UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOQUÍMICA E IMUNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOQUÍMICA E IMUNOLOGIA

Izabela Mamede Costa Andrade da Conceição

# MODULAÇÃO DA EXPRESSÃO DE ISOFORMAS DE LNCRNA NA RESPOSTA AO TRATAMENTO POR METFORMINA EM DIFERENTES TIPOS CELULARES

Belo Horizonte

Junho 2022

**Izabela Mamede Costa Andrade da Conceição**

# MODULAÇÃO DA EXPRESSÃO DE ISOFORMAS DE LNCRNA NA RESPOSTA AO TRATAMENTO POR METFORMINA EM DIFERENTES TIPOS CELULARES

Orientador: Profa. Dra. Glória Regina Franco

Orientador: Prof. Dr. Marcelo Rizzatti Luizon

Dissertação submetida ao Departamento de Bioquímica e Imunologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Mestre em Bioquímica e Imunologia.

Belo Horizonte

Junho 2022

043 Conceição, Izabela Mamede Costa Andrade da.
Modulação da expressão de isoformas de lncRNA na resposta ao tratamento por metformina em diferentes tipos celulares [manuscrito] / Izabela Mamede Costa Andrade da Conceição. – 2022.
92 f. : il. ; 29,5 cm.

Orientador: Profa. Dra. Glória Regina Franco. Coorientador: Prof. Dr. Marcelo Rizzatti Luizon.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Departamento de Bioquímica e Imunologia.

1. Bioquímica e imunologia. 2. Isoformas de RNA. 3. RNA Longo não Codificante. 4. Metformina. I. Franco, Glória Regina. II. Luizon, Marcelo Rizzatti. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 577.1

Ficha catalografica elaborada pela Bibliotecária Fabiane C M Reis – CRB 6/ 2680

**Universidade Federal de Minas Gerais**
*Curso de Pós-Graduação em Bioquímica e Imunologia ICB/UFMG*
*Av. Antônio Carlos, 6627 – Pampulha*
*31270-901 – Belo Horizonte – MG*
*e-mail: pg-biq@icb.ufmg.br (31)3409-2615*

**ATA DA DEFESA DA DISSERTAÇÃO DE MESTRADO DE IZABELA MAMEDE COSTA ANDRADE DA CONCEIÇÃO. Aos vinte dias do mês de julho de 2022 às 09:00 horas, reuniu-se, de forma "online" utilizando a plataforma "Zoom", no Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, a Comissão Examinadora da dissertação de Mestrado, indicada *ad referendum* do Colegiado do Curso, para julgar, em exame final, o trabalho intitulado "Modulação da expressão de isoformas de lncRNA na resposta ao tratamento por metformina em diferentes tipos celulares", requisito final para a obtenção do grau de Mestre em Bioquímica e Imunologia, área de concentração: Bioquímica. Abrindo a sessão, a Presidente da Comissão, Profa. Glória Regina Franco, da Universidade Federal de Minas Gerais, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores, com a respectiva defesa da candidata. Logo após a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações: Dr. Erich Birelli Tahara (Universidade Federal de Minas Gerais), aprovada; Dr. Marcelo Rizzatti Luizon - Orientador (Universidade Federal de Minas Gerais), aprovada; Dr. Renato Santana de Aguiar (Universidade Federal de Minas Gerais), aprovada; Dr. Paulo de Paiva Rosa Amaral (INSPER - SP), aprovada; Dra. Glória Regina Franco - Orientadora (Universidade Federal de Minas Gerais), aprovada. Pelas indicações a candidata foi considerada:**

**( X ) APROVADA**
**( ) REPROVADA**

**O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente da Comissão encerrou a reunião e lavrou a presente Ata que será assinada por todos os membros participantes da Comissão Examinadora. Belo Horizonte, 20 de julho de 2022.**

**Dr. Erich Birelli Tahara (UFMG)**

**Dr. Marcelo Rizzatti Luizon - Orientador (UFMG)**

**Dr. Renato Santana de Aguiar (UFMG)**

**Dr. Paulo de Paiva Rosa Amaral (INSPER - SP)**

Glória Regina Franco

**Dra. Glória Regina Franco - Orientadora (UFMG)**

# Agradecimentos

Agradeço em primeiro lugar à todas as minorias na ciência que pavimentaram o caminho para que hoje eu possa estar aqui escrevendo essa dissertação.

Agradeço ao meu namorado Gabriel que eu amo e que foi o meu principal ajudante emocionalmente falando nos últimos anos, sem seu apoio e da sua família eu não estaria terminando isso hoje, obrigada mesmo do fundo do meu coração. Agradeço ao Alesch que salvou minha vida doze anos atrás e que sem ele eu teria desistido de estar nesse mundo muito antes de chegar no mestrado. Agradeço ao Rafa que mesmo distante e sem contato é a pessoa da minha família que sempre me apoia.

Agradeço a todos os membros do LGB, em especial dos GLORIOSOS, atuais e passados. Agradeço ao pessoal que está presente lá todo dia por hoje: a Dani, ao Wesley, aos Zoretes (Jéssica e Darlan), aos membros do meu exército sensacional de ICs otakus (André, Gabrielle e Bárbara), ao Mendes, à técnica do lab, Lorrane, e aos demais professores do LGB, Andréa e Carlos Renato. Agradeço também aos GLORIOSOS antigos em especial à Nayara, por toda a amizade, ao Thomaz, por ter me auxiliado demais no início desse trabalho, ao Lúcio por ter me ensinado todo o básico e ao Herón, Stella, Bruno, pela amizade. Agradeço também ao Tiago Bruno pelas oportunidades e ao Hemanoel pelos conselhos e dicas, vocês s 10! Agradeço aos membros do Luizon lab, em especial à Rahyssa, Ju e Dani, por toda a amizade ao longo desses anos.

Agradeço aos meus amigos pessoais: A Ju e ao Daniel, a Amanda e a Ana, ao Henrique e a Carol, meus maiores amigos da graduação. Agradeço as três pessoas que ainda estão do meu lado desde o ensino médio: Isa, Iza e Gi, e aos meus amigos da bioquímica, Nat e Carlos e a equipe toda do CV bioinfo.

Um agradecimento especial a meus orientadores. Ao Professor Marcelo por todo o auxílio e amizade desde o início da minha iniciação científica, você é demais sério. Agradeço à Professora Glória que é um grande exemplo do tipo de pesquisadora que eu quero ser, realmente você a rainha da transcriptômica e a pessoa mais inteligente que já conheci. Agradeço também a Dra. Natália que foi minha primeira orientadora de fato e foi responsável não só por eu não desistir do curso de Biologia como também por me mostrar que eu amava bioinformática. Agradeço à Mari, que é a amizade mais sensacional que eu já fiz na minha vida.

Agradeço à banca que aceitou participar dessa defesa e ler essa dissertação.

*“Never be a spectator of unfairness or stupidity.*
*The grave will supply plenty of time for silence.”*

*“Your least favorite virtue, or nominee for the most overrated one?*
*Faith.*
*Closely followed—in view of the overall shortage of time—by patience.”*

*“The gods that we've made are exactly the gods*
*you'd expect to be made by a species*
*that's about half a chromosome away*
*from being chimpanzee.”*

**Christopher Hitchens (1949-2011)**

*“Creativity comes from applying*
*things you learn in other fields*
*to the field you work in.”*

**Aaron Swartz (1986-2013)**

**Resumo**

RNAs longos não codificadores (lncRNAs) são com mais de 200 nucleotídeos que não codificam proteínas por métodos canônicos de tradução. Eles sofrem *splicing*, já que possuem introns e exons, e possuem múltiplas isoformas transcritas. Em média, lncRNAs possuem o dobro de isoformas quando comparados com genes codificadores de proteínas. Mesmo assim, tanto para lncRNAs quanto para genes codificadores de proteínas, a medição de expressão diferencial é feita rotineiramente somente em nível do gene. A análise em nível de gene não considera as múltiplas isoformas transcritas, tornando os resultados de expressão diferencial incompletos.

Metformina é a primeira linha de tratamento para Diabetes tipo II e para várias outras doenças, incluindo o câncer. Contudo, seu mecanismo de ação ainda não foi completamente elucidado no nível molecular, como suas ações como regulador epigenético, anti-proliferação e anti-envelhecimento. Análises globais de transcriptoma utilizando metformina em diferentes tipos celulares revelam que apenas genes codificadores de proteína são levados em consideração para avaliar os efeitos desse fármaco na modulação da expressão gênica.

Nosso objetivo foi caracterizar isoformas de lncRNAs diferencialmente afetadas pelo tratamento com metformina em diversos tipos celulares humanos e sugerir como esse fármaco regula a expressão de lncRNAs. Buscamos dados de todas as bibliotecas de alta profundidade, com sequenciamento *paired-end* disponíveis publicamente. Seis séries foram selecionadas para exploração posterior de forma a realizar uma análise estatisticamente comparável de expressão diferencial e, posteriormente, inferir possíveis papeis biológicos para as isoformas de lncRNAs, utilizando ferramentas *in sílico*.

Para atingirmos os nossos objetivos, realizamos o pseudo-alinhamento de sequências com o *Salmon* (v 1.5.0) e expressão diferencial com o pacote em R Fishpond. Correlação, integração com bancos de dados epigenômicos disponíveis e anotação de transcritos foram feitas utilizando pacotes em R e o enriquecimento funcional foi feito com o pacote *fgsea* e comparado entre todas as séries.

Pelo nosso conhecimento, esse é o primeiro estudo que analisa isoformas de lncRNA responsivas ao tratamento por metformina. Encontramos uma mesma isoforma de lncRNA (AC016831.6-205) superexpressa em todas as seis séries de tratamento com metformina, sendo o seu segundo exon putativamente codificador de um peptídeo com relevância para a ação da droga em câncer. Além disso, sugerimos que outras duas isoformas de lncRNA (ZBED5-AS1-

207 e AC125807.2-201) encontradas nesse estudo podem agir como elementos regulatórios em *cis* para transcritos em sua vizinhança.

Nossos resultados reforçam a importância de se levar em consideração isoformas diferencialmente expressas de lncRNAs para o entendimento do mecanismo de ação da metformina no nível molecular.

Palavras chave: RNAs longos não-codificantes, bioinformática, trancriptômica.

**Abstract**

Long non-coding RNAs (lncRNAs) are molecules with more than 200 nucleotides which do not code for proteins in canonical ways. They undergo *splicing*, since they possess introns and exons, and have multiple transcribed isoforms. On average lncRNAs have twice as many isoforms when compared to protein-coding genes. Nevertheless, for lncRNAs, as well as for mRNA, measurements of differential expression are routinely performed only at the gene level. Gene level analysis does not consider the multiple isoforms which arise from the transcription of a single gene, making the final differential expression results incomplete.

