL2700. A) Número de sondas anotadas de acordo com os diferentes bancos de dados. B e D) Anotação por grupos do *Gencode*. C e E) Anotação por tipo segundo o *Gencode*.

4.2 LncRNAs Diferencialmente Expressos em Infecção por Dengue

Realizamos então a análise de expressão diferencial entre pacientes com infecção aguda inicial (*EarlyDengue*) ou aguda tardia (*LateDengue*) e pessoas sem a infecção aguda (*Control*). Nossas análises revelaram milhares de transcritos cuja expressão estava alterada com a infecção nos diferentes estudos (figura 8).

A infecção aguda tardia parece causar uma perturbação maior no transcrito do que a recente. Encontramos em todos os estudos analisados, um maior número de gene diferencialmente expressos (DEGs) na comparação *LateDengue vs Control* do que na *EarlyDengue vs Control* (figura 8).

Figura 8 – *Volcano plot* das sondas diferencialmente expressos nos estudos de transcrito para as comparações dos grupos *EarlyDengue* e *LateDengue* contra o grupo *Control*. As sondas hipo-expressas estão representadas em verde e as sondas hiper-expressas estão representadas em vermelho.

A próxima análise teve como objetivo identificar quais DEGs estavam consistentemente alterados nos diferentes estudos. A figura 9 mostra que há um número considerável de DEGs em comum entre os estudos. Novamente, notamos que há mais DEGs em comum na comparação com *LateDengue* do que *EarlyDengue* (figura 9).

Figura 9 – Diagrama de *Venn* dos genes com pelo menos uma sonda considerada diferencialmente expressa independentemente do sinal do $\text{Log}_2(\text{Fold-change})$. A) Comparação *EarlyDengue vs Control*. B) Comparação *LateDengue vs Control*.

Apesar de encontrar um número total maior de DEGs consistentes entre os estudos na comparação de *LateDengue*, identificamos mais lncRNAs diferencialmente expressos na *EarlyDengue*. Enquanto 185 lncRNAs estavam diferencialmente expressos em *EarlyDengue*, apenas 19 o estavam na comparação com *LateDengue* (figura 10).

Figura 10 – Número de genes diferencialmente expressos consistentemente em pelo menos 60% dos estudos em cada grupo de transcritos.

As tabelas 1 e 2 abaixo mostram as listas completas de lncRNAs diferencialmente expressos em *EarlyDengue* e *LateDengue* comparados com Controle, respectivamente.

Tabela 1 – Lista de lncRNAs com expressão alterada na comparação *EarlyDengue vs Control*. Estão incluídos lncRNAs com expressão aumentada (*UP*) ou diminuída (*DOWN*) na infecção por dengue. A coluna “*Rank*” indica o nível de expressão do lncRNA em amostras de sangue avaliadas com o método de RNA-seq de acordo com a seção “Análise Comparativa entre microarray e RNA-seq”.

EarlyDengue vs Control

LncRNA ID	Direção	Rank	LncRNA ID	Direção	Rank	LncRNA ID	Direção	Rank	LncRNA ID	Direção	Rank	LncRNA ID	Direção	Rank
LINC00926	DOWN	6000	AC008742.1	DOWN		gene_id_G058868	DOWN		NONHSAG028467	UP	2000	gene_id_G070677	UP	
NONHSAG021967	DOWN	9000	AC009041.2	DOWN		gene_id_G060695	DOWN		PCBP1-AS1	UP	2000	GUSBP11	UP	
NONHSAG032753	DOWN	9000	AC010503.4	DOWN		gene_id_G072681	DOWN		NONHSAG016111	UP	6000	IDH1-AS1	UP	
LRRC75A-AS1	DOWN	10000	AC012146.1	DOWN		gene_id_G086287	DOWN		C1RL-AS1	UP	13000	LGALS8-AS1	UP	
PRKCQ-AS1	DOWN	10000	AC019080.1	DOWN		gene_id_G087614	DOWN		NONHSAG093642	UP	16000	LINC00487	UP	
ITGB2-AS1	DOWN	11000	AC020916.1	DOWN		gene_id_G087840	DOWN		NONHSAG032746	UP	17000	LINC00528	UP	
AC135050.6	DOWN	12000	AC021016.2	DOWN		IQCH-AS1	DOWN		JPX	UP	18000	LINC00900	UP	
MAPKAPK5-AS1	DOWN	12000	AC090204.1	DOWN		KANSL1-AS1	DOWN		NONHSAG020666	UP	18000	LINC01260	UP	
NONHSAG053410	DOWN	12000	AC093010.3	DOWN		LINC002481	DOWN		AL021707.6	UP	19000	LINC01504	UP	
AC010186.3	DOWN	13000	AC093157.1	DOWN		LINC00494	DOWN		U62317.2	UP	19000	LINC02555	UP	
CHRM3-AS2	DOWN	13000	AC104809.2	DOWN		LINC00852	DOWN		AL157394.1	UP	20000	MGC16275	UP	
TSPOAP1-AS1	DOWN	13000	AC107952.2	DOWN		LINC01215	DOWN		NONHSAG081930	UP	20000	MIR34AHG	UP	
AC026401.3	DOWN	14000	AC125257.1	DOWN		LINC01857	DOWN		SP2-AS1	UP	20000	MIR3945HG	UP	
LINC01588	DOWN	14000	AF131215.5	DOWN		LINC01963	DOWN		AC004771.1	UP		NONHSAG004922	UP	
SNHG8	DOWN	14000	AF131215.6	DOWN		LINC02035	DOWN		AC007608.3	UP		NONHSAG006481	UP	
AC093323.1	DOWN	15000	AL121839.2	DOWN		LOXL1-AS1	DOWN		AC009948.1	UP		NONHSAG006898	UP	
AC100810.1	DOWN	15000	AL133355.1	DOWN		MINCR	DOWN		AC083862.2	UP		NONHSAG013184	UP	
C1orf132	DOWN	16000	AL137003.2	DOWN		MIR3142HG	DOWN		AC087672.3	UP		NONHSAG021744	UP	
EBLN3P	DOWN	16000	AL355488.2	DOWN		NONHSAG000447	DOWN		AC092718.4	UP		NONHSAG029691	UP	
FGD5-AS1	DOWN	16000	AL513477.2	DOWN		NONHSAG007679	DOWN		AC093278.2	UP		NONHSAG035025	UP	
LINC01550	DOWN	16000	AL591895.1	DOWN		NONHSAG011964	DOWN		AC099343.3	UP		NONHSAG039503	UP	
TTC28-AS1	DOWN	16000	AL691432.2	DOWN		NONHSAG020736	DOWN		AC147651.3	UP		NONHSAG039504	UP	
AC005674.2	DOWN	17000	AP002478.1	DOWN		NONHSAG025645	DOWN		AL137145.1	UP		NONHSAG051497	UP	
OTUD6B-AS1	DOWN	17000	AP002840.2	DOWN		NONHSAG030295	DOWN		AL356356.1	UP		NONHSAG053607	UP	
AC055822.1	DOWN	18000	AP002884.3	DOWN		NONHSAG046805	DOWN		AL359532.1	UP		NONHSAG073818	UP	
AC096921.2	DOWN	18000	CERNA1	DOWN		NONHSAG046849	DOWN		AP000757.1	UP		NONHSAG094043	UP	
LINC01089	DOWN	18000	DANCR	DOWN		NONHSAG053848	DOWN		AP003774.1	UP		NONHSAG097384	UP	
NONHSAG025200	DOWN	18000	ENSG00000188206	DOWN		NONHSAG076336	DOWN		ATP2A1-AS1	UP		PITRM1-AS1	UP	
LEF1-AS1	DOWN	19000	ENSG00000270189	DOWN		PITPNA-AS1	DOWN		FAM225A	UP		PYCARD-AS1	UP	
NONHSAG000329	DOWN	19000	gene_id_G002292	DOWN		PRKAG2-AS1	DOWN		FOXN3-AS1	UP		SCAMP1-AS1	UP	
AC024075.2	DOWN	20000	gene_id_G012662	DOWN		PWAR6	DOWN		gene_id_G002122	UP		SERPINB9P1	UP	
AL138756.1	DOWN	20000	gene_id_G035489	DOWN		SLC25A25-AS1	DOWN		gene_id_G012396	UP		USP30-AS1	UP	
LINC01278	DOWN	20000	gene_id_G035923	DOWN		SMIM27	DOWN		gene_id_G036337	UP				
NONHSAG030333	DOWN	20000	gene_id_G040474	DOWN		SNHG19	DOWN		gene_id_G043680	UP				
AC004656.1	DOWN		gene_id_G044490	DOWN		SNHG9	DOWN		gene_id_G063244	UP				
AC004918.1	DOWN		gene_id_G054037	DOWN		ST20-AS1	DOWN		gene_id_G063246	UP				
AC005842.1	DOWN		gene_id_G054809	DOWN		TRIM52-AS1	DOWN		gene_id_G063810	UP				
AC008124.1	DOWN		gene_id_G057040	DOWN		ZBED5-AS1	DOWN		gene_id_G070068	UP				
						ZNF793-AS1	DOWN							

Tabela 2 – Lista de lncRNAs com expressão alterada na comparação *LateDengue vs Control*. Estão incluídos lncRNAs com expressão aumentada (*UP*) ou diminuída (*DOWN*) na infecção por dengue. A coluna “*RNA-seq Rank*” indica o nível de expressão do lncRNA em amostras de sangue avaliadas com o método de RNA-Seq de acordo com a seção “Análise Comparativa entre microarray e RNA-seq”

LateDengue vs Control

LncRNA ID	Direção	RNA-seq Rank	LncRNA ID	Direção	RNA-seq Rank
LINC00926	DOWN	6000	FAM30A	UP	13000
NONHSAG043790	DOWN	7000	NONHSAG093072	UP	15000
AC093323.1	DOWN	15000	Z93930.2	UP	16000
LINC00877	DOWN	15000	ARRDC1-AS1	UP	
MCM3AP-AS1	DOWN	19000	DANCR	UP	
gene_id_G078254	DOWN		LINC01679	UP	
ITPR1-AS1	DOWN		MIR503HG	UP	
LINC00944	DOWN				
LINC00954	DOWN				
MIR100HG	DOWN				
NONHSAG001220	DOWN				
NONHSAG048350	DOWN				

4.3 Análise Comparativa entre Microarray e RNA-seq

Para confirmar que os lncRNAs são expressos em sangue, nós re-analisamos dados de RNA-seq de sangue disponíveis no GEO e obtivemos a expressão dos genes em 109 amostras saudáveis. O valor de TPM (*Transcripts Per Million*) foi utilizado para representar a expressão de um gene. A média de TPM entre todas as amostras para cada gene foi calculada, os genes contendo sondas confiáveis nas plataformas de *microarray* foram obtidos e foi observado que a expressão do grupo *lncoding*, que representa os lncRNAs, possui expressão menor que do grupo *coding*, que representa os genes codificadores de proteínas (figura 11 A). Além disso, o valor de expressão de todos os genes em cada amostra foi ordenado em ordem decrescente e grupos contendo os 1000, 2000, 3000 até 20000 genes com maior expressão foram criados. Então, os genes foram classificados como *Top_1000* se ele estivesse no grupo 1000 em pelo menos 90% das amostras, o mesmo foi realizado para os outros grupos (figura 11 B e 11 C). Desta maneira podemos classificar os lncRNAs de acordo com o seu nível de expressão em RNA-seq.

Figura 11 – Comparação entre transcritômica por *microarray* e RNAseq A) Densidade da média de expressão pela média de TPM dos genes de transcritos codificantes e de lncRNAs. B) Número de genes da plataforma de *microarray* com maior expressão nos dados de RNAseq. Os genes foram ordenados de acordo com o valor de TPM para cada amostra e se um gene pertencer aos 1000 genes com maior expressão em pelo menos 90 por cento das amostras este gene é considerado pertencente ao grupo *Top_1000*, este mesmo procedimento foi repetido de 1000 até 20000 genes. C) Similar ao item B, porém representando o número de lncRNAs

4.4 Análise de Enriquecimento

Com o objetivo de entendermos melhor a função dos genes consistentemente diferencialmente expressos nas comparações anteriores, foi realizada uma análise de enriquecimento funcional utilizando o banco de dados *Blood Transcriptional Modules*. Os módulos foram ordenados por significância estatística ($-\text{Log}_{10}(\text{p-valor ajustado})$) e estão representados na figura 12. Na comparação *EarlyDengue vs Control* os módulos mais significativos com menor expressão em pacientes com dengue estão relacionados à células T, ativação de células T, diferenciação de células T, células B. Enquanto que os genes com maior expressão em pacientes com dengue estão relacionados com ativação de células dendríticas, monócitos, resposta à *interferon* do tipo 1, resposta inata antiviral e divisão celular estimulada em células T CD4.

Na comparação *LateDengue vs Control* os genes com menor expressão estão relacionados à células dendríticas, monócitos, neutrófilos, sinalização inflamatória e *Toll Like Receptors* (TLR) e ativação de células B. Já os genes com maior expressão em pacientes

infectados os módulos mais relevantes estão relacionados à ciclo celular, divisão celular estimulada em células T CD4+, eventos de sinalização PLK1, alvos de E2F1.

Figura 12 – Análise de enriquecimento das comparações com *Blood Transcriptional Modules*. Em cada gráfico de barras está presente os dez módulos com maior significância estatística. Comparação *EarlyDengue vs Control* (A e B) e Comparação *LateDengue vs Control* (C e D). Em verde são vias enriquecidas dos genes com expressão diminuída na infecção (A e C) e em vermelho são vias enriquecidas dos genes com expressão aumentada.

A análise de enriquecimento foi realizada também utilizando o banco de dados Reactome (figura 13). Na comparação *EarlyDengue vs Control* nenhuma via foi enriquecida para os genes com expressão diminuída. Para os genes com expressão aumentada nesta comparação identificamos como significativas as vias relacionadas ao sistema imunológico, sistema imunológico inato, ciclo celular, tradução mitocondrial, fase M do ciclo celular, degranulação de neutrófilos, sinalização por citocinas no sistema imune e sinalização por *interferon* gama.

Em contraste, na comparação *LateDengue vs Control* foi possível obter vias enriquecidas para os genes com expressão diminuída, sendo eles, sistema imunológico, sistema imunológico inato, degranulação de neutrófilos, sinalização por citocinas no sistema

imunológico, sinalização por interleucinas, metabolismo de fosfolípidos, biossíntese de glicerofosfolípidos, sinalização por interleucinas 3, 5 e GM-CSF, sinalização *DC-SIGN* e interações de superfície celular na parede vascular. Já os genes com expressão aumentada possuem as seguintes vias significativas relacionadas ao ciclo celular e mitose.

Figura 13 – Análise de enriquecimento das comparações com *Reactome Pathways*. Em cada gráfico de barras está presente os dez módulos com maior significância estatística. Comparação EarlyDengue vs Control (A e B) e Comparação LateDengue vs Control (C e D). Em verde são vias enriquecidas dos genes hipo-expressos (A e C) e em vermelho são vias enriquecidas dos genes sobre expressos (B e D).

4.5 Gene Set Enrichment Analysis

Também foi realizada o *Gene Set Enrichment Analysis* (GSEA) para verificar módulos do banco de dados *Blood Transcriptional Modules* alterados nas comparações realizadas.

Na comparação *EarlyDengue vs Control* (figura 14) os módulos enriquecidos positivamente são ciclo celular e transcrição, células dendríticas ativadas, superfície de células dendríticas, superfície de células dendríticas ativadas por LPS, monócitos, sinalização inflamatória e *Toll-like receptors*, detecção e imunidade viral, rede de alvos de IRF2, grupo de quimiocinas, assinatura de *interferon* antiviral e ativação imunológica. Já os módulos

enriquecidos negativamente são módulos de células B e células secretoras de anticorpo, imunoglobulinas, superfície de células B, células B, ativação de células T e células T .

Na comparação *Late Dengue vs Control* (figura 15) os módulos enriquecidos positivamente são todos relacionados ao ciclo celular e reparo de DNA. Já os módulos enriquecidos negativamente são módulos de ativação dos sistema imunológico, proteínas de membrana, coagulação sanguínea, transporte transmembranar, sinalização inflamatória e *Toll-like receptors*, superfície de monócitos, monócitos e neutrófilos.

Figura 14 – *Gene Set Enrichment Analysis* para comparação entre *EarlyDengue* e *Control* com o banco de dados *Blood Transcriptional Modules*. O tamanho de cada círculo representa o valor absoluto de *Normalized Enrichment Score*.

Figura 15 – *Gene Set Enrichment Analysis* para comparação *LateDengue vs Control* com o banco de dados *Blood Transcriptional Modules*. O tamanho de cada círculo representa o valor absoluto de *Normalized Enrichment Score*.

4.6 Correlação em Cis

A fim de identificar possíveis regulações da expressão dos genes pelos lncRNAs em cis foi calculado o coeficiente de correlação de Spearman dos pares de lncRNAs e mRNAs próximos. Foram identificados 10 pares consistentemente correlacionados nos quatro estudos (figura 16).

Figura 16 – Correlação dos lncRNAs com os respectivos genes que podem estar sendo regulados em cis. As letras ao lado dos genes representam se o gene está diferencialmente expresso na comparação *EarlyDengue vs Control*, representado pela letra I (Agudo Inicial), ou na comparação *LateDengue vs Control*, representado pela letra T (Agudo Tardio). A cor vermelha representa que o gene está com maior expressão na classe Teste e azul representa maior expressão na classe Controle. As caixas em cor verde com o texto “xK”, onde x pode valer de 1 à 20, representam a expressão do gene em amostras de RNA-seq de acordo com a figura 11-B.

Um exemplo interessante é do par formado pelo gene EBNA1BP2 e lncRNA NONHSAG001220 que possui correlação negativa nos quatro estudos (figura 17). Além disso, o gene EBNA1BP2 é diferencialmente expresso com expressão aumentada na classe Agudo Tardio.

Figura 17 – Correlação entre EBNA1BP2 e o lncRNA NONHSAG001220 em quatro estudos de microarray. A linha azul representa a reta proveniente da regressão linear. Pontos vermelhos representam amostras de pacientes com dengue e pontos azuis as amostras de pessoas não infectadas com dengue.

4.7 Análise de Redes de Co-expressão

Esta análise de redes de co-expressão nos permite definir quais genes possuem um mesmo padrão de expressão em todas as amostras e identificar quais módulos estão enriquecidos nas comparações realizadas na análise de expressão diferencial.

Para definir nossos módulos de co-expressão utilizamos a ferramenta *CEMiTool* nos quatro estudos. Para o estudo GSE28405 não foi possível obter o valor de Beta e portanto a análise não foi finalizada. Os módulos obtidos nos três estudos foram associados e o método GSEA foi utilizado para identificar o enriquecimento dos módulos nas comparações realizadas na análise de expressão diferencial. Na comparação *EarlyDengue vs Control* os módulos enriquecidos positivamente nos três estudos são Mod3, Mod8, Mod9 e Mod10 e os módulos enriquecidos negativamente são Mod12 e Mod6 (figura 18). Na comparação

LateDengue vs Control os módulos enriquecidos positivamente nos quatro estudos são Mod3 e Mod4. Enquanto que os módulos enriquecidos negativamente são Mod6 e Mod5. É interessante notar que o módulo Mod5 está enriquecido negativamente na comparação *LateDengue vs Control* enquanto que na comparação *EarlyDengue vs Control* este módulos está enriquecimento positivamente em dois dos três estudos.

Figura 18 – Análise de enriquecimento dos módulos de coexpressão. O tamanho de cada círculo representa o valor absoluto de *Normalized Enrichment Score* (NES). A) Comparação *EarlyDengue vs. Control*. B) Comparação entre *LateDengue vs Control*.

A fim de entender melhor a função de cada módulo identificado nós realizamos a análise de enriquecimento (figura 19). Os resultados para os módulos que estavam enriquecidos na comparação de classes estão presentes na figura 19. O módulo Mod3

está relacionado com células dendríticas ativadas, resposta inata antiviral, assinatura de *interferon* antiviral, resposta a *interferon* do tipo 1, receptor RIG-1 *like*. O módulo Mod4 não possui enriquecimento de nenhum módulo. O módulo Mod5 possui enriquecidos módulos relacionados à neutrófilos, ativação imunológica, receptores *Toll-like* e sinalização inflamatória, monócitos, superfície de monócitos, receptores acoplados à proteína G e recrutamento de neutrófilos. O módulo Mod6 está relacionado à células B, desenvolvimento de células T e B, imunoglobulinas, células plasmáticas e células B, diferenciação de células T, redes de TCR, IL2 e IL7 e ativação de células T. O módulo Mod8 está relacionado à monócitos, transcrição e ciclo celular, transportadores e receptores em células mielóides, assinatura de superfície de monócitos, receptores *Toll-like* e sinalização inflamatória, ativação imunológica e resposta inflamatória. O módulo Mod9 possui os seguintes módulos enriquecidos: resposta a *interferon* do tipo 1, monócitos, *interferon* antiviral, ativação do sistema complemento, células dendríticas ativadas, receptor RIG-1 *like*, redes alvo de IRF2, ativação da imunidade inata através da percepção de DNA citosólico, células dendríticas, transportadores e receptores enriquecidos em células mielóides. O módulo Mod10 possui enriquecido o módulo de apresentação de antígenos. O módulo Mod12 possui enriquecido os módulos de células T e ativação de células T.

Figura 19 – Análise de enriquecimento dos módulos de co-expressão. Os módulos com enriquecimento consistente entre todos os estudos através da análise de GSEA foram selecionados..

Para identificar possíveis regulações dos lncRNAs nos módulos encontrados nós calculamos a correlação dos lncRNAs com a média da expressão dos componentes de cada módulo bem como a correlação de cada componente do módulo com o todos os lncRNAs . Os resultados dessa análise estão resumidos na figura 20. O módulo com mais lncRNAs correlacionados é o Mod1 contendo 6 lncRNAs correlacionados. O lncRNA correlacionado com mais módulos é o DANCR, totalizando cinco módulos. Está correlacionado positivamente com o expressão média do módulo Mod1 e possui mais genes correlacionados

com este módulo. O lncRNA DANCN está hipo-expresso na comparação *EarlyDengue vs Control* e hiper-expresso na comparação *LateDengue vs Control*. O lncRNA FAM30A está hiper-expresso na comparação *LateDengue vs Control* e está correlacionado positivamente com o Mod1 e Mod4 e negativamente com o Mod10.

Figura 20 – Correlação entre os lncRNAs e módulos de co-expressão. A) Correlação entre a média de cada módulo de co-expressão e lncRNA. A caixa em verde representa a expressão do lncRNA na análise de RNA-seq. A caixa com as letras I ou T representam que o lncRNA está diferencialmente expresso na comparação *EarlyDengue vs Control* ou *LateDengue vs Control*, respectivamente. A cor vermelha representa que o lncRNA é hiper-expresso e em verde representa que o lncRNA é hipo-expresso B) Número de genes correlacionados com o lncRNA em cada módulo de co-expressão.