Metformin is the first-line oral therapy for type II Diabetes and for several other diseases, including cancer. However, its mechanism of action remains not thoroughly explained, with its multiple effects as epigenetic regulator, anti-proliferation and anti-aging drug not explored at the molecular level. Global transcriptomic analyses using metformin in different cell types reveals that only protein-coding genes are normally taken into consideration.

Our aim was to globally characterize lncRNA isoforms that were differentially affected by metformin treatment on multiple human cell types, and to provide insights into the lncRNA regulation by this drug. We selected data from all higher depth and paired-end stranded libraries publicly available. We selected six series for further exploration to perform a statistically comparable differential expression (DE) isoform analysis. We also inferred the biological roles for lncRNA DE isoforms using in silico tools.

Sequence pseudo-alignment was performed with Salmon (v.1.5.0) and differential expression using Fishpond. Correlation, integration with epigenomic available datasets and transcript annotation were performed using packages in R and functional enrichment was performed with fgsea and integrated between all series.

To our knowledge, this is the first study that globally analyzed lncRNA isoforms responsive to metformin treatment. We found the same isoform of a lncRNA (AC016831.6-205) highly expressed in all six-metformin series, which has a second exon putatively coding for a peptide with relevance to the drug action in cancer. Moreover, other two lncRNA isoforms (ZBED5-AS1-207 and AC125807.2-201) may also behave as cis-regulatory elements to the expression of transcripts in their vicinity.

Our results strongly reinforce the importance of taking into consideration DE isoforms of lncRNA for the understanding of metformin mechanisms at molecular level.

Keywords: Long non-coding RNAs, bioinformatics, transcriptomics.

## LISTA DE FIGURAS

**1.** Mecanismo de ação da metformina no Hepatócito.....21

**2.** Mecanismos propostos anti-envelhecimento da metformina.....22

**3.** Estrutura tridimensional do Spliceossomo.....24

**4.** Representação esquemática dos tipos de transcrito.....25

**5.** lncRNAs regulando a expressão gênica de diferentes formas.....28

**6.** Visão global do pipeline utilizado.....30

**7.** Contagem de isoformas de lncRNA diferencialmente expressas entre as séries.....38

**8.** Comparação entre os *log2FoldChanges* dos transcritos encontrados a depender do método de *bootstrap* utilizando análise de correlação.....40

**9.** Comparação entre os valores de TPM de duas execuções independentes do Salmon para as amostras de hepatócitos primários.....41

**10.** Gráfico de comparação entre o número total de correlações obtidas e os valores de corte utilizados.....42

**11.** Contagem de transcritos ao redor das regiões de cada um dos lncRNAs cujas isoformas foram selecionadas para análises posteriores.....43

**12.** Contagem de isoformas de lncRNA diferencialmente expressas nas séries.....44

**13.** Intercessão entre as isoformas diferencialmente expressas presentes nas seis séries representada em formato de upsetplots.....46

**14.** Heatmap das isoformas de lncRNA diferencialmente expressas em, ao menos, quatro séries com seus metadados.....47

**15.** Região genômica dos transcritos de NEAT1 do Ensembl Genome Browser.....50

**16.** Região genômica de alguns dos transcritos de LINC00511 do Ensembl Genome Browser.....51

**17.** Região genômica de alguns dos transcritos de GAS5 do Ensembl Genome Browser.....52

**18.** Região genômica do transcrito AL133243.2-201 no Ensembl Genome Browser.....53

**19.** Região genômica ao redor dos transcritos de AC016831.6-205 e sua sobreposição com LINC-PINT.....54

**20.** Contagem de pares de correlação significativos de acordo com os valores de corte por isoforma selecionada.....56

**21.** Número de transcritos presentes na região genômica de 1 Mb, ao redor de cada lncRNA analisado, ordenado por localização cromossomal.....57

**22.** Gráfico de barras do log2FC em cada série dos pares de isoforma de lncRNA-mRNA que possivelmente agem em *cis*.....58

**23.** Figura combinada do UCSC Genome Browser e do Ensembl Genome Browser da região ao redor do AC125807.2-201.....59

**24.** Figura combinada do UCSC Genome Browser e do Ensembl Genome Browser da região ao redor do ZBED-AS1-207.....60

**25.** Isoformas do gene FOXM1 com seus éxons e íntrons se sobrepondo na região genômica específica.....61

**26.** Algumas isoformas de transcrito do gene EIFG2 com seus éxons e íntrons se sobrepondo na região genômica específica.....62

**27.** Enriquecimento funcional de transcritos codificadores de proteínas alvos da isoforma de lncRNA AC016831.6-205.....63

**28.** Rede de interação dos transcritos alvo de AC016831.6-205 que também aparecem diferencialmente expressos em mais de quatro séries.....65

**29.** Enriquecimento funcional de transcritos codificadores de proteínas alvos da isoforma de lncRNA NEAT1-202.....66

**30.** Região genômica do lncRNA LINC-PINT (A) e circularização de seu segundo exon por *backsplicing* (B).....71

**31.** *Heatmap* de transcritos presentes em quatro ou mais bibliotecas incluindo todas as bibliotecas com número mínimo necessário de réplicas encontradas na literatura.....76

# LISTA DE TABELAS

**Tabela 1:** Descrição de todas as bibliotecas utilizadas no estudo.....36

**Tabela 2:** Séries de Metformina selecionadas para análises posteriores.....38

**Tabela 3:** Resultados da busca por revisões sistemáticas seguidas ou não de meta-análise de lncRNAs em vários bancos de dados.....68

# LISTA DE ABREVIATURAS

**2DD:** *Normal human dermal fibroblasts* (fibroblastos humanos normais)

**786-O:** 786-O *renal cell carcinoma tumor model* (células renais de carcinoma).

**A549:** *Adenocarcinomic human alveolar basal epithelial cells* (células epiteliais basais de adenocarcinoma humano)

**AMPK:** *Adenosine Monophosphate-activated Protein Kinase* (Proteina kinase adenosina monofosfato)

**-AS1:** Isoforma Antisenso

**ATF3:** *Activating Transcription Factor 3* (Fator de transcrição ativador 3)

**BIRC6:** *Baculoviral IAP Repeat Containing 6* (Repetição IAP contida em Baculovirus)

**CAL27:** *Oral squamous epithelial carcinoma cells* (Células de carcinoma escamoso epitelial oral)

**CALM2:** *Calmodulin 2* (Calmodulina 2)

**cAMP:** AMP cíclico

**CD5:** *Cluster of Differentiation 5* (Conjunto de diferenciação 5)

**ChIP-seq:** *Chromatin Immunoprecipitation Followed by Sequencing* (Imunoprecipitação de cromatina seguida de sequênciamento)

**CTCF:** *CCCTC-Binding Factor* (Fator de ligação de CCCTC)

**DM1:** Distrofia Miotônica Tipo I

**DM2:** Diabetes Mellitus tipo II

**DMG:** Diabetes Mellitus Gestacional

**DNA**: *Deoxyribonucleic acid* (ácido desoxirribonucleico)

**DOG:** *Downstream of Gene* (A jusante do gene)

**-DT:** *Divergent Transcript* (Transcrito divergente)

**DTE:** *Differential Transcript Expression* (Expressão diferencial de transcrito)

**E2F:** *E2 Transcription Factor* (Fator de transcrição E2)

**E2F3:** E2 *Transcription Factor 3* (Fator 3 de transcrição E2)

**E-BOX:** *Enhancer box* (Box acentuador)

**EGF:** *Epidermal Growth Factor* (Fator de crescimento epidermal)

**EGFR:** *Epidermal Growth Factor Receptor* (Receptor do fator de crescimento epidermal)

**ENCODE:** *Encyclopedia of DNA Elements*

**EIF1:** *Eukaryotic Translation Initiation Factor 1* (Fator eucariótico de início da tradução 1)

**eIF3d:** *Eukaryotic Translation Initiation Factor 3 Subunit D* (Fator eucariotico de início da tradução 3 subunidade D)

**EIF4G2:** *Eukaryotic Translation Initiation Factor 4 Gamma 2* (Fator eucariotico de início da tradução 4 gamma 2)

**EZH2:** *Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit* (Acentuador de Zeste 2 subunidade do complexo repressivo da polycomb 2)

**FDA:** *Food and Drug Administration*

**FGSEA:** *Fast Gene Set Enrichment Analysis* (Análise rápida de enriquecimento de grupos de genes)

**FOXM1:** *Forkhead box protein M1* (Proteína M1 de forkhead box)

**GAM:** *Genome Architecture Mapping* (Mapeamento de arquitetura do genoma)

**GAS5:** *Growth Arrest Specific 5* (Especifico da parada de crescimento 5)

**GPD2:** Glicerol-3-fosfato desidrogenase

**H1-HeSC:** *Human Embryonic Stem Cell line* (Linhagem de células embrionárias humanas)

**HCC:** *Hepatocellular Carcinoma* (Células de carcinoma Hepatico)

**HEAL:** *HIV-1-enhanced lncRNA* (LncRNA aprimorado de HIV-1)

**HEPG2**: *Human liver cancer cells* (Celulas humanas de câncer de fígado)

**HeSC**: *Human embryonic stem cells* (Celulas-tronco embrionárias humanas)

**HFFc6:** *Human Foreskin Fibroblasts* (Fibroblasto humano epitelia)

**HI-C:** *High throughput chromossome conformation capture* (Captura conformacional de cromossomos de alto desempenho)

**HIF1A:** *Hypoxia Inducible Factor 1 subunit Alpha* (Fator 1 de indução de subxia subunidade alfa)

**HIF2A:** *Hypoxia Inducible Factor 2 subunit Alpha* (Fator 2 de indução de subxia subunidade alfa)

**HMBS:** *Hydroxymethylbilane Synthase* (Sintase de hidroximetilbilano)

**HSC2**: *Oral squamous carcinoma cells* (Celulas de carcinoma escamoso oral)

**HULC:** *Highly Up-regulated in Liver Cancer* (Altamente superexpresso em câncer de fígado)

**HUVEC:** *Human umbilical vascular endothelial cells* (Célula endotelial vascular umbilical humana)

**IRES:** *Internal Ribossome Entry Site* (Sitio de entrada interna do ribossomo)

**LILACS:** *Latin American and Caribbean Health Sciences*

**LINC00511:** *Long Intergenic Non-Protein Coding RNA 511* (RNA longo intergênico não codificador de proteínas 511)

**LINC-PINT:** *Long Intergenic Non-Protein Coding RNA, P53 Induced Transcript* (RNA longo intergênico não codificador de proteínas, induzido por P53)

**lncRNA:** *Long noncoding RNA* (RNA longo não codificador)

**MALAT1:** *Metastasis Associated Lung Adenocarcinoma* (Adenocarcinoma de pulmão associado à metástase)

**MAP4K3:** *Mitogen-Activated Protein 4 Kinase 3* (Proteína kinase 3 ativada pelo mitógeno 4)

**MAP3K20:** *Mitogen-Activated Protein 3 Kinase 20* (Proteína kinase 20 ativada pelo mitógeno 3)

**MCM3AP:** *Minichromosome Maintenance Complex Component 3 Associated Protein* (Proteina associada ao complex de manutenção de minicromossomos)