5 Discussão

Apesar de vários estudos de transcrito em pacientes com dengue terem sido publicados, nenhum deles teve como foco a análise dos lncRNAs. Em nossa meta-análise, encontramos cerca de 200 lncRNAs possivelmente relacionados com a infecção por dengue.

A análise de expressão diferencial por estudo mostrou que em ambas as comparações (*EarlyDengue vs Control* e *LateDengue vs Control*), havia um número muito grande de genes diferencialmente expressos, com uma média de 4.173,5 DEGs/estudo. É interessante notar que o maior número de genes diferencialmente expressos ocorre na plataforma GPL570 podendo indicar algum efeito técnico. Isto reforça a necessidade de associar os resultados de diferentes estudos com o objetivo de mitigar estes efeitos. Sendo assim, consideramos somente como diferencialmente expressos aqueles genes que eram DEGs em pelo menos 60% dos estudos. A comparação *LateDengue vs Control* apresentou um número ligeiramente maior de genes diferencialmente expressos, 2.991 DEGs, em comparação à *EarlyDengue vs Control*, 2.856, após a combinação dos quatro estudos. Entretanto, o maior número de lncRNAs diferencialmente expressos, 185, ocorreu na comparação *EarlyDengue vs Control* contra 19 lncRNAs na comparação *LateDengue vs Control*. Isto pode indicar a maior importância dos lncRNAs na imunidade inata induzida pela infecção pelo vírus da dengue.

A expressão do lncRNA LRRC75A-AS1 está diminuída na comparação *EarlyDengue vs Control*, este gene está localizado no braço curto do cromossomo 17 próximo ao gene LRRC75A porém na fita oposta. Both et al. (2012) identificaram o LRRC75A-AS1 como um possível oncogene com papel importante em osteossarcomas, este gene possui maior expressão em tumores e a alteração na sua expressão está associada à alterações de número de cópias. Zhang et al. (2016) analisaram a expressão gênica em sinoviócitos de pacientes com artrite reumatóide e identificaram centenas de lncRNAs com expressão alterada entre eles o LRRC75A-AS1 que está hiper-expresso podendo indicar a relação deste gene com o processo patológico da artrite reumatóide. Apesar destes estudos a função biológica do LRRC75A-AS1 ainda não foi definida.

Já a expressão do lncRNA PCBP1-AS1 está aumentada tanto na comparação *EarlyDengue vs Control* quanto em carcinoma de células escamosas bucal (FENG et al., 2017). Mo et al. (2017) realizaram uma análise de redes de co-expressão em células

mononucleares do sangue periférico e identificaram que o lncRNA PCBP1-AS1 é um *hub* em um dos módulos identificados e está relacionado às vias do banco de dados KEGG de Lisossomos, receptores RIG-1 like, via de sinalização do receptor *NOD-like* e via de sinalização mTOR.

Na comparação *LateDengue vs Control* o lncRNA diferencialmente expresso com maior número de artigos o citando é o DANCR, o qual está com maior expressão na classe *LateDengue*. Alguns estudos indicam que o DANCR possui papel importante papel importante em diversos tipos de tumores (MAO et al., 2017) (JIANG et al., 2017) (SHA et al., 2017). Mao et al. (2017) mostraram que o DANCR está com maior expressão em células de câncer gástrico e a associação de DANCR com EZH2 e HDAC3 leva ao silenciamento do gene lncRNA-LET levando migração celular e invasão.

O fato de lncRNAs induzidos pela infecção também terem sua expressão aumentada em tumores pode estar relacionado à proliferação de células do sistema imune que são ativadas pela infecção por Dengue. Estudos já mostraram que na infecção por Dengue há uma expansão da população de monócitos (KWISSA et al., 2014b) e de células B secretoras de anticorpos (plasmablastos) (GARCIA-BATES et al., 2013). É possível que os lncRNAs possam possuir um papel nesta expansão de tipos celulares específicos.

A análise de enriquecimento realizada utilizando os DEGs nos permitiu o melhor entendimento dos módulos e vias sendo afetados durante a infecção. Na comparação *EarlyDengue vs Control* os genes com expressão aumentada na infecção estão relacionados com ciclo celular o que já foi estabelecido por Simmons et al. (2007), que realizaram um dos primeiros estudos do transcrito de pacientes infectados pelo vírus da dengue, indicando que a divisão celular e processos relacionados estão envolvidos na resposta inicial a dengue. Além disso, o módulo de eventos de sinalização por PLK1 está enriquecido positivamente (isto é, atividade aumentada na infecção). Este módulo possui papel já conhecido no controle do ciclo celular. Pohl et al. (2017) mostraram que o *knockdown* dos genes PLK1, PLK2 e PLK3 ou a inativação da função de quinase PLK leva a redução da replicação do vírus influenza. Além disso este estudo mostrou que a atividade da PLK é necessária em estágios iniciais do ciclo de replicação viral. Sendo assim, as *Polo-like kinases* (PLK) estão sendo consideradas como potenciais alvos terapêuticos para a infecção pelo vírus influenza. Outro módulo enriquecido para os genes hiper-expressos em *EarlyDengue vs Control* são módulos de células dendríticas ativadas e monócitos, essas células são alvos iniciais do vírus da dengue (COSTA et al., 2013). Os módulos de resposta à *interferon* do tipo 1

e assinatura de *interferon* antiviral estão enriquecidos, o que já foi descrito em diversos artigos que avaliaram o perfil de expressão gênica em pacientes com dengue (SIMMONS et al., 2007), (SUN et al., 2013), (AFROZ et al., 2016). Diversas moléculas responsáveis por identificar partículas virais são ativadas e levam a ativação de duas cascatas de regulação, sendo elas fatores regulatórios de *interferon* e NF-KB, que levam a produção de *interferon* do tipo 1 e citocinas inflamatórias (COSTA et al., 2013).

Na análise de enriquecimento realizada com os genes hipo-expressos na comparação *EarlyDengue vs Control*, ou seja, que possuem expressão diminuída na infecção, estão muitos módulos relacionados a células T e células B, o que pode sugerir a diminuição dessas células no sangue em estágios iniciais da infecção ou sua migração para órgãos linfoides saindo portanto da circulação periférica.

O módulo de degranulação de neutrófilos está enriquecido entre os genes hiper-expressos na comparação *EarlyDengue vs Control*, isto também foi identificado no estudo de Banerjee et al. (2017). Os autores deste estudo sugerem que o processo de degranulação e ativação de neutrófilos está relacionado com a severidade da dengue.

Na comparação *LateDengue vs Control* grande parte dos módulos enriquecidos positivamente estão relacionados com ciclo celular, igualmente visto na comparação *EarlyDengue vs Control*. Em contraste com a comparação *EarlyDengue vs Control*, entre os módulos enriquecidos negativamente na comparação *LateDengue vs Control* estão células dendríticas, monócitos, neutrófilos, receptores *Toll-like* e sinalização inflamatória. Isto sugere que vias relacionadas com imunidade inata e adaptativa aparecem enriquecidas de forma diferente em dengue aguda inicial e tardia. Como esperado, módulos relacionados a imunidade inata estão positivamente enriquecidos nos estágios iniciais da infecção e negativamente enriquecidos em estágios tardios. Enquanto que módulos relacionados a imunidade adaptativa estão enriquecidos positivamente.

As análises de GSEA corroboram os resultados obtidos pelas análises de enriquecimentos citadas anteriormente. Novamente percebemos que módulos relacionados a imunidade inata estão enriquecidos positivamente no estágio inicial da infecção, como *interferon* do tipo 1 e sinalização inflamatória e módulos relacionados a imunidade adaptativa, como ciclo celular estimulado em células T CD4+, estão enriquecidos positivamente no estágio tardio. Uma diferença importante entre as duas análises é que a análise de enriquecimento parte de uma lista de genes diferencialmente expressos na qual temos que definir através de limiares de p-valor e *fold-change*. Na análise de GSEA, interrogamos todos

os genes de acordo com sua posição no *rank* ordenado pelo $\log_2(\text{fold-change})$, por exemplo (SUBRAMANIAN et al., 2005). Esses resultados similares entre as duas análises nos dá mais confiança de que os limiares que utilizamos para definir os genes diferencialmente expressos não afetaram os resultados de enriquecimento de vias de sinalização.

Através das análises de correlação em cis foi possível revelar potenciais lncRNAs reguladores de genes próximos a eles. Podemos citar o exemplo do par de lncRNA NONHSAG001220 e o gene EBNA1BP2 (*EBNA1 binding protein*). Este par possui correlação negativa consistente em quatro estudos sendo que o lncRNA está hipo-expresso e o gene EBNA1BP2 está hiper-expresso em amostras com infecção tardia. O gene EBNA1BP2 é conhecido por se associar ao gene EBNA1 (*Epstein Barr nuclear antigen 1*), este gene é proveniente do vírus *Epstein Barr* e possui papel fundamental na replicação do genoma viral (KAPOOR; FRAPPIER, 2003). O gene EBNA1BP2 também se ligar ao proto-oncogene c-Myc podendo regular a proliferação celular e a tumorigênese (KAPOOR; FRAPPIER, 2003). Tanto o c-Myc quanto o EBNA1 são conhecidos também por suas funções no aumento da proliferação de células B (De Alboran et al., 2001). Já foi mostrado que lncRNAs poderiam inibir a transcrição de genes próximos por mecanismos epigenéticos (O'LEARY et al., 2015), através da associação com outras proteínas como a RNA Polimerase II (MARINER et al., 2008) ou degradação da molécula de RNA (GONG; MAQUAT, 2011). Sendo assim, o lncRNA NONHSAG001220 pode ser um potencial regulador da expressão do gene EBNA1BP2, inibindo sua expressão. Entretanto, experimentos ainda precisam ser realizados para confirmar esta possível regulação.

Para identificar possíveis regulações em trans de lncRNAs, executamos primeiro uma análise de redes de co-expressão com a ferramenta desenvolvida por nosso laboratório chamada CEMiTool (manuscrito submetido) (RUSSO et al., 2017). A ferramenta está publicamente disponível no repositório do projeto Bioconductor e no repositório do GitHub¹. O pacote CEMiTool é baseado na ferramenta WGCNA (LANGFELDER; HORVATH, 2008), o qual possui como parâmetro o valor de Beta utilizado para obter o peso de cada aresta de uma rede de co-expressão. O pacote CEMiTool é capaz de detectar o valor de Beta automaticamente quando possível, caso contrário, a análise não é executada. Sendo assim, foi possível realizar a análise para três dos quatro estudos. Não foi possível obter o valor de Beta para o estudo GSE28405, isto poderia ser devido ao menor número de amostras utilizados, porém este estudo contém cerca de 80 amostras e é o segundo

¹ <https://github.com/csbl-usp/CEMiTool/>

estudo com maior número de amostras. Outra possível explicação para isso pode ser alta variabilidade entre as amostras devido a efeitos biológicos ou problemas técnicos como efeito de lote ².

O CEMiTool, assim como o WGCNA, software no qual ele foi baseado (LANGFELDER; HORVATH, 2008), após a obtenção dos módulos resulta em um grafo completo para cada módulo, ou seja, cada gene é considerado co-expresso com todos os genes do seu módulo. Assim, quando executamos a ferramenta em três estudos, cada análise gerou um conjunto diferente de módulos. A fim de obter módulos consenso entre todos os estudos, um grafo contendo a união das arestas e vértices presentes nos três estudos foi criado e foi atribuído ao peso de cada aresta o número de estudos em que a aresta estava presente. Posteriormente, o algoritmo de detecção de comunidades em grafos, Walktrap (PONS; LATAPY, 2006), foi aplicado resultando em grupos de vértices que chamamos de módulo consenso. Outra alternativa a abordagem utilizada que consideramos foi utilizar apenas as arestas que estavam presentes nos três estudos e então obter os componentes conectados do grafo. No entanto, esta análise (não mostrada no texto) gerou poucos módulos grandes e muitos módulos contendo apenas dois ou três genes. Por isso, resolvemos adotar a estratégia descrita nos métodos.