**MeSH:** *Medical Subject Headings*

**miRNA:** *microRNAs*

**mRNA:** RNA mensageiro

**MSigDB:** Molecular Signature Database

**MTOR:** *Mammalian Target of Rapamycin* (Alvo da rapamicina em mamíferos)

**MYCN:** *N MYC proto-oncogene* (Proto-oncogene N-MYC)

**NEAT1:** *Nuclear Paraspeckle Assembly Transcript 1* (Transcrito de montage das *paraspeckles* nucleares 1)

**NES:** *Normalized Enrichment Score* (Pontuação normalizada de enriquecimento)

**NIFK:** *Nucleolar Protein Interacting with The FHA Domain Of MKI67* (Proteína nucleolar interagindo com o domínio FHA do MKI67)

**NRG1:** Neuregulina 1

**OCT1:** *Organic Cation Transporter 1* (Transportador orgânico de cátions 1)

**ORFs:** *Open Reading Frames* (Regiões abertas de leitura)

**PAAN:** *PA-associated noncoding RNA* (RNA não codificante associado a PA)

**PANC-1:** *Pancreatic Duct Epithelioid Cells* (Células epiteliais de ducto pancreático)

**PBMC:** *Peripheral blood mononuclear cells* (Células periféricas sanguíneas de núcleo único)

**PCOS:** *Polycystic ovary syndrome* (Síndrome do Ovário Policístico)

**PDB:** *Protein Database*

**PROSPERO:** *Prospectively registered systematic reviews*

**PSMG1:** *Proteasome Assembly Chaperone 1* (Chaperona 1 de montagem do proteassoma)

**RAB11B:** *Ras-related Protein B* (Proteína relacionada ao RAS B)

**ReN:** *Neural progenitor cells* (Células progrnitoras neuronais)

**RNA:** *Ribonucleic Acid* (Ácido ribonucleico)

**RNA-seq:** *Sequenciamento de RNA*

**SERPINE1:** *Serpin Family E Member 1* (Membro 1 da família das Serpinas)

**SLC22A1:** *Solute Carrier Family 22 (Organic Cation Transporter), Member 1* (Membro 1 da família dos carreadores de soluto)

**SNHG15:** *Small Nucleolar RNA Host Gene 15* (Gene hospedeiro de RNAs nucleolares 15)

**snoRNAs:** *Small Nucleolar RNAs* (RNAs pequenos nucleolares)

**snRNAs:** *Small Nuclear RNAs* (RNAs pequenos nucleares)

**snRNPs:** *Small Nuclear Ribonucleoprotein* (Ribonucleoproteínas pequenas nucleares)

**sORFs:** *small ORFs* (pequenas ORFs)

**SOX4:** *SRY-Box Transcription Factor 4* (Fator de transcrição SRY-box)

**SRA:** *Sequence Read Archive*

**TADs:** *Topologically Associated Domains* (Domínios associados topologicamente)

**TAP-RNAs:** *Topological Anchor Point RNAs* (RNAs de ponto de ancoragem de domínios)

**TNF:** *Tumor Necrosis Factor* (Fator de necrose tumoral)

**THAP9:** *THAP domain containing 9* (Contenedor do domínio THAP 9)

**TPM:** *Transcripts Per Million* (Transcritos por Milhão)

**UCSC:** *University of California, Santa Cruz*

**uORFs:** *upstream ORFs* (ORFs a montante)

**VEGFA:** *Vascular Endothelial Growth Factor A* (Fator de crescimento endotelial vascular A)

**ZBED5:** *Zinc finger BED domain containing Protein 5* (Proteína contenedora de domínio SINC finger BED 3)

**ZBTB11:** *Zinc Finger and BTB Domain Containing 11* (Proteína contenedora de domínio zinc finger e BTB 11)

**ZNF213**: *Zinc Finger Factor 213* (Fator de zinc finger 213)

# SUMÁRIO

**1. Introdução.....20**
1.1 Metformina.....20
1.2 O *Splicing* e a variabilidade do transcriptoma e proteoma além das ORFs.....23
1.3 RNAs longos não-codificadores.....24
**2. Objetivos.....29**
2.1 Objetivo geral.....29
2.2 Objetivos específicos.....29
**3. Métodos.....30**
3.1 Seleção de bibliotecas públicas.....30
3.2 Alinhamento e quantificação de transcritos.....31
3.3 Análise de expressão diferencial de transcritos.....32
3.4 Correlação da expressão de transcritos.....32
3.5 Anotação funcional e filtragem de regiões genômicas.....33
3.6 Enriquecimento funcional de transcritos.....34
**4. Resultados.....36**
4.1 Bibliotecas selecionadas.....36
4.2 Otimização do pipeline.....39
4.3 Transcritos diferencialmente expressos.....44
4.4 Intercessão entre as séries.....45
4.5 Isoformas com possível ação em *cis*.....54
4.6 Isoformas com possível ação em *trans*.....62
**5. Discussão.....68**
**6. Conclusão.....79**

**Referências.....80**

**Apêndice A.....91**
Email resposta da equipe Gencode a respeito do lncRNA AC016831.6-205

# 1. Introdução

## 1.1 Metformina

A metformina (dimetilbiguanida) é o fármaco mais prescrito no mundo como primeira escolha para o tratamento de Diabetes tipo 2 (DM2), fazendo parte da lista de medicamentos essenciais da Organização Mundial da Saúde(BAILEY, 2017). Em 1994, a metformina foi aprovada pela *Food and Drug Administration* (FDA), sendo introduzida no mercado estadunidense no ano seguinte (BAILEY, 2017). Novos benefícios de tal droga foram relatados: a terapia medicamentosa "precoce" com metformina pode reduzir a mortalidade por eventos cardiovasculares e elevar a sobrevida de pacientes obesos com DM2, apresenta baixos riscos de causar subglicemia, além de não causar ganho de peso (BAILEY, 2017)e pode ser associada com moderada perda de peso devido a seus efeitos colaterais gastrointestinais (PENZIAS *et al.*, 2017).

O uso da metformina provoca a diminuição dos níveis sanguíneos de glicose pela inibição da gliconeogênese hepática. Além disso, o fármaco inibe a glicogenólise, reduz a resistência à insulina nos tecidos hepático e muscular, diminui a absorção gastrointestinal de glicose, facilita seu transporte nos tecidos periféricos, e melhora indiretamente a resposta de células beta à glicose (RENA; HARDIE; PEARSON, 2017).

A metformina ainda é utilizada no tratamento de diversas doenças além da DM2, incluindo a Síndrome do Ovário Policístico (PCOS, Polycystic ovary syndrome) e a Diabetes Mellitus Gestacional (DMG) (TODD; FLOREZ, 2014), além de ser um dos fármacos mais sugeridos hoje em dia para o tratamento de outras doenças cardiovasculares e cânceres (MA *et al.*, 2020) .

Importante notar que existe variabilidade de resposta ao tratamento com o fármaco, sendo que mais de 30% dos pacientes que recebem a monoterapia com metformina são classificados como não responsivos (COOK *et al.*, 2007) . Neste contexto, a variação genética desempenha um papel importante na resposta à metformina, com polimorfismos comuns explicando entre 21 e 34% da variabilidade de resposta do paciente (ZHOU *et al.*, 2014).

Molecularmente, o fármaco pode atuar tanto de maneira dependente da proteína quinase ativada por AMP (AMPK) quanto independentemente dela (RENA; HARDIE; PEARSON, 2017) e seu mecanismo de ação mais conhecido passa pela ativação do complexo 1 da cadeia de transporte de elétrons da mitocôndria causando a ativação da AMPK, redução do AMP

cíclico, promoção de vias glicolíticas, inibição de vias da gliconeogênese **(Figura 1)**. Outro mecanismo de ação proposto e que tem ganhado força nos últimos anos é de que o efeito da metformina na Glicerol-3-fosfato desidrogenase (GPD2) poderia ser o potencializador dos efeitos da diminuição de glicose pela droga, por alterações causadas no estado redox do fígado (MADIRAJU *et al.*, 2018). Recentemente foi sugerido que essa inibição indireta da GPD2 é causada por efeitos do fármaco no complexo IV da cadeia de transporte de elétrons mitocondrial e *in vivo* essa seria a causa mais provável da redução de gliconeogênese provocada pelo fármaco (LAMOIA *et al.*, 2023).

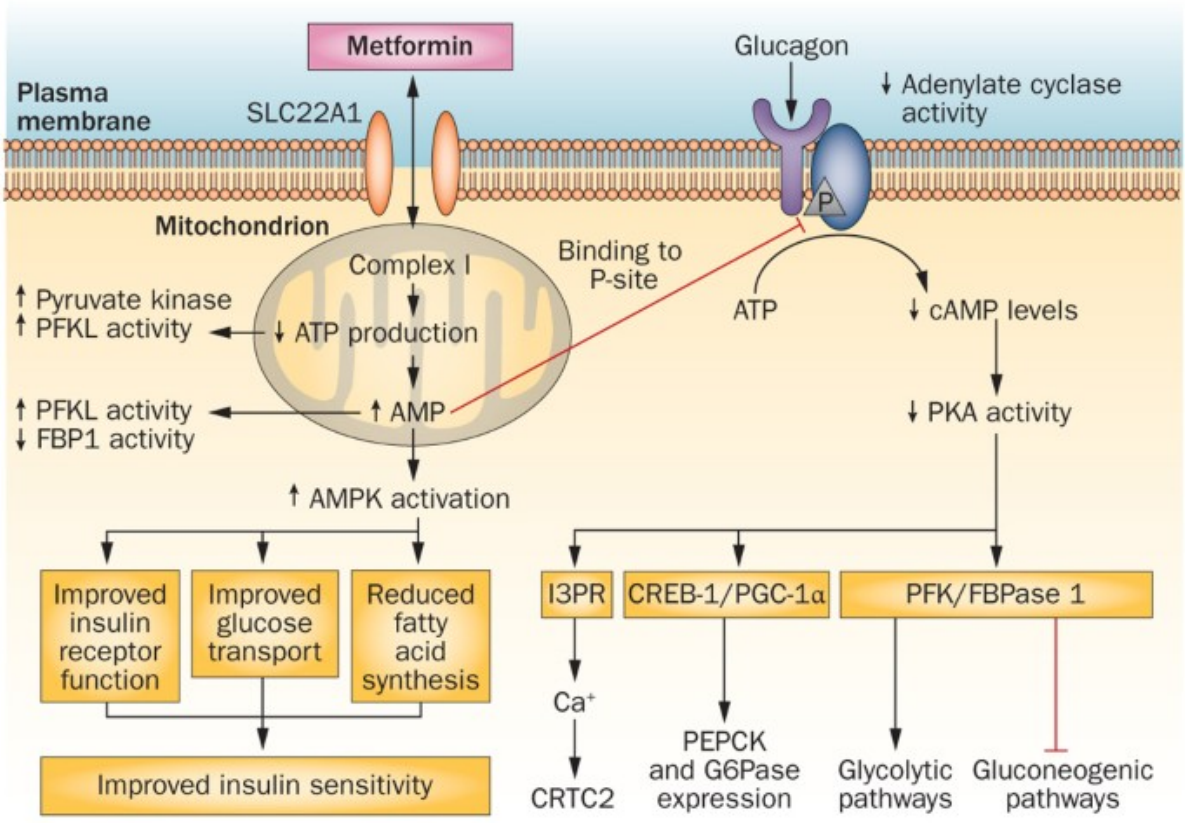


***Figura 1:*** *Modo de ação da metformina no hepatócito. Metformina entra na célula pelo receptor SLC22A1, também conhecido como OCT1, onde age no complexo I da mitocôndria, ativando o AMPK, que diminui os níveis circulantes de AMP cíclico (cAMP). Fonte: Pernicova (2017).*