A abordagem de *Gene Set Enrichment Analysis* utilizando módulos consenso revelou módulos com atividade induzida ou reprimida nos grupos de amostras de dengue aguda inicial e tardia em relação a amostras controle. Muitos dos módulos estavam aumentados ou diminuídos em ambas as condições. No entanto, o módulo 5 estava com a atividade aumentada na dengue aguda inicial mas diminuída na aguda tardia. Através da análise de enriquecimento foi verificado que o módulo 5 contém genes relacionados à neutrófilos, monócitos, receptores *Toll-like* e sinalização inflamatória e receptores acoplados à proteína G. Esse resultado é corroborado pelos achados de Banerjee et al. (2017) que encontraram genes hiper-expressos relacionados a neutrófilos e demonstraram uma possível associação da imunidade mediada por neutrófilos e a severidade da dengue.

Cabe ressaltar que vários dos módulos consensos estavam enriquecidos por vias relacionadas com imunidade inata e adaptativa. O módulo 12, relacionado à células T e ativação de células T, está enriquecido negativamente na classe Aguda Inicial em comparação à classe controle. O módulo 9, relacionado à *interferon* do tipo 1, monócitos, ativação do sistema complemento e células dendríticas ativadas, está enriquecido positiva-

² <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html>

mente na comparação na classe Agudo Inicial em relação a classe controle. Isto é esperado, já que, monócitos e células dendríticas são alvos iniciais da infecção pelo vírus da dengue. Além disso, *interferon* do tipo 1 e suas vias relacionadas são conhecidos por participar nas primeiras respostas à infecções virais (KWISSA et al., 2014b) (RODENHUIS-ZYBERT; WILSCHUT; SMIT, 2010).

Comparamos então o perfil de expressão dos lncRNAs com os genes dos módulos para achar possíveis regulações em trans. Nossas análises revelaram candidatos interessantes que parecem regular ou serem regulados da mesma forma que os genes dos módulos identificados. O lncRNA correlacionado com mais módulos é o DANCR, totalizando cinco módulos. DANCR está correlacionado negativamente com os módulos 5, 9 e 10, que estão enriquecidos positivamente na classe Agudo Inicial. Além disso o lncRNA DANCR está hiper-expresso na classe Agudo Tardia e hipo-expresso na classe Agudo Inicial. O módulo 9 contém genes relacionados a *interferon* do tipo 1, monócitos, ativação do sistema complemento, células dendríticas ativadas. Já, o módulo 10 está relacionado a apresentação de antígenos. Sendo assim, a baixa expressão do lncRNA DANCR na classe Agudo Inicial pode ser responsável por pela ativação dos genes dos módulos 5, 9, 10.

6 Conclusão

Neste trabalho nós mostramos uma análise sistemática de lncRNAs possivelmente envolvidos com a infecção por dengue. Nossos achados revelaram que vários lncRNAs possuem a expressão alterada em sangue de pacientes com dengue, tanto na fase aguda inicial quanto na tardia. Mostramos evidência que alguns dos lncRNAs podem agir em cis, regulando a expressão do gene codificador próximo a eles. Esse foi o caso do lncRNA NONHSAG001220 correlacionado com o gene EBNA1BP2. Para alguns lncRNAs, procuramos uma possível ação em trans, realizando análises de co-expressão. Encontramos lncRNAs relacionados a resposta de *interferon* tipo 1, sinalização inflamatória e receptores *Toll-like* e células dendríticas ativadas. Isso indica que eles podem estar regulando genes dessas vias ou estão sendo regulados de uma forma similar pelos mesmos fatores que regulam essas vias. Mesmo que novos experimentos sejam necessários para validar esses achados, nós mostramos evidências de que lncRNAs possam ter um papel importante na regulação gênica durante a infecção por dengue. Por fim, acreditamos que este trabalho serviu para revelar novas intuições sobre os mecanismos regulatórios da infecção por dengue.

Referências

- AFROZ, S. et al. Transcriptome meta-analysis reveals a dysregulation in extra cellular matrix and cell junction associated gene signatures during Dengue virus infection. *Scientific reports*, Nature Publishing Group, v. 6, n. January, p. 33752, sep 2016. ISSN 2045-2322. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/27651116http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5030657>. Citado 2 vezes nas páginas 25 e 66.
- ALLONSO, D. et al. Dengue Virus NS1 Protein Modulates Cellular Energy Metabolism by Increasing Glyceraldehyde-3-Phosphate Dehydrogenase Activity. *Journal of Virology*, v. 89, n. 23, p. 11871–11883, 2015. ISSN 0022-538X. Disponível em: <http://jvi.asm.org/lookup/doi/10.1128/JVI.01342-15>. Citado na página 22.
- ANDREWS, S. *FastQC: a quality control tool for high throughput sequence data*. 2010. Disponível em: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Citado na página 45.
- AUNE, T. M.; SPURLOCK, C. F. Long non-coding RNAs in innate and adaptive immunity. *Virus Research*, Elsevier B.V., v. 212, p. 146–160, 2016. ISSN 18727492. Disponível em: <http://dx.doi.org/10.1016/j.virusres.2015.07.003>. Citado na página 30.
- AUNE, T. M.; Spurlock Iii, C. F. Long non-coding RNAs in innate and adaptive immunity. *Virus Research*, Elsevier B.V., v. 212, p. 146–160, 2016. Citado na página 29.
- BALL, C. A. et al. Submission of microarray data to public repositories. *PLoS Biology*, v. 2, n. 9, p. 1276–1277, 2004. ISSN 15449173. Citado na página 32.
- BANERJEE, A. et al. RNA-Seq analysis of peripheral blood mononuclear cells reveals unique transcriptional signatures associated with disease progression in dengue patients. *Translational Research*, Elsevier, v. 186, p. 62–78.e9, 2017. ISSN 18781810. Disponível em: <http://dx.doi.org/10.1016/j.trsl.2017.06.007>. Citado 3 vezes nas páginas 25, 66 e 68.
- BAYTAK, E. et al. Whole transcriptome analysis reveals dysregulated oncogenic lncRNAs in natural killer/T-cell lymphoma and establishes MIR155HG as a target of PRDM1. *Tumor Biology*, v. 39, n. 5, p. 101042831770164, 2017. ISSN 1010-4283. Disponível em: <http://journals.sagepub.com/doi/10.1177/1010428317701648>. Citado na página 31.
- BENJAMINI, Y.; HOCHBERG, Y. *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. 1995. 289–300 p. Citado na página 42.
- BHATT, S. et al. The global distribution and burden of dengue. *Nature*, Nature Publishing Group, v. 496, n. 7446, p. 504–7, apr 2013. ISSN 1476-4687. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3651993{&}tool=pmcentrez{&}rendertype=ab>. Citado 3 vezes nas páginas 11, 19 e 21.
- BLAKE, J. a. et al. Gene ontology consortium: Going forward. *Nucleic Acids Research*, v. 43, n. D1, p. D1049–D1056, 2015. ISSN 13624962. Citado na página 42.
- BOLSTAD, B. M. et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, v. 19, n. 2, p. 185–193, 2003. ISSN 1367-4803. Disponível em: <http://bioinformatics.oxfordjournals.org/content/19/2/185.long>. Citado na página 40.

BOTH, J. et al. Identification of novel candidate oncogenes in chromosome region 17p11.2-p12 in human osteosarcoma. *PLoS One*, v. 7, n. 1, p. 15–19, 2012. ISSN 1932-6203. Citado na página 64.

BRUEGGE, J. zur; EINSPANIER, R.; SHARBATI, S. A Long Journey Ahead: Long Non-coding RNAs in Bacterial Infections. *Frontiers in Cellular and Infection Microbiology*, v. 7, n. March, p. 1–6, 2017. ISSN 2235-2988. Disponível em: <http://journal.frontiersin.org/article/10.3389/fcimb.2017.00095/full>. Citado na página 33.

BYK, L. A. et al. Dengue Virus Genome Uncoating Requires Ubiquitination. *mBio*, v. 7, n. 3, p. e00804–16, 2016. ISSN 2150-7511. Disponível em: <http://mbio.asm.org/lookup/doi/10.1128/mBio.00804-16>. Citado na página 22.

CARMO-FONSECA, M.; RINO, J. RNA seeds nuclear bodies. *Nature cell biology*, Nature Publishing Group, v. 13, n. 2, p. 110–2, feb 2011. ISSN 1476-4679. Disponível em: <http://dx.doi.org/10.1038/ncb2161><http://www.ncbi.nlm.nih.gov/pubmed/10606244><http://www.ncbi.nlm.nih.gov/pubmed/21283118>. Citado na página 27.

CARRIERI, C. et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, Nature Publishing Group, v. 491, n. 7424, p. 454–457, 2012. ISSN 0028-0836. Disponível em: <http://www.nature.com/doi/10.1038/nature11508>. Citado na página 27.

CAZALLA, D.; YARIO, T.; STEITZ, J. A. Down-Regulation of a Host MicroRNA by a Herpesvirus saimiri Noncoding RNA. *Science*, v. 328, n. 5985, p. 1563–1566, jun 2010. ISSN 0036-8075. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.1187197>. Citado na página 29.

COSTA, V. V. et al. Inflammatory and innate immune responses in dengue infection: Protection versus disease induction. *American Journal of Pathology*, American Society for Investigative Pathology, v. 182, n. 6, p. 1950–1961, 2013. ISSN 00029440. Disponível em: <http://dx.doi.org/10.1016/j.ajpath.2013.02.027>. Citado 2 vezes nas páginas 65 e 66.

CROFT, D. et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, v. 42, n. D1, p. 481–487, 2014. ISSN 03051048. Citado na página 42.

CSÁRDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal Complex Systems*, v. 1695, p. 1–9, 2006. ISSN 0095-1137. Citado na página 44.

De Alboran, I. M. et al. Analysis of c-Myc function in normal cells via conditional gene-targeted mutation. *Immunity*, v. 14, n. 1, p. 45–55, 2001. ISSN 10747613. Citado na página 67.

DEVIGNOT, S. et al. Genome-wide expression profiling deciphers host responses altered during dengue shock syndrome and reveals the role of innate immunity in severe dengue. *PloS one*, v. 5, n. 7, p. e11671, jan 2010. ISSN 1932-6203. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2907396&tool=pmcentrez&rendertype=ab>. Citado 2 vezes nas páginas 23 e 24.

DIMITROVA, N. et al. LincRNA-p21 Activates p21 In cis to Promote Polycomb Target Gene Expression and to Enforce the G1/S Checkpoint. *Molecular Cell*, Elsevier Inc., v. 54, n. 5, p. 777–790, 2014. ISSN 10974164. Disponível em: <http://dx.doi.org/10.1016/j.molcel.2014.04.025>. Citado na página 43.

DONG, Y. et al. Long noncoding RNAs coordinate functions between mitochondria and the nucleus. *Epigenetics & Chromatin*, BioMed Central, v. 10, n. 1, p. 41, 2017. ISSN 1756-8935. Disponível em: <http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-017-0149-x>. Citado na página 26.

DURINCK, S. et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, v. 21, n. 16, p. 3439–3440, 2005. ISSN 13674803. Citado na página 43.

DURINCK, S. et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, v. 4, n. 8, p. 1184–1191, 2009. ISSN 1754-2189. Citado na página 43.

ETEBARI, K. et al. Dengue virus infection alters post-transcriptional modification of microRNAs in the mosquito vector *Aedes aegypti*. *Scientific Reports*, Nature Publishing Group, v. 5, n. 1, p. 15968, 2015. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/srep15968>. Citado na página 31.

EWELS, P. et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, v. 32, n. 19, p. 3047–3048, 2016. ISSN 14602059. Citado na página 45.

FAGHIHI, M. A. et al. Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biology*, v. 11, n. 5, p. R56, 2010. ISSN 1465-6906. Disponível em: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-5-r56>. Citado na página 27.

FENG, J. et al. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes and Development*, v. 20, n. 11, p. 1470–1484, 2006. ISSN 08909369. Citado na página 25.

FENG, L. et al. Transcriptome analysis reveals differentially expressed lncRNAs between oral squamous cell carcinoma and healthy oral mucosa. *Oncotarget*, v. 8, n. 19, p. 31521–31531, may 2017. ISSN 1949-2553. Citado na página 64.

FENG, Y.; ZHANG, L. *Regulatory Non- Coding RNAs IN Series Editor*. [S.l.: s.n.]. ISBN 9781493913688. Citado na página 32.

FORTES, P.; MORRIS, K. V. Long noncoding RNAs in viral infections. *Virus Research*, Elsevier B.V., v. 212, p. 1–11, 2016. ISSN 18727492. Disponível em: <http://dx.doi.org/10.1016/j.virusres.2015.10.002>. Citado na página 29.

GARCIA-BATES, T. M. et al. Association between Magnitude of the Virus-Specific Plasmablast Response and Disease Severity in Dengue Patients. *The Journal of Immunology*, v. 190, n. 1, p. 80–87, jan 2013. ISSN 0022-1767. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20402989><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2858352><http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1103350>. Citado na página 65.

GAUTIER, L. et al. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, v. 20, n. 3, p. 307–315, 2004. ISSN 13674803. Citado na página 40.

GEISLER, S.; COLLIER, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*, v. 14, n. 11, p. 699–712, 2013. ISSN 1471-0072. Disponível em: <http://www.nature.com/doi/10.1038/nrm3679>. Citado na página 26.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, v. 5, n. 10, p. R80, jan 2004. ISSN 1465-6914. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545600&tool=pmcentrez&rendertype=ab>. Citado na página 39.

GEZER, U. et al. Short Communication Cell Biology International. *Cell Biology International*, n. April 2014, p. 1–20, 2014. Citado na página 26.

GOMES, A. L. V. et al. Classification of dengue fever patients based on gene expression data using support vector machines. *PloS one*, v. 5, n. 6, p. e11267, jan 2010. ISSN 1932-6203. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2890409&tool=pmcentrez&rendertype=ab>. Citado na página 23.

GONG, C.; MAQUAT, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, Nature Publishing Group, v. 470, n. 7333, p. 284–8, feb 2011. ISSN 1476-4687. Disponível em: <http://www.nature.com/doi/10.1038/nature09701><http://www.ncbi.nlm.nih.gov/pubmed/21307942><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3073508>. Citado 2 vezes nas páginas 26 e 67.

GRAVITZ, L. Therapy: This Time it's Personal. *Scientific American*, v. 311, n. 1, p. 106–108, jun 2014. ISSN 0036-8733. Disponível em: <http://www.nature.com/doi/10.1038/scientificamerican0714-S6>. Citado na página 31.

GROEN, A. K. The pros and cons of gene expression analysis by microarrays. *Journal of Hepatology*, v. 35, n. 2, p. 295–296, 2001. ISSN 01688278. Citado na página 32.

GROTE, P.; HERRMANN, B. G. The long non-coding RNA *Fendrr* links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biology*, v. 10, n. 10, p. 1579–1585, 2013. ISSN 1547-6286. Disponível em: <http://www.tandfonline.com/doi/abs/10.4161/rna.26165>. Citado na página 26.

HANSEN, T. B. et al. Natural RNA circles function as efficient microRNA sponges. *Nature*, Nature Publishing Group, v. 495, n. 7441, p. 384–388, 2013. ISSN 0028-0836. Disponível em: <http://www.nature.com/doi/10.1038/nature11993>. Citado na página 27.

HARROW, J. et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, v. 22, n. 9, p. 1760–1774, 2012. ISSN 10889051. Citado na página 36.

HEWSON, C. et al. Extracellular vesicle associated long non-coding RNAs functionally enhance cell viability. *Non-coding RNA Research*, Elsevier Ltd, v. 1, n. 1, p. 3–11, 2016. ISSN 24680540. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S2468054016300014>. Citado na página 26.

HINRICHS, A. S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic acids research*, v. 34, n. Database issue, p. D590–8, jan 2006. ISSN 1362-4962. Citado na página 37.

HOANG, L. T. et al. The early whole-blood transcriptional signature of dengue virus and features associated with progression to dengue shock syndrome in Vietnamese children and young adults. *Journal of virology*, v. 84, n. 24, p. 12982–94, dec 2010. ISSN 1098-5514. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3004338&tool=pmcentrez&rendertype=ab>. Citado na página 23.

HU, G. et al. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nature immunology*, v. 14, n. 11, p. 1190–8, 2013. ISSN 1529-2916. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3805781&tool=pmcentrez&rendertype=ab>. Citado na página 29.

IIOTT, N. E. et al. Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nature Communications*, v. 5, 2014. ISSN 2041-1723. Disponível em: <http://www.nature.com/doifinder/10.1038/ncomms4979>. Citado na página 33.

IRIZARRY, R. a. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, v. 4, p. 249–264, 2003. ISSN 1465-4644. Citado na página 40.

IYER, M. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*, v. 47, n. 3, p. 199–208, 2015. ISSN 1061-4036. Disponível em: <http://dx.doi.org/10.1038/ng.3192>. Citado 2 vezes nas páginas 33 e 36.

JIANG, N. et al. lncRNA DANCER promotes tumor progression and cancer stemness features in osteosarcoma by upregulating AXL via miR-33a-5p inhibition. *Cancer Letters*, Elsevier Ltd, v. 405, n. June, p. 46–55, 2017. ISSN 18727980. Disponível em: <http://dx.doi.org/10.1016/j.canlet.2017.06.009>. Citado na página 65.

JULIA, A. et al. An eight-gene blood expression profile predicts the response to infliximab in rheumatoid arthritis. *PLoS ONE*, v. 4, n. 10, 2009. ISSN 19326203. Citado na página 31.

KANLAYA, R. et al. The ubiquitin-proteasome pathway is important for dengue virus infection in primary human endothelial cells. *Journal of proteome research*, v. 9, n. 10, p. 4960–71, oct 2010. ISSN 1535-3907. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/pr100219yhttp://www.ncbi.nlm.nih.gov/pubmed/20718508>. Citado na página 22.

KAPOOR, P.; FRAPPIER, L. EBNA1 partitions Epstein-Barr virus plasmids in yeast cells by attaching to human EBNA1-binding protein 2 on mitotic chromosomes. *Journal of virology*, v. 77, n. 12, p. 6946–6956, 2003. ISSN 0022-538X. Citado na página 67.

KAUFFMANN, A.; GENTLEMAN, R.; HUBER, W. arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics (Oxford, England)*, v. 25, n. 3, p. 415–6, feb 2009. ISSN 1367-4811. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2639074&tool=pmcentrez&rendertype=ab>. Citado na página 46.

- KELSEN, J. R. et al. Maintaining Intestinal Health: The Genetics and Immunology of Very Early Onset Inflammatory Bowel Disease. *Cmgh*, Elsevier Inc, v. 1, n. 5, p. 462–476, 2015. ISSN 2352345X. Disponível em: <http://dx.doi.org/10.1016/j.jcmgh.2015.06.010>. Citado na página 31.
- KENT, W. J. BLAT — The BLAST -Like Alignment Tool. *Genome research*, v. 12, p. 656–664, 2002. ISSN 1088-9051. Citado na página 37.
- KHATRI, P.; SIROTA, M.; BUTTE, A. J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, v. 8, n. 2, 2012. ISSN 1553734X. Citado na página 43.
- KRETZ, M. et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, Nature Publishing Group, v. 493, n. 7431, p. 231–235, 2012. ISSN 0028-0836. Disponível em: <http://www.nature.com/doi/10.1038/nature11661>. Citado na página 27.
- KRUEGER, F. *Trim Galore!* 2012. Disponível em: https://www.bioinformatics.babraham.ac.uk/projects/trim{_}galo. Citado na página 45.
- KUNG, J. T. Y.; COLOGNORI, D.; LEE, J. T. Long Noncoding RNAs: Past, Present, and Future. *Genetics*, v. 193, n. 3, p. 651–669, 2013. ISSN 0016-6731. Disponível em: <http://www.genetics.org/cgi/doi/10.1534/genetics.112.146704>. Citado na página 26.
- KWISSA, M. et al. Dengue virus infection induces expansion of a CD14+CD16+ monocyte population that stimulates plasmablast differentiation. *Cell Host and Microbe*, Elsevier Inc., v. 16, n. 1, p. 115–127, 2014. ISSN 19346069. Disponível em: <http://dx.doi.org/10.1016/j.chom.2014.06.001>. Citado na página 23.
- KWISSA, M. et al. Dengue Virus Infection Induces Expansion of a CD14(+)CD16(+) Monocyte Population that Stimulates Plasmablast Differentiation. *Cell host & microbe*, Elsevier Inc., p. 1–13, jun 2014. ISSN 1934-6069. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24981333>. Citado 3 vezes nas páginas 22, 65 e 69.
- LAM, P. K. et al. The value of daily platelet counts for predicting dengue shock syndrome: Results from a prospective observational study of 2301 Vietnamese children with dengue. *PLoS Neglected Tropical Diseases*, v. 11, n. 4, p. 1–20, 2017. ISSN 19352735. Citado na página 21.
- LAM, P. K. et al. Clinical characteristics of dengue shock syndrome in vietnamese children: A 10-year prospective study in a single hospital. *Clinical Infectious Diseases*, v. 57, n. 11, p. 1577–1586, 2013. ISSN 10584838. Citado na página 21.
- LANGFELDER, P.; HORVATH, S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, v. 19, n. 2, p. 200–6, 2004. ISSN 0884-0431. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/14969389>. Citado na página 47.
- LANGFELDER, P.; HORVATH, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, v. 9, p. 559, 2008. ISSN 1471-2105. Citado 3 vezes nas páginas 47, 67 e 68.

LAU, C. C. et al. Viral-human chimeric transcript predisposes risk to liver cancer development and progression. *Cancer Cell*, Elsevier Inc., v. 25, n. 3, p. 335–349, 2014. ISSN 18783686. Disponível em: <http://dx.doi.org/10.1016/j.ccr.2014.01.030>. Citado na página 29.

LEE, N. et al. EBV noncoding RNA binds nascent RNA to drive host PAX5 to viral DNA. *Cell*, Elsevier Inc., v. 160, n. 4, p. 607–618, 2015. ISSN 10974172. Disponível em: <http://dx.doi.org/10.1016/j.cell.2015.01.015>. Citado na página 29.

LEE, S. et al. Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell*, Elsevier Inc., v. 164, n. 1-2, p. 69–80, 2016. ISSN 10974172. Disponível em: <http://dx.doi.org/10.1016/j.cell.2015.12.017>. Citado na página 27.

LEE, T. et al. Altered expression of cellular proliferation, apoptosis and the cell cycle-related genes in lung cancer cells with acquired resistance to EGFR tyrosine kinase inhibitors. *Oncology Letters*, p. 2191–2197, 2017. ISSN 1792-1074. Disponível em: <http://www.spandidos-publications.com/10.3892/ol.2017.6428>. Citado na página 31.