Os possíveis efeitos positivos da metformina na prevenção e tratamento de diversos tipos de cânceres são ainda mais discutidos por diversas revisões nos últimos anos. A metformina influenciaria a gênese de tumores tanto diretamente, pela indução de estresse energético, quanto indiretamente, pela redução sistêmica de níveis de insulina e de vias imuno-

metabólicas associadas à progressão e morte destes tumores (HECKMAN-STODDARD et al., 2017; MA et al., 2020; PERNICOVA; KORBONITS, 2014).

Estudos clínicos em fase II ou III que obtiveram resultados positivos da ação do fármaco em tumores estão em andamento (ClinicalTrial.govNCT01864096, ClinicalTrial.gov NCT01101438). No entanto, a maior parte da literatura disponível atualmente na área apresenta benefícios apenas em pacientes já diabéticos ou não explora mais detalhadamente as possíveis causas moleculares dos efeitos do fármaco.

Outros estudos tem proposto a ação da metformina como regulador epigenético, que poderia ser a causa tanto de seus efeitos benéficos antitumorais quanto de um possível efeito como agente antienvelhecimento (CUYÀS *et al.*, 2016; ROMERO *et al.*, 2017; TRIGGLE *et al.*, 2022). Esse efeito do fármaco afetaria a epigenética da célula a partir do controle da disponibilidade de certos nutrientes essenciais para as enzimas modificadoras da cromatina, como as histonas acetiltransferases e metiltransferases, que utilizam metabolitos como NAD+, acetyl-CoA e alfa-cetoglutarato, todos tendo suas concentrações afetadas pelos efeitos do fármaco (CUYÀS *et al.*, 2016; MENENDEZ, 2020)**(Figura 2)**.

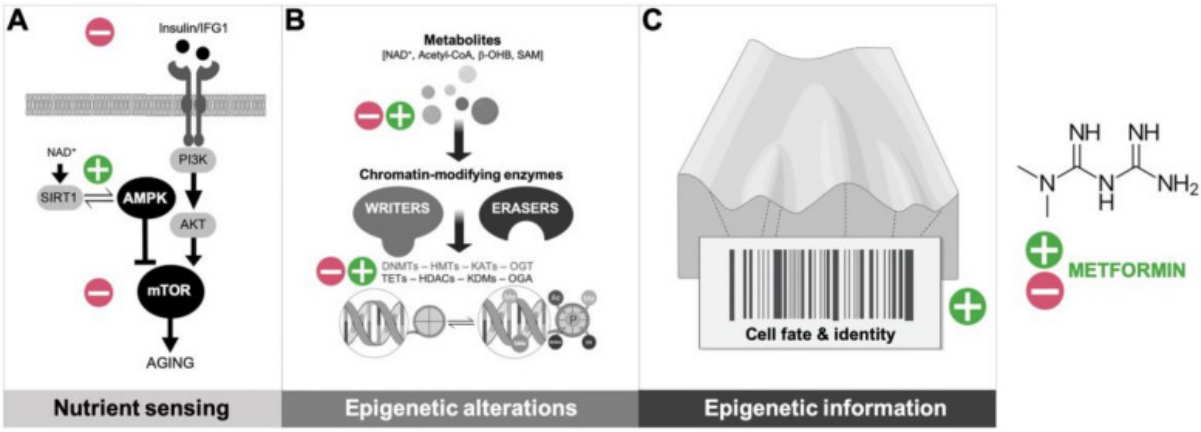


***Figura 2:*** *Mecanismos anti-envelhecimento da metformina: Em **A** é possível ver o método de ação canônico por ativação da AMPK que por sua inibição de MTOR influência no envelhecimento celular. Em **B** sua ação sugerida no controle de metabolitos necessários como substratos e cofatores para certas enzimas epigenéticas. Em **C** a mostra da paisagem epigenética sugerida por Waddington com a metformina agindo como preservador do epigenoma global. Fonte: Menendez (2020).*

## 1.2 O *Splicing* e a variabilidade do transcriptoma e do proteoma

O genoma humano possui menos de 1,5% de sua totalidade caracterizado como genes codificadores de proteína (BEJERANO *et al.*, 2004) . Essa caracterização tem como base a presença ou ausência de janelas abertas de leitura (ORFs, *Open Reading Frames*). As ORFs são genericamente definidas em uma sequência de ácidos nucleicos como regiões entre o códon de início e o códon de término da tradução, sendo essa região transcrita e traduzida pelos modos canônicos de tradução (SIEBER; PLATZER; SCHUSTER, 2018). No entanto, o processo entre a transcrição e a síntese da proteína é muito mais complexo, havendo o processamento do RNA, assim como uma possível tradução de regiões fora da ORF verdadeira pelo ribossomo ou por outros complexos proteicos (KWAN; THOMPSON, 2019; YUANYUAN; XINQIANG, 2022).

O processamento do RNA em eucariotos começa pela remoção dos introns e junção dos exons no processo de *splicing*, um processo ultraconservado desde genomas ancestrais (BEJERANO *et al.*, 2004) . Atualmente já sabemos que a relação direta de um gene codificador de proteína gerando um RNA mensageiro que será traduzido em um único tipo de proteína é falsa (ULE; BLENCOWE, 2019) . No processo de *splicing* são gerados vários transcritos, tanto aqueles codificadores de proteína com ligeiras diferenças em relação à proteína canônica, usualmente chamados de produtivos, quanto os não-codificadores de proteínas, ou não-produtivos; o que chamamos, respectivamente, de *splicing* canônico e *splicing* alternativo. Essa definição de produtivo e não produtivo também é feita pela presença ou ausência de ORFs nas anotações do transcriptoma (HARROW *et al.*, 2012).

Em mais detalhes, o processo de *splicing* e de *splicing* alternativo necessitam do spliceossomo, um complexo ribonucleoproteico que catalisa as reações de *splicing*. Neste complexo estão incluídos uma diversa gama de fatores, tais como snRNPs (*small nuclear ribonucleoprotein*) e snRNAs (*small nuclear RNAs*); U1, U2, U4, U6 e U5, porções do spliceossomo que executam o processo de *splicing* e mais de 150 proteínas adicionais (ULE; BLENCOWE, 2019; WANG, E.; AIFANTIS, 2020). Todo o processo de *splicing* já foi muito bem elucidado por diversas imagens de criomicroscopia eletrônica, que estão disponíveis no *Protein Data Bank* (PDB) pelos códigos: 5NRL, 5O9Z, 5XJC, 5YZG, 6FF4, 5MQF, 5UZ5, 5GM6, 5MPS, 4PJO, 6G90 e 5JL5 (KASTNER *et al.*, 2019; PLASCHKA; NEWMAN; NAGAI, 2019) (**Figura 3**).

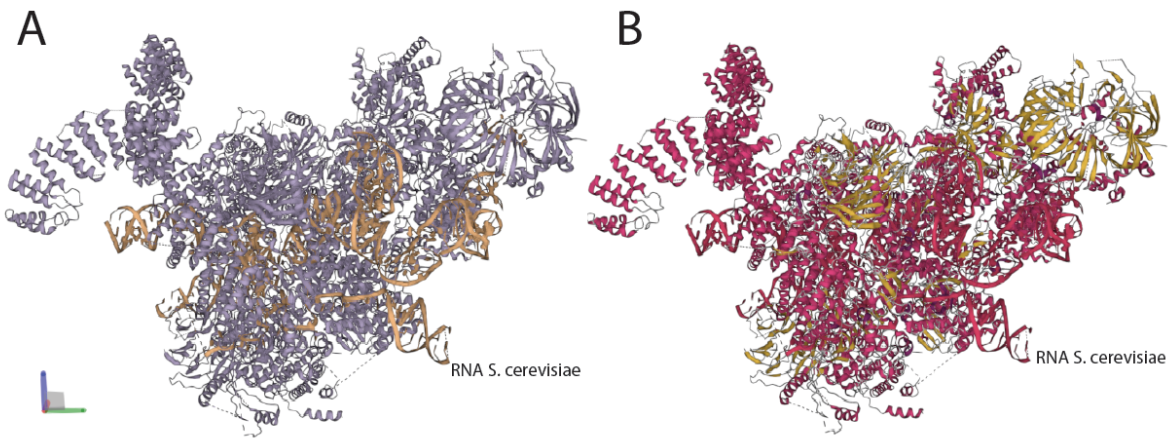


***Figura 3**: Spliceossomo remodelado para ligação de exons em transcritos de S. cerevisiae - PDB 5MPS. Em **A** os tipos de estrutura estão coloridos, em cinza o Spliceossomo e em laranja o RNA sendo processado e em **B** é mostrada a composição de estruturas secundárias, se alfa-hélice ou folha-beta.*

Quando pretende-se analisar a expressão gênica a partir do transcriptoma de células, tecidos, ou organismos, pode-se realizar a montagem *de novo* deste transcriptoma, ou utilizar anotações prévias em nível dos transcritos ou em nível dos genes. A análise do transcriptoma baseada na anotação somente do gene não incluirá toda a diversidade de transcritos produtivos e não produtivos produzidos a partir da transcrição desse gene. Por outro lado a anotação baseada nos transcritos é muito mais informativa (LOVE *et al.*, 2020).

Na anotação de transcritos, aqueles que não contêm uma ORF completa são classificados como transcritos processados (*processed transcript*) (FRANKISH *et al.*, 2019) . Dentro da categoria dos transcritos processados se encontra a subcategoria de intron-retido (*retained-intron*), outra categoria comum entre os não produtivos é a de e decaimento mediado pela presença de códons de parada precoces (*nonsense-mediated decay* ou NMD)(FRANKISH *et al.*, 2019). Cada uma dessas subcategorias tem uma definição própria, sendo que intron-retido e NMD são categorias de transcritos que sofreram *splicing* alternativo, só que a primeira é composta de transcritos que  contém sequências intronicas se comparados às variantes de transcritos codificadores derivados do gene e a segunda categoria é composta de transcritos que possuem um intron retido que carrega um códon de parada traducional a cerca de 50 nucleotídeos de um sítio de *splicing* à jusante (FRANKISH *et al.*, 2019)(**Figura 4**). As demais categorias de transcritos não serão abordadas nessa dissertação.