LEEUWEN, S. van; MIKKERS, H. Long non-coding RNAs: Guardians of development. *Differentiation*, Elsevier, v. 80, n. 4-5, p. 175–183, 2010. ISSN 03014681. Disponível em: <http://dx.doi.org/10.1016/j.diff.2010.07.003>. Citado na página 26.

LI, B.; DEWEY, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, v. 12, n. 1, p. 323, 2011. ISSN 1471-2105. Disponível em: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323>. Citado na página 46.

LI, S. et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature immunology*, v. 15, n. 2, p. 195–204, 2014. ISSN 1529-2916. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3946932&tool=pmcentrez&rendertype=ab>. Citado na página 42.

LI, W. et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, v. 498, n. 7455, p. 516–20, 2013. ISSN 1476-4687. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3718886&tool=pmcentrez&rendertype=ab>. Citado na página 43.

LIAO, Q. et al. NcFANs: A web server for functional annotation of long non-coding RNAs. *Nucleic Acids Research*, v. 39, n. SUPPL. 2, p. 118–124, 2011. ISSN 03051048. Citado na página 32.

LOKE, P. et al. Gene expression patterns of dengue virus-infected children from nicaragua reveal a distinct signature of increased metabolism. *PLoS neglected tropical diseases*, v. 4, n. 6, p. e710, jan 2010. ISSN 1935-2735. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2886038&tool=pmcentrez&rendertype=ab>. Citado na página 23.

LONG, H. T. et al. Patterns of gene transcript abundance in the blood of children with severe or uncomplicated dengue highlight differences in disease evolution and host response to dengue virus infection. *The Journal of infectious diseases*, v. 199, n. 4, p. 537–46, feb 2009. ISSN 0022-1899. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19138155>. Citado na página 23.

- MAASS, P. G.; LUFT, F. C.; BÄHRING, S. Long non-coding RNA in health and disease. *Journal of Molecular Medicine*, v. 92, n. 4, p. 337–346, 2014. ISSN 14321440. Citado na página 26.
- MAO, Z. et al. LncRNA DANCER Promotes Migration and Invasion through Suppression of LncRNA-LET in Gastric Cancer Cells. *Bioscience Reports*, p. BSR20171070, 2017. ISSN 0144-8463. Disponível em: <http://bioscirep.org/lookup/doi/10.1042/BSR20171070>. Citado na página 65.
- MARINER, P. D. et al. Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. *Molecular Cell*, v. 29, n. 4, p. 499–509, 2008. ISSN 10972765. Citado 2 vezes nas páginas 26 e 67.
- MARQUES-TOLEDO, C. d. A. et al. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLOS Neglected Tropical Diseases*, v. 11, n. 7, p. e0005729, jul 2017. ISSN 1935-2735. Disponível em: <http://dx.plos.org/10.1371/journal.pntd.0005729>. Citado na página 19.
- MERCER, T. R.; DINGER, M. E.; MATTICK, J. S. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, v. 10, n. 3, p. 155–159, 2009. ISSN 1471-0056. Disponível em: <http://www.nature.com/doi/10.1038/nrg2521>. Citado na página 26.
- MO, X.-B. et al. Identification and evaluation of lncRNA and mRNA integrative modules in human peripheral blood mononuclear cells. *Epigenomics*, p. epi-2016-0178, 2017. ISSN 1750-1911. Citado na página 64.
- MORAN, V. a.; PERERA, R. J.; KHALIL, A. M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Research*, v. 40, n. 14, p. 6391–6400, 2012. ISSN 03051048. Citado na página 26.
- MUKHOPADHYAY, S.; KUHN, R. J.; ROSSMANN, M. G. A structural perspective of the flavivirus life cycle. *Nature Reviews Microbiology*, v. 3, n. 1, p. 13–22, 2005. ISSN 1740-1526. Disponível em: <http://www.nature.com/doi/10.1038/nrmicro1067>. Citado na página 21.
- NAKAYA, H. I. et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome biology*, v. 8, n. 3, p. R43, jan 2007. ISSN 1465-6914. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1868932&tool=pmcentrez&rendertype=ab>. Citado na página 25.
- NASCIMENTO, E. J. M. et al. Gene expression profiling during early acute febrile stage of dengue infection can predict the disease outcome. *PLoS ONE*, v. 4, n. 11, 2009. ISSN 19326203. Citado na página 23.
- O'LEARY, V. B. et al. PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation. *Cell Reports*, v. 11, n. 3, p. 474–485, 2015. ISSN 22111247. Citado 2 vezes nas páginas 26 e 67.
- OSANAI, C. H. et al. Surto De Dengue Em Boa Vista, Roraima. Nota Previa. *Revista do Instituto de Medicina Tropical de Sao Paulo*, v. 25, n. 1, p. 53–54, 1983. ISSN 00364665. Citado 2 vezes nas páginas 19 e 22.

POHL, M. O. et al. Identification of Polo-like kinases as potential novel drug targets for influenza A virus. *Scientific reports*, v. 7, n. 1, p. 8629, aug 2017. ISSN 2045-2322. Citado na página 65.

PONS, P.; LATAPY, M. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, v. 10, n. 2, p. 191–218, 2006. ISSN 07403194. Disponível em: <http://arxiv.org/abs/physics/0512106>. Citado 2 vezes nas páginas 44 e 68.

QUINLAN, A. R.; HALL, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, v. 26, n. 6, p. 841–842, 2010. ISSN 13674803. Citado na página 38.

RANZANI, V. et al. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nature immunology*, v. 16, n. 3, p. 318–25, 2015. ISSN 1529-2916. Disponível em: <http://www.nature.com/pros.lib.unimi.it/ni/journal/v16/n3/full/ni.3093.html>. Citado na página 30.

RITCHIE, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, v. 43, n. 7, p. e47, 2015. ISSN 13624962. Citado 2 vezes nas páginas 41 e 46.

ROBINSON, D. R. et al. Integrative clinical genomics of metastatic cancer. *Nature*, 2017. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/nature23306>. Citado na página 31.

RODENHUIS-ZYBERT, I. A.; WILSCHUT, J.; SMIT, J. M. Dengue virus life cycle: Viral and host factors modulating infectivity. *Cellular and Molecular Life Sciences*, v. 67, n. 16, p. 2773–2786, 2010. ISSN 1420682X. Citado 3 vezes nas páginas 22, 23 e 69.

RUSSO, P. et al. *CEMiTool: Co-expression Modules identification Tool. R package version 0.99.11*. 2017. Citado 2 vezes nas páginas 44 e 67.

SAAYMAN, S. et al. An HIV-Encoded Antisense Long Noncoding RNA Epigenetically Regulates Viral Transcription. *Molecular Therapy*, The American Society of Gene & Cell Therapy, v. 22, n. 6, p. 1164–1175, 2014. ISSN 15250016. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1525001616307043>. Citado na página 29.

SCHMITZ, S. U.; GROTE, P.; HERRMANN, B. G. Mechanisms of long noncoding RNA function in development and disease. *Cellular and Molecular Life Sciences*, Springer International Publishing, v. 73, n. 13, p. 2491–2509, 2016. ISSN 14209071. Citado na página 28.

SCREATON, G. et al. New insights into the immunopathology and control of dengue virus infection. *Nature Publishing Group*, v. 15, n. 12, p. 745–759, 2015. ISSN 1474-1733. Disponível em: <http://dx.doi.org/10.1038/nri3916>. Citado 4 vezes nas páginas 19, 20, 21 e 23.

SEAN, D.; MELTZER, P. S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, v. 23, n. 14, p. 1846–1847, 2007. ISSN 13674803. Citado 2 vezes nas páginas 39 e 46.

SERGUSHICHEV, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, p. 60012, 2016. Citado na página 43.

SESSIONS, O. M. et al. Host cell transcriptome profile during wild-type and attenuated dengue virus infection. *PLoS neglected tropical diseases*, v. 7, n. 3, p. e2107, jan 2013. ISSN 1935-2735. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3597485&tool=pmcentrez&rendertype=ab>. Citado na página 31.

SHA, S. et al. Targeting long non-coding RNA DANCR inhibits triple negative breast cancer progression. *Biology Open*, v. 6, n. 9, p. 1310–1316, 2017. ISSN 2046-6390. Disponível em: <http://bio.biologists.org/lookup/doi/10.1242/bio.023135>. Citado na página 65.

SIMMONS, C. P. et al. Patterns of host genome-wide gene transcript abundance in the peripheral blood of patients with acute dengue hemorrhagic fever. *The Journal of infectious diseases*, v. 195, n. 8, p. 1097–1107, apr 2007. ISSN 0022-1899. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4042601&tool=pmcentrez&rendertype=ab>. Citado 4 vezes nas páginas 23, 24, 65 e 66.

SMYTH, G. K. Limma : Linear Models for Microarray Data. In: GENTLEMAN, R. et al. (Ed.). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. [S.l.]: Springer, 2005. p. 397–420. Citado na página 41.

SPURLOCK, C. F. et al. Expression and functions of long noncoding RNAs during human T helper cell differentiation. *Nature Communications*, Nature Publishing Group, v. 6, apr 2015. ISSN 2041-1723. Citado na página 29.

SUBRAMANIAN, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, v. 102, n. 43, p. 15545–50, 2005. ISSN 0027-8424. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16199517>. Citado 3 vezes nas páginas 42, 43 e 67.

SUN, P. et al. Sequential Waves of Gene Expression in Patients with Clinically Defined Dengue Illnesses Reveal Subtle Disease Phases and Predict Disease Severity. *PLoS Neglected Tropical Diseases*, v. 7, n. 7, 2013. ISSN 1935-2727. Citado 3 vezes nas páginas 23, 24 e 66.

SUN, Y. et al. Identification of long non-coding RNAs biomarkers associated with progression of endometrial carcinoma and patient outcomes. *Oncotarget*, v. 8, n. 32, p. 52604–52613, aug 2017. ISSN 1949-2553. Disponível em: <http://www.oncotarget.com/fulltext/17537>. Citado na página 26.

TARCA, A. L.; ROMERO, R.; DRAGHICI, S. Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics and gynecology*, v. 195, n. 2, p. 373–88, aug 2006. ISSN 1097-6868. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2435252&tool=pmcentrez&rendertype=ab>. Citado na página 41.

TEO, C. S. H.; CHU, J. J. H. Cellular Vimentin Regulates Construction of Dengue Virus Replication Complexes through Interaction with NS4A Protein. *Journal of Virology*, v. 88, n. 4, p. 1897–1913, 2014. ISSN 0022-538X. Disponível em: <http://jvi.asm.org/cgi/doi/10.1128/JVI.01249-13>. Citado na página 22.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, v. 25, n. may, p. 25–29, 2000. ISSN 1061-4036. Disponível em: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Gene+Ontology:+tool+for+the+unification+of+bio>. Citado na página 42.

THOREAU, M. et al. Vaccine-induced tumor regression requires a dynamic cooperation between T cells and myeloid cells at the tumor site. *Oncotarget*, v. 6, n. 29, p. 27832–27846, 2015. ISSN 1949-2553 (Electronic). Citado na página 31.

TIAN, X. et al. Differentially Expressed lncRNAs in Gastric Cancer Patients: A Potential Biomarker for Gastric Cancer Prognosis. *Journal of Cancer*, v. 8, n. 13, p. 2575–2586, 2017. ISSN 1837-9664. Disponível em: <http://www.jcancer.org/v08p2575.htm>. Citado na página 26.

TRIPATHI, V. et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell*, Elsevier Ltd, v. 39, n. 6, p. 925–938, 2010. ISSN 10972765. Disponível em: <http://dx.doi.org/10.1016/j.molcel.2010.08.011>. Citado 2 vezes nas páginas 25 e 26.

UBOL, S. et al. Differences in global gene expression in peripheral blood mononuclear cells indicate a significant role of the innate responses in progression of dengue fever but not dengue hemorrhagic fever. *The Journal of infectious diseases*, v. 197, n. 10, p. 1459–67, may 2008. ISSN 0022-1899. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18444802>. Citado na página 23.

VOLDERS, P. J. et al. LNCipedia: A database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research*, v. 41, n. D1, p. 246–251, 2013. ISSN 03051048. Citado na página 36.

VOLDERS, P. J. et al. An update on LNCipedia: A database for annotated human lncRNA sequences. *Nucleic Acids Research*, v. 43, n. D1, p. D174–D180, 2015. ISSN 13624962. Citado na página 36.

WANG, K. C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, Nature Publishing Group, v. 472, n. 7341, p. 120–4, 2011. ISSN 1476-4687. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3670758&tool=pmcentrez&rendertype=ab>. Citado na página 25.

WANG, W.-T. et al. Genome-wide Long Non-coding RNA Analysis Identified Circulating lncRNAs as Novel Non-invasive Diagnostic Biomarkers for Gynecological Disease. *Scientific Reports*, v. 6, n. 1, p. 23343, 2016. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/srep23343>. Citado na página 26.

WARKE, R. V. et al. TRAIL is a novel antiviral protein against dengue virus. *Journal of virology*, v. 82, n. 1, p. 555–64, jan 2008. ISSN 1098-5514. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2224358&tool=pmcentrez&rendertype=ab>. Citado na página 23.

WEI, T.; SIMKO, V. *corrplot: Visualization of a Correlation Matrix*. 2016. Citado na página 47.

- WESTERMANN, A. J. et al. Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions. *Nature*, Nature Publishing Group, v. 529, n. 7587, p. 496–501, 2016. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/nature16547>. Citado na página 31.
- WICKHAM, H. *ggplot2*. New York, NY: Springer New York, 2009. ISBN 978-0-387-98140-6. Disponível em: <http://link.springer.com/10.1007/978-0-387-98141-3>. Citado na página 47.
- WOHLKONIG, A. et al. Structural basis of quinolone inhibition of type IIA topoisomerases and target-mediated resistance. *Nature Structural & Molecular Biology*, Nature Publishing Group, v. 17, n. 99, p. 1152–1153, 2010. ISSN 1545-9993. Disponível em: <http://www.nature.com/doifinder/10.1038/nsmb.1892>. Citado na página 26.
- WONG, L. H. et al. Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. *Genome Research*, v. 17, n. 8, p. 1146–1160, 2007. ISSN 10889051. Citado na página 25.
- World Health Organization. Dengue vaccine: WHO position paper, July 2016 – recommendations. *Vaccine*, v. 35, n. 9, p. 1200–1201, 2017. ISSN 18732518. Citado na página 21.
- XIE, C. et al. NONCODEv4: Exploring the world of long non-coding RNA genes. *Nucleic Acids Research*, v. 42, n. D1, p. 98–103, 2014. ISSN 03051048. Citado na página 36.
- YANG, L. et al. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature*, Nature Publishing Group, v. 500, n. 7464, p. 598–602, 2013. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/nature12451>. Citado na página 26.
- YU, G. et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, v. 16, n. 5, p. 284–287, 2012. ISSN 1536-2310. Citado na página 42.
- YU, W. et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature*, v. 451, n. 7175, p. 202–6, 2008. ISSN 1476-4687. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2743558&tool=pmcentrez&rendertype=ab>. Citado na página 43.
- ZHANG, L. et al. Whole transcriptome microarrays identify long non-coding RNAs associated with cardiac hypertrophy. *Genomics Data*, Elsevier B.V., v. 5, p. 68–71, 2015. ISSN 22135960. Disponível em: <http://dx.doi.org/10.1016/j.gdata.2015.05.014>. Citado na página 32.
- ZHANG, Y. et al. Long noncoding RNA expression profile in fibroblast-like synoviocytes from patients with rheumatoid arthritis. *Arthritis Research & Therapy*, Arthritis Research & Therapy, v. 18, n. 1, p. 227, 2016. ISSN 1478-6362. Disponível em: <http://arthritis-research.biomedcentral.com/articles/10.1186/s13075-016-1129-4>. Citado na página 64.

Anexo A – Artigos Publicados

Cell Host & Microbe
Short Article

CellPress

Sequential Infection with Common Pathogens Promotes Human-like Immune Gene Expression and Altered Vaccine Response

Tiffany A. Reese,^{1,*} Kevin Bi,² Amal Kambal,³ Ali Filali-Mouhim,⁴ Lalit K. Beura,⁵ Matheus C. Bürger,⁶ Bali Pulendran,⁷ Rafick-Pierre Sekaly,⁴ Stephen C. Jameson,⁸ David Masopust,⁵ W. Nicholas Haining,² and Herbert W. Virgin^{3,*}¹Departments of Immunology and Microbiology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA²Department of Pediatric Oncology, Dana-Farber Cancer Institute, and Pediatric Hematology and Oncology, Children's Hospital, Boston, MA 02115, and the Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, MA 02142, USA³Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA⁴Department of Pathology, Case Western Reserve University, Cleveland, OH 44106, USA⁵Department of Microbiology and Immunology, Center for Immunology, University of Minnesota, Minneapolis, MN 55455, USA⁶Department of Clinical Analyses and Toxicology, School of Pharmaceutical Science at University of São Paulo, São Paulo 05508, Brazil⁷Emory Vaccine Center, Yerkes National Primate Research Center and Department of Pathology, Emory University, Atlanta, GA 30329, USA⁸Department of Laboratory Medicine and Pathology, Center for Immunology, University of Minnesota, Minneapolis, MN 55455, USA*Correspondence: tiffany.reese@utsouthwestern.edu (T.A.R.), virgin@wustl.edu (H.W.V.)<http://dx.doi.org/10.1016/j.chom.2016.04.003>

SUMMARY

Immune responses differ between laboratory mice and humans. Chronic infection with viruses and parasites are common in humans, but are absent in laboratory mice, and thus represent potential contributors to inter-species differences in immunity. To test this, we sequentially infected laboratory mice with herpesviruses, influenza, and an intestinal helminth and compared their blood immune signatures to mock-infected mice before and after vaccination against yellow fever virus (YFV-17D). Sequential infection altered pre- and post-vaccination gene expression, cytokines, and antibodies in blood. Sequential pathogen exposure induced gene signatures that recapitulated those seen in blood from pet store-raised versus laboratory mice, and adult versus cord blood in humans. Therefore, basal and vaccine-induced murine immune responses are altered by infection with agents common outside of barrier facilities. This raises the possibility that we can improve mouse models of vaccination and immunity by selective microbial exposure of laboratory animals to mimic that of humans.

INTRODUCTION

Substantial variation in human immune responses is due to environmental influences (Brodin et al., 2015; Roederer et al., 2015). Potential variables include nutritional status, different health practices, age, socioeconomic status, and geographic location. In addition, the bacterial microbiome influences immune and inflammatory responses (Honda and Littman, 2012; Hooper et al., 2012). An added, but less well understood, envi-

ronmental contributor to variation is the history of infection with acute and chronic pathogens, including herpesviruses and intestinal parasites (Foxman and Iwasaki, 2011; Furman et al., 2015; Salgame et al., 2013; Virgin, 2014; Virgin et al., 2009). Persistent infections change the immune response to unrelated pathogens and vaccines (Furman et al., 2015; Oldstone, 2005; Osborne et al., 2014; Reese et al., 2014; Salgame et al., 2013; Selin et al., 2006; Slifka et al., 2003; Virgin, 2014). Some chronic co-infections enhance, while others inhibit, immunity to secondary challenge (Barton et al., 2007; MacDuff et al., 2015; Stelekati et al., 2014; Stelekati and Wherry, 2012). Moreover, humans are frequently infected with acute viral pathogens, which may change the immune system (Foxman and Iwasaki, 2011).

There is concern that rodent models do not faithfully predict human immune responses (Mestas and Hughes, 2004; Seok et al., 2013; Takao and Miyakawa, 2015), limiting the value of this powerful model system. However, mouse models are indispensable for biomedical studies and play a significant role in the development of vaccines and therapeutics. This highlights the need for studies that identify environmental variables that might, in addition to chromosomal genetic variation, contribute to species-specific immune response differences between mice and humans. Notably, barrier-raised mice are free of many acute and chronic infections that are recognized to contribute to human immune variation (Salgame et al., 2013; Virgin, 2014). For example, people chronically infected with intestinal helminths have lower responses to vaccination with *Bacillus Calmette-Guérin* (BCG) (Elias et al., 2001), cholera (Cooper et al., 2001) and tetanus toxoid (Nookala et al., 2004; Sabin et al., 1996). Furthermore, chronic infection with the herpesvirus cytomegalovirus (HCMV) alters responses to human influenza vaccination (Furman et al., 2015), and infection of mice with murine CMV (MCMV) and/or a murine γ -herpesvirus (MHV68) alters bacterial immunity and reverses inherited immunodeficiency (Barton et al., 2007; MacDuff et al., 2015). We therefore sought to test the hypothesis that infection history, and in particular

JVI Accepted Manuscript Posted Online 7 December 2016

J. Virol. doi:10.1128/JVI.01844-16

Copyright © 2016, American Society for Microbiology. All Rights Reserved.

1 Adjuvanting an SIV vaccine with TLR ligands encapsulated in nanoparticles
2 induces persistent antibody responses and enhanced protection in TRIM5α
3 restrictive macaques

4

5 Sudhir Pai Kasturi¹, Pamela A. Kozlowski², Helder I Nakaya^{1,3}, Matheus C. Burger³,
 6 Pedro Russo³, Mathew Pham¹, Yevgeniy Kovalenkov¹, Eduardo L. V. Silveira^{1,3}, Colin
 7 Havenar-Daughton⁴, Samantha L. Burton¹, Katie M. Kilgore¹, Mathew J. Johnson¹, Rafiq
 8 Nabi², Traci Legere¹, Zarpheen Jinnah Sher¹, Xuemin Chen⁵, Rama R. Amara^{1,6}, Eric
 9 Hunter^{1,7}, Steven E. Bosinger¹, Paul Spearman^{1,5}, Shane Crotty⁴, Francois Villinger^{1,8},
 10 Cynthia A. Derdeyn^{1,7}, Jens Wrannert^{1,5} and Bali Pulendran^{1,7}

11

12 Affiliations

13 ¹ Emory Vaccine Center/Yerkes National Primate Research Center at Emory
 14 University, 954, Gatewood Rd, Atlanta, Georgia, USA

15 ² Department of Microbiology, Immunology, and Parasitology, Louisiana State
 16 University Health Sciences Center, New Orleans, Louisiana, USA 70112

17 ³ Faculty of Pharmaceutical Sciences, Department of Clinical and Toxicological
 18 Analyses, University of Sao Paulo, Brazil

19 ⁴ Division of Vaccine Discovery, La Jolla Institute of Allergy and Immunology, La Jolla,
 20 California, USA

21 ⁵ Department of Pediatrics, Emory University School of Medicine, 2015
 22 Uppergate Drive, Atlanta, GA 30322, USA

23 ⁶ Division of Microbiology and Immunology, Emory University, Rollins Research
 24 Center, 1510 Clifton Road, Atlanta GA 30322

25 ⁷ Department of Pathology and Laboratory Medicine, 1364, Clifton Rd, Atlanta GA