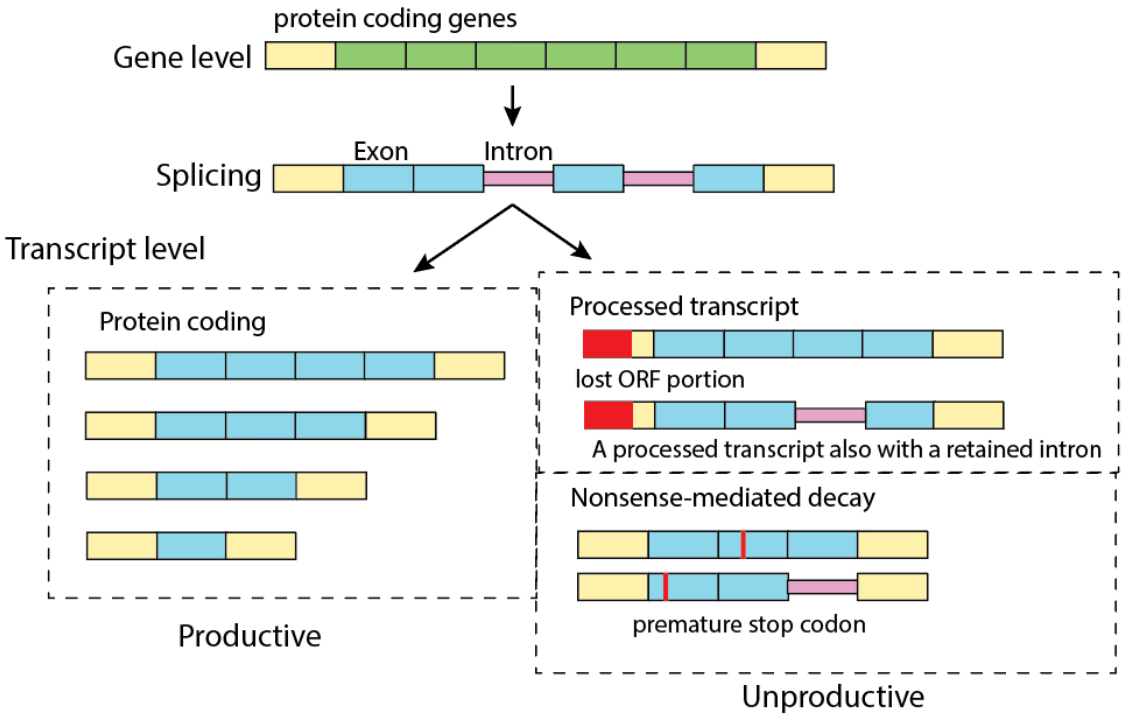


***Figura 4**: Representação esquemática dos tipos de transcritos gerados a partir de splicing. O processo de splicing e splicing alternativo dão origem à diversos tipos de isoformas de transcritos, produtivas e não-produtivas. As isoformas produtivas são todas codificadoras de proteína. Já as isoformas não produtivas são aquelas que retem introns, NMD e transcritos sem ORF.*

O *splicing* alternativo é altamente conservado desde eucariotos basais (ZEA *et al.*, 2021) , sendo um controlador fino da expressão gênica e funciona como regulador do nível produção proteica (BEJERANO *et al.*, 2004) e um dos maiores responsáveis pela variabilidade de RNAs e proteínas (ULE; BLENCOWE, 2019). Em algumas formas de controle da expressão gênica, a célula não interrompe a transcrição da região, apenas modifica o processamento a posteriori do RNA para que transcritos codificadores de proteína não sejam gerados ou não cheguem ao ribossomo para a tradução(ULE; BLENCOWE, 2019).

A presença ou ausência de ORF completa em um transcrito não determina, efetivamente, se um peptídeo pode ser produzido a partir desse transcrito por tradução não-canônica (KWAN; THOMPSON, 2019). A tradução não canônica é muito utilizada por mRNAs virais para explorar a maquinaria do hospedeiro e é melhor descrita nestes casos (CÁCERES *et al.*, 2018; MIRAS *et al.*, 2017; SOROKIN *et al.*, 2021). Entretanto, a tradução não canônica também é um processo constitutivo utilizado pela célula, especialmente em casos de estresse celular, sendo sugerida como um dos motivos para as grandes diferenças na

expressão gênica encontradas entre análises de transcriptoma e proteoma de uma mesma amostra (KWAN; THOMPSON, 2019). Especificamente, células de tumor são conhecidas por explorar métodos de tradução alternativos(SRIRAM; BOHLEN; TELEMAN, 2018; WALTERS; THOMPSON, 2016). Todavia, muito pouco dos mecanismos de tradução alternativa já foi explorado em outras condições.

Existem três mecanismos mais comuns de tradução alternativa: o da tradução por m6a (6-metil adenosina), o da tradução por IRES (*internal ribosome entry site*) e os mecanismos de tradução independente de ribossomos. Os mecanismos independentes de ribossomos normalmente ocorrem por ligações de complexos proteicos específicos que podem ter função de tradução, como o eIF3d (LEE, A. S. Y. *et al.*, 2016). Já o mecanismo de IRES ocorre por recrutamento da subunidade 40S do ribossomo para certas regiões do RNA, levando a uma tradução não dependente do processamento completo de mRNA e uma entrada do ribossomo independente da presença do CAP 5' (GEBAUER; HENTZE, 2016; MARTINEZ-SALAS *et al.*, 2018) . Já a tradução utilizando m6a foi descoberta mais recentemente e ocorre quando mRNAs não capeados podem ser traduzidos devido à presença de m6a na sua extremidade 5' (SRIRAM; BOHLEN; TELEMAN, 2018). Por meio desses processos de tradução não canônicos, muitos mRNAs não considerados produtores de proteínas podem traduzir, normalmente, pequenos peptídeos (XING, J. *et al.*, 2021).

### 1.3 RNAs longos não-codificadores

Os estudos de sequenciamento de nova geração, no início dos anos 2000, revelaram a presença de milhares de RNAs longos não-codificadores (lncRNA) de diversos tipos, tais como intronicos, intergênicos e antisenso(JARROUX; MORILLON; PINSKAYA, 2017; STARK; GRZELAK; HADFIELD, 2019). O lncRNA se refere a qualquer RNA caracterizado como não-codificador de proteína que não é produto de *splicing* alternativo de um transcrito proveniente de um gene codificador de proteínas (FRANKISH *et al.*, 2019). A diversidade de lncRNAs ainda não é completamente conhecida devido a muitos destes RNAs serem expressos de modo condição específico e ao problema do tamanho das leituras de sequenciamento versus a profundidade do sequenciamento de RNA. Métodos recentes de sequenciamento de leituras longas auxiliam na identificação de novos lncRNAs (AFFYMETRIX; COLD SPRING HARBOR LABORATORY ENCODE TRANSCRIPTOME PROJECT, 2009) e suas isoformas. No entanto, o sequenciamento de leituras longas ainda não possui profundidade para as análises de expressão de lncRNAs entre diferentes condições, como ocorre com o

sequenciamento de leituras curtas e, portanto, ainda são pouco utilizados para caracterizar expressão diferencial (MANTERE; KERSTEN; HOISCHEN, 2019).

Os lncRNAs podem atuar nos mais variados processos celulares e possíveis novas ações de lncRNA são descobertas a cada ano, sendo que eles podem agir em processos pré-transcricionais, controlando a abertura e fechamento de cromatina ao redor de certos genes (LOGSDON; VOLLGER; EICHLER, 2020), transcricionais, servindo de arcabouço para facilitar a ligação de fatores de transcrição na região promotora de genes, processos co-transcricionais regulando *splicing* canônico e *splicing* alternativo (ENGREITZ *et al.*, 2016), processos pós-transcricionais, controlando a meia vida e a taxa traducional de mRNAs no citoplasma e até mesmo codificando micropeptídeos (RAN *et al.*, 2022; TAN *et al.*, 2021) (**Figura 5**). Estes peptídeos codificados por lncRNA, inclusive, tem funções biológicas preditas em diversas doenças, inclusive o câncer (LI, Z.-X. *et al.*, 2018).

lncRNAs também sofrem *splicing* e *splicing* alternativo, já que possuem introns e exons. No entanto, produzem apenas transcritos canonicamente definidos como não produtivos (STANĚK, 2021).

A modulação que fármacos realizam na expressão de lncRNAs tem sido explorada atualmente, entretanto, a maioria dos estudos são focados em lncRNAs específicos (WEI *et al.*, 2020). Além do mais, estudos que abordam isoformas de lncRNA são raríssimos e normalmente são feitos somente em lncRNAs muito explorados na literatura, como o NEAT1 (ADRIAENS *et al.*, 2019), e estes usualmente abordam diferenças entre apenas poucos tipos de isoformas, sendo que lncRNAs possuem em média muito mais isoformas do que genes codificadores de proteínas.

Pela nossa revisão da literatura, este é o primeiro trabalho que aborda a caracterização de isoformas de lncRNA diferencialmente expressas após o tratamento com um fármaco a partir de dados de transcriptoma global e de forma sistemática, utilizando o mesmo pipeline em diversos experimentos independentes.

Dessa forma, a principal hipótese deste trabalho é de que a metformina regula isoformas de lncRNA que são centrais e essenciais para os efeitos moleculares conhecidos do fármaco.