LETTER

doi:10.1038/nature19330

Defining CD8⁺ T cells that provide the proliferative burst after PD-1 therapySe Jin Im¹, Masao Hashimoto¹, Michael Y. Gerner², Junghwa Lee¹, Haydn T. Kissick³, Matheus C. Burger⁴, Qiang Shan⁵, J. Scott Hale¹, Judong Lee¹, Tahseen H. Nasti¹, Arlene H. Sharpe⁷, Gordon J. Freeman⁸, Ronald N. Germain², Helder I. Nakaya⁴, Hai-Hui Xue^{5,6} & Rafi Ahmed¹

Chronic viral infections are characterized by a state of CD8⁺ T cell dysfunction that is associated with expression of the programmed cell death 1 (PD-1) inhibitory receptor^{1–4}. A better understanding of the mechanisms that regulate CD8⁺ T cell responses during chronic infection is required to improve immunotherapies that restore function in exhausted CD8⁺ T cells. Here we identify the population of virus-specific CD8⁺ T cells that proliferate after blockade of the PD-1 inhibitory pathway in mice chronically infected with lymphocytic choriomeningitis virus (LCMV). These LCMV-specific CD8⁺ T cells expressed the PD-1 inhibitory receptor but at the same time also expressed several co-stimulatory molecules such as ICOS, OX-40 and CD28. This CD8⁺ T cell subset was characterized by a unique gene signature that was related to CD4⁺ T follicular helper (Tfh) cells, CD8⁺ T cell memory precursors and hematopoietic stem cell progenitors, but that was distinct from CD4⁺ Th1 cells and CD8⁺ terminal effectors. This CD8⁺ T cell population was only found in lymphoid tissues and resided predominantly in the T cell zones along with naïve CD8⁺ T cells. These PD-1⁺ CD8⁺ T cells were stem cell-like during chronic LCMV infection, undergoing self-renewal and also differentiating into the terminally exhausted CD8⁺ T cells that were present in both lymphoid and non-lymphoid tissues. The proliferative burst after PD-1 blockade came almost exclusively from this CD8⁺ T cell subset. Importantly, the transcription factor TCF1 played a cell-intrinsic and essential role in the generation of this CD8⁺ T cell subset. These findings provide a better understanding of T cell exhaustion and have implications towards optimizing PD-1-directed immunotherapy in chronic infections and cancer.

Functional exhaustion of antigen-specific CD8⁺ T cells has been well documented during persistent infections^{1,2} and cancer³. A hallmark of exhausted CD8⁺ T cells is expression of various inhibitory receptors most notably PD-1⁴. Several studies have shown that the pool of exhausted CD8⁺ T cells is phenotypically and functionally heterogeneous^{5–8}. Our goal here was to better characterize the CD8⁺ T cells that are present during chronic viral infection. We were intrigued by a study showing that a subset of human CD8⁺ T cells express CXCR5⁹, a chemokine receptor, that is normally present on B cells and CD4⁺ T follicular helper (Tfh) cells. Another study described CXCR5⁺ CD8⁺ T cells that regulate autoimmunity in mice¹⁰. So we asked the question whether CXCR5⁺ CD8⁺ T cells were also generated during persistent viral infections. We addressed this issue using the mouse model of LCMV infection where T cell exhaustion was first documented¹. We found that there was a distinct population of CXCR5⁺ LCMV GP33-specific CD8⁺ T cells in the spleens of chronically infected mice (LCMV done 13) whereas GP33-specific memory CD8⁺ T cells in mice that had cleared the infection (LCMV Armstrong) did not express CXCR5

(Fig. 1a). The CXCR5⁺ CD8⁺ T cells in chronically infected mice also expressed the CD4⁺ Tfh markers, ICOS and Bcl-6, and were negative for Tim-3, a marker associated with CD4⁺ Th1 cells¹¹. In contrast, the CXCR5⁺ GP33-specific CD8⁺ T cells in chronically infected mice expressed Tim-3 and were negative for ICOS and Bcl-6. Both subsets of GP33-specific CD8⁺ T cells in chronically infected mice expressed high levels of the PD-1 inhibitory receptor with the CXCR5⁺ cells showing slightly higher levels (Fig. 1a). An identical pattern of expression of these molecules was seen with CD8⁺ T cells recognizing another LCMV epitope GP276 (Extended Data Fig. 1a). Thus, this novel population of CXCR5⁺ cells was seen with both tetramer positive CD8⁺ T cells and these cells were detectable as early as day 8 after infection and were stably maintained in mice with high levels of viremia (Fig. 1b, Extended Data Fig. 1b). To determine if the generation of these cells was due to antigen persistence or due to different tropism of LCMV done 13¹², mice were infected with either low dose (2 × 10² pfu) of done 13 that is controlled within a week or with high dose (2 × 10⁶ pfu) that causes a persistent infection. CXCR5⁺ LCMV-specific CD8⁺ T cells were only generated in the chronically infected mice, showing that antigen persistence drives the generation of this novel CD8⁺ T cell subset (Extended Data Fig. 2).

Transcriptional profiling revealed that the PD-1⁺ CXCR5⁺ and PD-1⁺ CXCR5⁺ CD8⁺ T cells in chronically infected mice had distinct gene signatures (Extended Data Fig. 3a). A striking finding was that the CXCR5⁺ CD8⁺ T cells expressed higher levels of several costimulatory molecules (*Cd28*, *Icos*, *Tnfrsf14* (LIGHT), *Tnfrsf4* (OX-40)) and lower levels of inhibitory receptors (*Cd244* (2B4), *Havcr2* (Tim-3), *Entpd1* (CD39), *Lag3*) compared to CXCR5⁺ cells (Fig. 1c, Extended Data Fig. 3b). These two CD8⁺ T cell populations also showed differences in the expression of effector molecules, chemokines and chemokine receptors, TLRs, transcription factors and memory markers (Fig. 1c and Extended Data Fig. 3b). CXCR5⁺ CD8⁺ T cells had higher levels of several effector molecules (perforin, granzymes, etc.), but did not express IL-2 or TNF, suggesting a more terminally differentiated state (Fig. 1c, Extended Data Fig. 4), confirming and extending earlier results with PD-1⁺ Tim-3⁺ CD8⁺ T cells⁷. A particularly interesting finding was that the two CD8⁺ T cell subsets expressed different *Tlr*s (Fig. 1c). TLRs are key molecules associated with innate immune responses but their role on CD8⁺ T cells is not well understood¹³. *Tlr3* and *Tlr7* were selectively upregulated by CXCR5⁺ CD8⁺ T cells and this was corroborated by enrichment of TLR cascade genes and interferon signaling pathways in this subset (Extended Data Fig. 5a). Regarding transcription factors, the CXCR5⁺ CD8⁺ T cells expressed *Bcl6*, *Tcf7* and *Plagl1* that are typically associated with CD4⁺ Tfh cells¹⁴ while CXCR5⁺ cells expressed *Prdm1* (Blimp-1) that is linked with CD4⁺ Th1 cells and effector CD8⁺ T cells highlighting the distinct transcriptional fates

¹Emory Vaccine Center and Department of Microbiology and Immunology, Emory University School of Medicine, Atlanta, Georgia, USA. ²Lymphocyte Biology Section, Laboratory of Systems Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA. ³Department of Urology, Emory University School of Medicine, Atlanta, Georgia, USA. ⁴School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil. ⁵Department of Microbiology, Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA. ⁶Interdisciplinary Immunology Graduate Program, Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA. ⁷Department of Pathology, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts, USA. ⁸Department of Medical Oncology, Dana-Farber Cancer Institute, Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA.

Induction of Cell Cycle and NK Cell Responses by Live-Attenuated Oral Vaccines against Typhoid Fever

Christoph J. Blohmke¹, Jennifer Hill¹, Thomas C. Darton^{1,2*}, Matheus Carvalho-Burger³, Andrew Eustace¹, Claire Jones¹, Fernanda Schreiber⁴, Martin R. Goodier⁵, Gordon Dougan⁴, Helder I. Nakaya³ and Andrew J. Pollard¹

¹Oxford Vaccine Group, Department of Paediatrics, University of Oxford, NIHR Oxford Biomedical Research Centre, Oxford, United Kingdom, ²Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield Medical School, Sheffield, United Kingdom, ³School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil, ⁴Microbial Pathogenesis Group, The Wellcome Trust Sanger Institute, Hinxton, United Kingdom, ⁵Faculty of Infectious and Tropical Diseases, Department of Immunology and Infection, London School of Hygiene & Tropical Medicine, London, United Kingdom

OPEN ACCESS

Edited by:

Fabio Bagnoli,
GlaxoSmithKline, Italy

Reviewed by:

Estrella Mariel Levy,
Consejo Nacional de Investigaciones
Científicas y Técnicas
(CONICET), Argentina
Annalisa Ciabattini,
University of Siena, Italy

*Correspondence:

Thomas C. Darton
thomas.darton@paediatrics.ox.ac.uk

Specialty section:

This article was submitted to
Vaccines and Molecular
Therapeutics,
a section of the journal
Frontiers in Immunology

Received: 01 August 2017

Accepted: 25 September 2017

Published: 12 October 2017

Citation:

Blohmke CJ, Hill J, Darton TC,
Carvalho-Burger M, Eustace A,
Jones C, Schreiber F, Goodier MR,
Dougan G, Nakaya HI and Pollard AJ
(2017) Induction of Cell Cycle and NK
Cell Responses by Live-Attenuated
Oral Vaccines against Typhoid Fever.
Front. Immunol. 8:1276.
doi: 10.3389/fimmu.2017.01276

The mechanisms by which oral, live-attenuated vaccines protect against typhoid fever are poorly understood. Here, we analyze transcriptional responses after vaccination with Ty21a or vaccine candidate, M01ZH09. Alterations in response profiles were related to vaccine-induced immune responses and subsequent outcome after wild-type *Salmonella* Typhi challenge. Despite broad genetic similarity, we detected differences in transcriptional responses to each vaccine. Seven days after M01ZH09 vaccination, marked cell cycle activation was identified and associated with humoral immunogenicity. By contrast, vaccination with Ty21a was associated with NK cell activity and validated in peripheral blood mononuclear cell stimulation assays confirming superior induction of an NK cell response. Moreover, transcriptional signatures of amino acid metabolism in Ty21a recipients were associated with protection against infection, including increased incubation time and decreased severity. Our data provide detailed insight into molecular immune responses to typhoid vaccines, which could aid the rational design of improved oral, live-attenuated vaccines against enteric pathogens.

Keywords: typhoid, typhoid vaccines, functional genomics, vaccine immunity, Ty21a, NK cell, cell cycle regulation

INTRODUCTION

Salmonella enterica serovar Typhi (*S. Typhi*) is the predominant cause of enteric fever, a non-specific febrile infection affecting between 9.1 and 17.8 million people globally each year (1–3). Transmission between humans occurs by ingestion of faecally contaminated food or water, after which *S. Typhi* invades the gut mucosa and may be taken up by phagocytic cells, before asymptomatic systemic dissemination to the reticuloendothelial system. Individuals presenting with typhoid fever develop non-specific symptoms including fever, abdominal pain and headache, and bacteremia ensues (4, 5). It remains unclear how innate immunity or the adaptive immune systems (following vaccination or prior infection) eradicates *S. Typhi* after infection. Detailed investigation of the molecular host responses to live-attenuated vaccines and interpretation in the context of responses seen after human exposure to virulent wild-type *S. Typhi* may provide useful insights informing future vaccine design.

Ty21a is an oral, live-attenuated typhoid vaccine that has been licensed in 1989 and administered to a large number of people (6), demonstrating a protective efficacy of 42–96% after three or four doses in clinical trials performed in typhoid-endemic regions (7–9). Ty21a has been attenuated