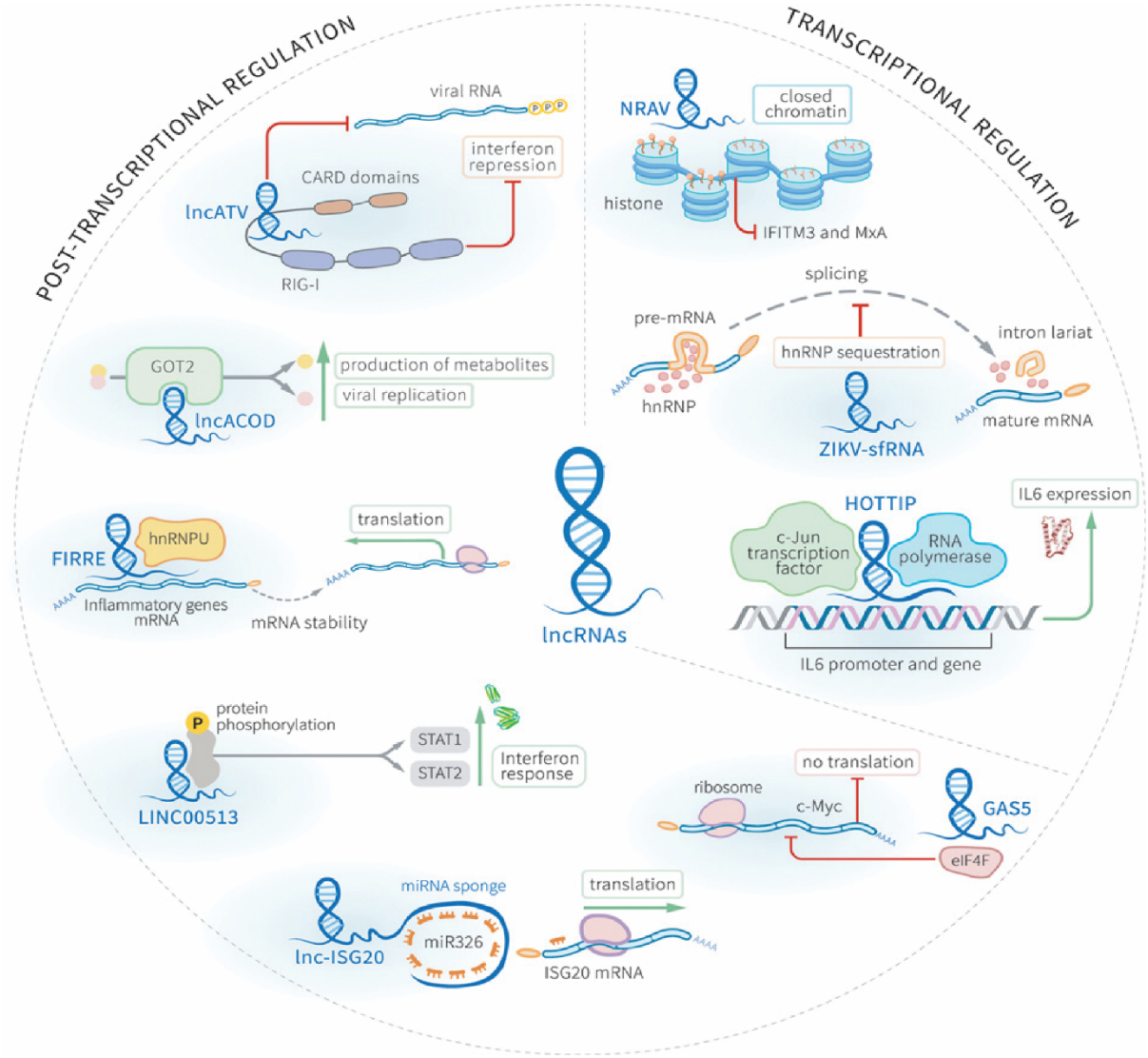


***Figura 5:*** *lncRNAs regulam a expressão gênica de diversas maneiras. Eles podem atuar se ligando à DNA, RNA e proteínas. Na figura estão presentes tanto as possibilidades de lncRNA agindo na regulação transcricional quanto na pós-transcricional. Os lncRNAs exemplo usados são lnc-ISG20, LINC00513, GAS5, FIRRE, lncACOD, lncATV, NRAV, ZIKV-sfRNA e HOTTIP Fonte: Luscher-Dias (2021).*

# 2.Objetivos

## 2.1 Objetivo Geral

Identificar isoformas de lncRNAs diferencialmente expressos em diversos tipos celulares tratados com metformina e inferir a função biológica dessas isoformas por meio de análises *in silico*.

## 2.2 Objetivos específicos

- Explorar bancos de dados de forma sistemática buscando experimentos de RNA-Seq que apresentem diferentes tipos celulares humanos tratados somente com metformina e grupos controles pareados;
- Reanálise dos experimentos de RNA-Seq com mapeamento no transcriptoma referência e análise de expressão diferencial em nível de isoforma dos transcritos;
- Comparação das isoformas de transcritos resultantes das análises e sugestão de mecanismos de ação biológicos dessas isoformas utilizando métodos *in silico*.

## 3. Métodos

Uma visão global de todo o pipeline utilizado, desde a seleção de bibliotecas até a anotação funcional dos lncRNA pode pode ser vista na **Figura 6**.

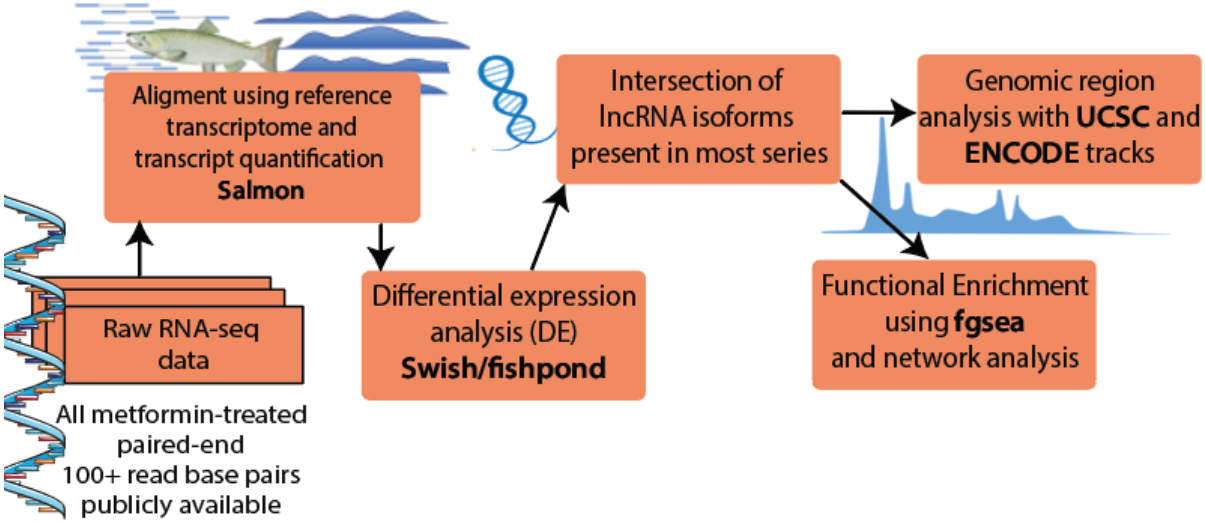


***Figura 6:*** *Visão global do pipeline utilizado desde a seleção de bibliotecas até a anotação funcional das isoformas de lncRNA. Começando a partir da seleção das bibliotecas, passando pelo pseudo-alinhamento no transcriptoma referência, pela expressão diferencial e intercessão entre as séries, até a anotação funcional das isoformas.*

### 3.1 Seleção de bibliotecas públicas

As bibliotecas foram selecionadas utilizando um modo sistemático de busca no *Sequence Read Archive* (SRA), o banco de dados de sequências em formato bruto com suas informações de metadados. O SRA é o maior e mais usado banco de resultados de sequenciamento disponíveis e atualmente é exigido que os resultados brutos de qualquer tipo de sequenciamento sejam depositados no SRA para publicação na maioria das revistas indexadas.

A busca no SRA foi feita utilizando os termos MeSH (*Medical Subject Headings*) associados a metformin, que são utilizados como uma forma reprodutível de se catalogar e buscar por informações relacionadas a assuntos integrados no Pubmed. A busca foi realizada por todas os experimentos que possuem dados brutos de sequenciamento de RNA (RNA-Seq) disponíveis do mais antigo até novembro de 2021, o que resultou em 14 experimentos de humanos, células ou pacientes, tratados com metformina com controles pareados que estão na Tabela 1 na sessão de resultados.

As bibliotecas que foram obtidas na busca possuíam características muito diferentes, tais como tamanho de leitura, concentração de metformina, número de pacientes, profundidade e tecnologia de sequenciamento utilizada. Todas elas foram analisadas individualmente pelo nosso pipeline para seleção posterior de resultados comparáveis.

Para todos os 14 experimentos foi checada a qualidade das bibliotecas, realizada a poda (*trimming*) (quando necessário) e realizado o pseudo-alinhamento no transcriptoma humano referência. Bibliotecas *single-end* e aquelas que possuíam menos de três replicatas biológicas de cada condição experimental foram excluídas antes da análise de expressão diferencial de transcritos (DTE, *differential transcript expression*). Bibliotecas *single-end* foram excluídas pois estas permitem um número muito menor de detecções, especialmente quando o objetivo são RNAs não produtivos, devido a nestas o sequenciamento das leituras ter sido feito apenas em uma direção. Bibliotecas sem o número mínimo de replicatas foram excluídas, pois um número inferior a três replicatas prejudica a maioria das análises posteriores e diminui a precisão estatística do cálculo de expressão diferencial. Após essa primeira seleção foi realizada a análise de DTE.

### 3.2 Alinhamento e quantificação de transcritos

A quantificação a nível de transcritos foi feita utilizando-se as leituras brutas de RNA-Seq e o software de pseudo-alinhamento Salmon (PATRO *et al.*, 2017) no transcriptoma humano referência do GENCODE versão 37(FRANKISH *et al.*, 2019). O Salmon foi rodado múltiplas vezes para comparação de resultados e estabelecimento de melhores ajustamentos.

Vários ajustes foram feitos com base em múltiplas execuções do Salmon para chagarmos na melhor forma possível do pseudo alinhamento das bibliotecas e estes ajustes podem ser encontrados na sessão dos resultados associada à otimização do pipeline.

O índice utilizado no programa para o pseudo-alinhamento foi executado com k_mer de tamanho 31, *bootstrap* utilizando método bayesiano e a quantificação no padrão. Sendo os comandos adicionados no pipeline: --k 31, --*validateMappings*, --*numGibbsSamples* 100 --*gcBias* and --*seqBias,*

O pacote em R Tximeta (LOVE *et al.*, 2020) foi utilizado para importar os resultados da quantificação e pseudo-alinhamento, levando em conta replicatas feitas pelo *bootstrap* e replicatas técnicas, e para sumarizá-las em contagem de transcritos por milhão de leituras.

Essas contagens de transcritos sumarizadas já vem incluídas nos metadados do transcriptoma referência utilizado na fase do pseudo-alinhamento. Esse transcriptoma referência também foi utilizado para a construção de um dicionário de conversão de transcritos para genes, que também continha os biotipos dos transcritos e dos genes, os IDs do Ensembl e os nomes oficiais de genes. Todas essas informações foram utilizadas posteriormente para a caracterização funcional dos transcritos.

### 3.3 Análise de expressão diferencial de transcritos

A expressão diferencial foi calculada a partir de contagens de transcrito e gene usando, primeiramente, o pacote DeSeq2 (LOVE; HUBER; ANDERS, 2014) e posteriormente a função swish (ZHU *et al.*, 2019) do pacote *fishpond* individualmente para cada série. O fishpond e sua função swish são um pacote recente que é feito especificamente para computar expressão diferencial à nível de transcrito e que integra melhor as replicatas técnicas produzidas a partir do *bootstrapping* do pseudo-alinhamento.

O valor de corte utilizado para expressão diferencial foi de p valor inferior a 0,05 e *log2FoldChange* absoluto superior a 0,5. Esse valor de corte foi decidido com base na literatura e observando os resultados em que lncRNA possuem níveis de expressão inferiores se comparados à genes codificadores de proteínas, e isso seria ainda menor quando observamos isoformas, sendo assim, os valores de corte usuais de 1 ou 2 *log2FoldChange* levariam à uma perda da maioria dos nossos resultados.

As isoformas de lncRNA diferencialmente expressas foram então comparadas entre as séries. Essa comparação foi feita utilizando upset-plots e heatmaps feitos, respectivamente, nos pacotes UpsetR e ComplexHeatmap. Os upset-plots são uma forma de visualizar intercessões das variáveis nos dados em grupos utilizando também um gráfico de barra para frequência, sendo uma forma diferente de um diagrama de Venn. O heatmap é uma forma bidimensional de representar magnitude de dados por meio de cores, aqui ele foi utilizado para mostrar os transcritos diferencialmente expressos entre as series e seus nomes.

### 3.4 Correlação da expressão de transcritos

Correlação da expressão foi feita utilizando-se as contagens de transcritos por milhão, que foram resultantes da importação pelo tximeta, e o pacote Rcorr utilizando o teste exato de Fisher para significância. A correlação foi feita entre os 36 transcritos diferencialmente

expressos em quatro ou mais series e todos os outros transcritos diferencialmente expressos para achar alvos que os lncRNA poderiam estar regulando. A correlação também foi feita entre os 36 transcritos e todos os transcritos presentes nas tabelas contagens de TPM para enriquecimento funcional posterior utilizando métodos dependentes do pano de fundo de expressão.

### 3.5 Anotação funcional e filtragem de regiões genômicas

A anotação funcional e a filtragem de regiões genômicas foram feitas utilizando os pacotes em R GenomicRanges e biomart e tracks genômicas dos bancos de dados Ensembl (HOWE *et al.*, 2021) e ENCODE (THE ENCODE PROJECT CONSORTIUM, 2012) submetidas para visualização genômica no *Ensembl Genome Browser* e no *UCSC Genome Browser*.

As tracks utilizadas tinham como objetivo visualizar os transcritos em seu contexto genômico adicionando também informações epigenéticas, tais como promotores, acentuadores e marcas de abertura e fechamento de cromatina. As tracks utilizadas do UCSC foram a anotação de transcritos do GENCODE versão 37, a mesma utilizada no alinhamento, para localizar as posições exatas de introns e exons dos transcritos selecionados; a track de elementos regulatórios do GeneHancer, para observar interações promotor-gene e promotor-acentuador em loops, e as tracks de HI-C (*High throughput chromossome conformation capture*) e micro-C em H1-hESC e HFFc6. As tracks utilizadas do ENSEMBL foram a anotação do GENCODE versão 37, para detecção e localização de isoformas específicas por nome, e a marcação de elementos regulatórios do Ensembl, para localização de acentuadores, promotores, região flanqueadora de promotor, CTCF e marcas de cromatina aberta.

Tracks de diversos inputs foram integradas de maneira exata por localização por base e posteriormente editadas para simplificação para gerar as figuras de localização genômica dos resultados. Essa simplificação foi, em sua maioria, a remoção de isoformas de outros genes presentes naquela região que possuíam um número muito alto de isoformase não eram o objetivo da figura, e a região de genes não relacionados à análise foi ofuscada levemente para melhor entendimento.

A distância possível de atuação de lncRNAs que agem em *cis* foi medida utilizando estimativas de tamanho de TADs (*Topologically Associating Domains*) humano de outros dados publicados que fizeram análises de GAM e HI-C de alta profundidade. A distância final

selecionada foi a de 1 megabase para cada lado, por 2 megabases ser próximo da distância máxima de TADs reportada previamente em células humanas.

### 3.6 Enriquecimento funcional de transcritos

O enriquecimento funcional foi feito utilizando o pacote em R *fgsea* para possíveis alvos de lncRNAs agindo em trans. O *fgsea* funciona comparando-se um valor ranqueado entre todos os transcritos, que normalmente é o valor de expressão ou *log2FoldChange*, e o utilizando de forma ordenada para enriquecer os dados com os valores de fundo, o que produz um enriquecimento normalizado. Esse valor representa a distribuição de categorias de grupos de genes, ou vias metabólicas, por uma lista de genes. O algoritmo primeiro ranqueia os genes baseados na medida de cada gene individual, utilizando também a condição observada, e então a lista completa ranqueada é utilizada para verificar como cada gene no grupo de genes está distribuído pela lista. O valor de enriquecimento (ES, *Enrichment Score*) observa o quanto os genes em um gene set estão mais presentes no início ou no final da lista ranqueada. A quantidade final de enriquecimento então é calculada a partir destes valores. O pacote *fgsea* em específico também estima a significância estatística do resultado por um teste de permutação repetindo o procedimento múltiplas vezes. O NES (*Normalized Enrichment Score*) é calculado a partir do ES ajustado para se levar conta as diferenças entre vários gene-sets.

Para esta análise utilizamos como valor de ranqueamento o valor médio de correlação entre a isoforma de lncRNA selecionado e todos os demais transcritos presentes na contagem final por série. Ou seja, todos os transcritos com contagem por milhão acima de 10 foram correlacionados, a partir de seu valor de TPM, com as 36 isoformas selecionadas utilizando o pacote Rcorr em cada uma das seis series analisadas.

O gene set usado foi o molecular signature database (MSigDB) C2 curated gene sets. Esse conjunto engloba diversos gene sets curados manualmente de diversas fontes, incluindo artigos, bancos de dados de vias metabólicas e pesquisadores experts em cada domínio. Esses genes sets foram expandidos para conter também os transcritos produtivos e não produtivos associados aos genes, permitindo-se assim um input a partir de correlação com transcritos.

A análise de enriquecimento foi feita individualmente para alvos das 36 isoformas de lncRNAs selecionadas e para cada uma das seis séries, utilizando um valor de repetição de 100.000 no programa.

O valor de corte de enriquecimento utilizado dependeu da quantidade total de gene sets enriquecidos variando de p ajustado inferior a 0,0001 até p ajustado inferior a 0,001. Para identificação de vias enriquecidas simultaneamente entre as séries foram selecionadas também somente aquelas presentes em ao menos dois experimentos independentes, sendo aqueles experimentos que possuem mais de uma série contado uma única vez.

Para gerar as figuras de enriquecimento as tabelas foram filtradas tendo em conta: vias metabólicas de humano possivelmente associadas ao fármaco e vias que não fossem extremamente gerais de processos metabólicos muito complexos (ex: Ciclo celular, Reparo). Todos os resultados foram mostrados em forma de gráficos de enriquecimento no formato pirulito, que permite a visualização de tamanho dos p valores, e intensidade do valor de NES para processos super e sub regulados.

# 4. Resultados

## 4.1 Bibliotecas selecionadas

A partir da seleção descrita nos Métodos, dos 14 experimentos iniciais obtidos (Tabela 1), todos foram passados pelo pipeline de expressão diferencial, no entanto, na análise de DTE foi visto que as bibliotecas cujas amostras eram de pacientes tratados e controles apresentavam um número muito menor de transcritos diferencialmente expressos se comparados com as bibliotecas de células (**Figura 7**). A partir desse resultado foi decidido que as análises posteriores seriam realizadas apenas nas bibliotecas de células pareadas tratadas e não tratadas com metformina que possuem no mínimo 20M de leituras de profundidade e tamanho de leitura de, ao menos, 100 bases.

*Tabela 1: Descrição dos dados de sequenciamento obtidos do SRA*

| **Número do experimento** | **Nome do primeiro autor** | **Tipo de sequenciamento** | **Profundidade do sequenciamento** | **Dosagem de Metformina em mM (mM = milimolar)** | **Tamanho da leitura** | **Número de bibliotecas** | **Cell or tissue type** |
|---|---|---|---|---|---|---|---|
| 1 | Rodriguez | Single | ~20M | 0,0003, 0,003 e 0,01 mM | 75 | 12 | Células ReN |
| 2 | Liu | Paired | ~30M | 10 mM | 150 | 4 | Células CAL27 e HSC2 |
| 3 | Kulkarni | Paired | ~40M | 0,01 a 0,04 mM (pacientes recebendo 1 a 2g/dia) | 150 | 136 | Biópsias musculares de pacientes |
| 4 | Guiliani | Single | ~20M | 0,2 mM | 100 | 34 | HUVEC |
| 5 | Xie | Paired | ~20M | 32,27 mM A549 / 69,97 mM 786-O | 150 | 24 | Células A549 e 786-O |
| 6 | Tan | Single | ~50M | não presente no artigo | 75 | 23 | Células CD4+T |
| 7 | Laustriat | Paired | ~60M | 25 nM e 10nM | 100 | 9 | Células HeSC |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 8 | Kulkarni | Paired | ~20M | 18 uM no sangue = 0,018 mM no sangue | 100 | 100 | Biópsias de músculo, esqueleto e tecido adiposo |
| 9 | Gillespie | Paired | ~30M | 1 mM e 0,5 mM | 130 | 6 | Células 2DD |
| 10 | Lachmandas | Paired | ~20M | 0,01 a 0,04 mM (pacientes recebendo 1 a 2g/dia) | 43 | 44 | Células PBMC |
| 11 | Luizon | Paired | ~30M | 2,5 mM | 100 | 15 | Hepatócitos primários |
| 12 | Monash | Single | ~30M | 2 mM | 40 | 12 | Células HAEC |
| 13 | Yue | Paired | ~20M | 5 mM | 100 | 32 | Células PANC-1 |
| 14 | Balliu | Paired | ~30M | 1mM | 150 | 9 | Células HEPG2 |

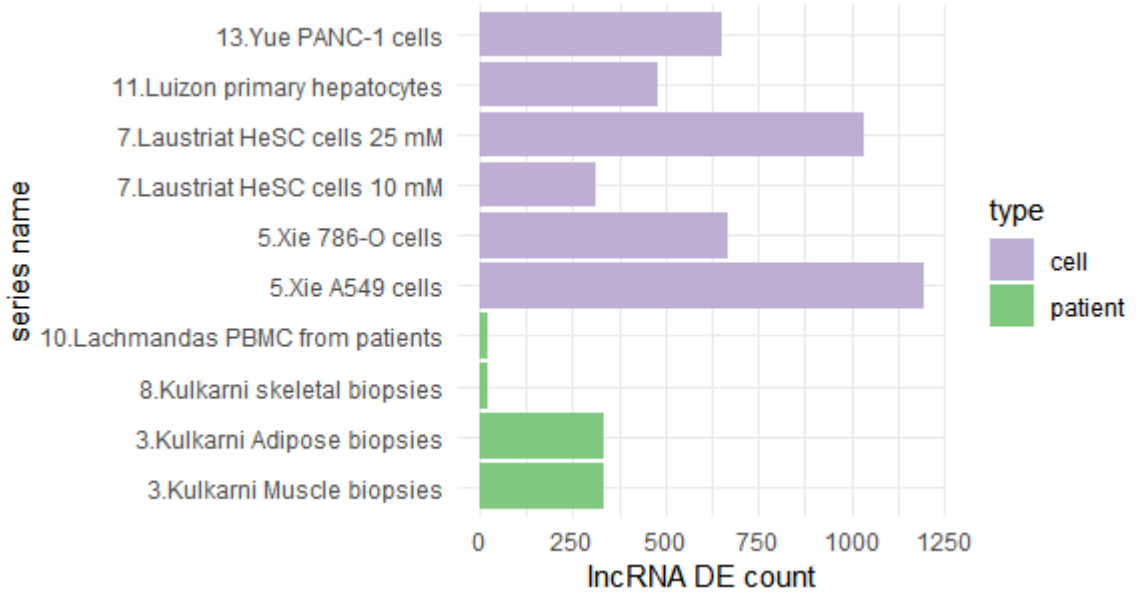


***Figura 7:*** *Contagem de isoformas de lncRNA diferencialmente expressas entre as séries. Em roxo os estudos de linhagens celulares ou de células primárias humanas e em verde de tecidos de biópsia ou pacientes. No eixo x a contagem de transcritos diferencialmente expressos e no*

*eixo y as séries analisadas. Em roxo linhagens celulares de células imortalizadas e de hepatócitos primários, em verde tecidos de biópsia e células de pacientes.*

Essa seleção resultou em quatro bibliotecas que foram separadas para análises comparativas posteriores e essas possuíam no total seis séries ou experimentos comparativos de concentrações de metformina em certo tipo celular contra o controle (Tabela 2).

*Tabela 2: Séries de Metformina selecionadas para análises posteriores*

| **Nome do artigo** | **ID do projeto no SRA** | **Tecnologia usada** | **Concentração de Metformina** | **Tipo celular** | **Profundidade da biblioteca** | **Tamanho das leituras** | **Horas pós tratamento** |
|---|---|---|---|---|---|---|---|
| Xie, 2020 | PRJNA612620 | Illumina NovaSeq 6000 | 69.97 mM | 786-O (câncer renal) | ~25M | 150 | 48 hours |
| Xie, 2020 | PRJNA612620 | Illumina NovaSeq 6000 | 32.27 mM | A549 (Celulas Epiteliais Pulmonares de tumor) | ~25M | 150 | 48 hours |
| Laustriat, 2015 | PRJNA577137 | Illumina HiSeq 2000 | 10 mM | HEsC (Células Tronco Embrionárias) | ~60M | 100 | 48 hours |
| Laustriat, 2015 | PRJNA577137 | Illumina HiSeq 2000 | 25 mM | HEsC (Células Tronco EMbrionárias) | ~60M | 100 | 48 hours |
| Yue, 2015 | PRJNA277028 | Illumina HiSeq 2500 | 5 mM | PANC-1 (células pancreáticas de carcinoma) | ~20M | 100 | 72 hours |
| Luizon, 2016 | PRJNA324847 | Illumina HiSeq 2000 | 2.5 mM | Hepatócitos Primários | ~30M | 100 | 8 hours |

Após a seleção, as séries foram analisadas de acordo com a sessão de métodos, a expressão diferencial foi calculada individualmente para cada série e apenas os lncRNAs foram selecionados para as análises posteriores. Essa seleção dos lncRNAs foi feita com base na anotação do transcriptoma humano do GENCODE versão 37 utilizando o critério gene_type ou tipo do gene e selecionando apenas transcritos cujos genes tinha tipo “lncRNA” de acordo com a versão 37.

### 4.2 Otimização do pipeline

Em busca de executarmos a análise bioinformática da melhor forma possível e que fazia sentido estatístico, biológico e de acordo com a literatura da área, aplicamos vários ajustes nos modelos predefinidos dos programas para otimização do pipeline.

Um ajuste a ser observado foi o k_mers utilizados no pseudo-alinhamento pelo *Salmon* pelo programa. O k_mer é o tamanho mínimo de mapeamento para que o pseudo-alinhamento seja considerado positivo. Então um valor menor poderia melhorar a sensibilidade quando se usa outras opções de validação de alinhamento e múltiplas execuções. Foram comparados os k_mers de tamanho 15, o mínimo recomendado para leituras de 100 ou mais nucleotídeos, e de tamanho 31 que é o padrão do programa para leituras de 75+ pares de base. O tamanho do k_mer não influenciou a contagem de transcritos por milhão (TPM, *Transcripts Per Million*) ou de DETs. Assim, o k_mer de 31, que é o modelo, foi mantido para todas as análises.

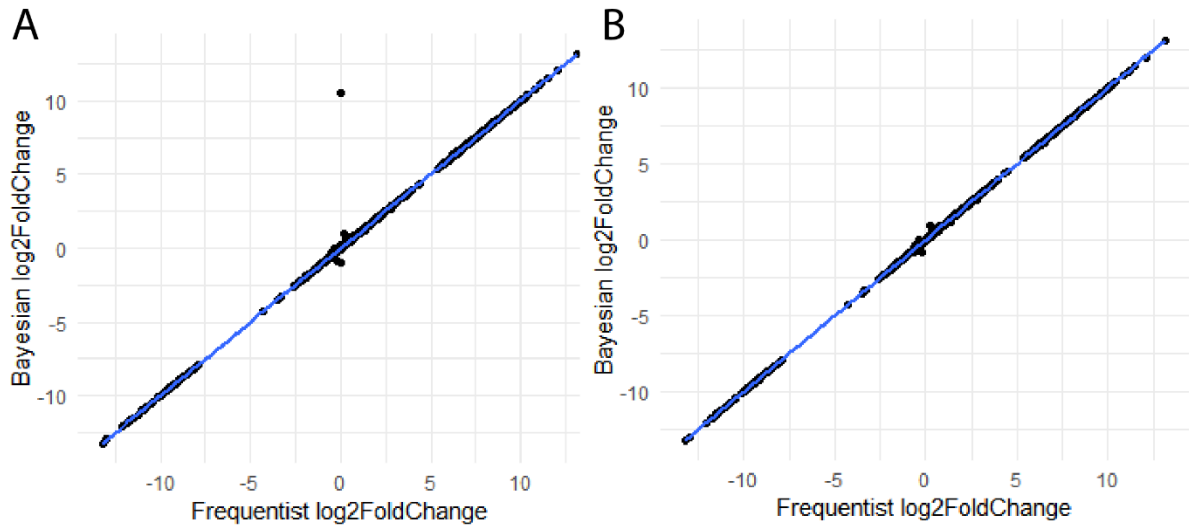


***Figura 8:*** *Comparação entre os log2FoldChanges dos transcritos encontrados a depender do método de bootstrap usado utilizando análise de correlação. Em A o resultado em log2FoldChange dos transcritos do bootstrap por inferência bayesiana (eixo y) e inferência frequentista (eixo x) em todos os transcritos que tiveram leituras contadas na análise. Já em B o mesmo resultado somente com transcritos que passaram no corte de expressão diferencial.*

Outro ajuste feito foi o método de *bootstrapping* a ser utilizado. O *bootstrap* é um procedimento de reamostragem das contagens obtidas utilizando diferentes classes de equivalência e procedimento de otimização para cada amostra rodada. O *bootstrap* permite computar a variância técnica da estimativa de contagem produzida pelo programa e tal variância pode ser utilizada para análises posteriores, tais como expressão diferencial e correlação. Quanto maior o número de *bootstraps*, melhor a estimativa da variância, mas maior poder de processamento gasto. O programa permite dois métodos de *bootstrapping*: o método frequentista e o método bayesiano. Ambos os métodos foram comparados utilizando-se a quantidade recomendada de *bootstraps* (100) o que levou a um resultado muito parecido

quando foi calculada a expressão diferencial (**Figura 8**). Então o método bayesiano foi selecionado pelo seu menor tempo gasto para o processamento.

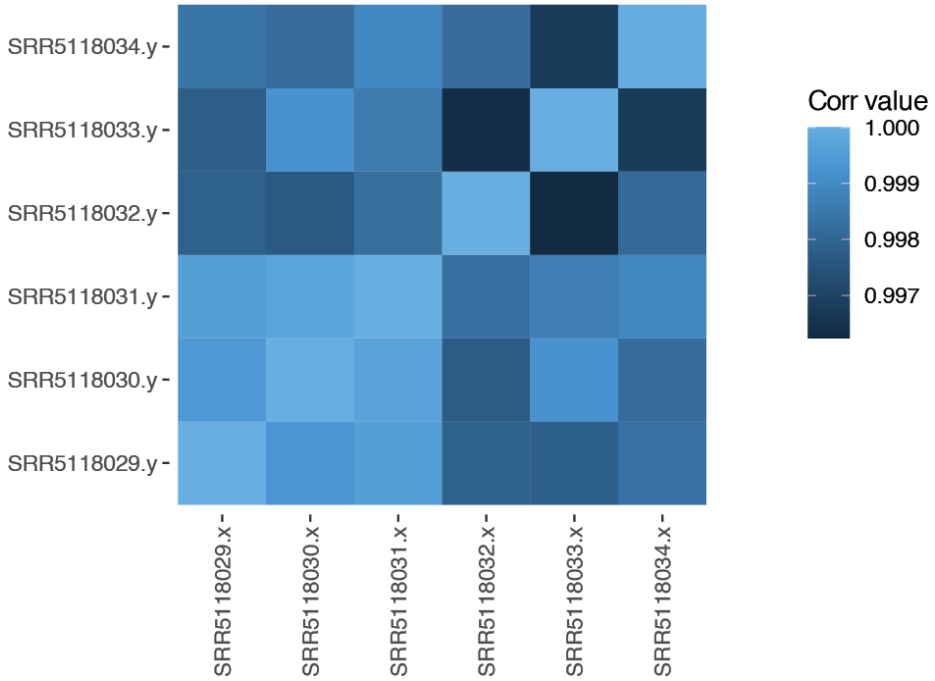


***Figura 9:*** *Comparação entre os valores de TPM de duas rodagens independentes do Salmon para as amostras de hepatócitos primários. As bibliotecas com terminação .x são da primeira rodagem e as. Y da segunda.*

O Salmon também foi rodado utilizando as mesmas condições diversas vezes para garantir que o programa estava obtendo os mesmos resultados a partir do mesmo input. Os resultados de TPM para a biblioteca de hepatócitos primários estão presentes na **Figura 9**, onde pode ser observada a correlação próxima a 1 entre duas rodagens independentes do programa, o que foi observado para as demais bibliotecas analisadas.

A expressão diferencial também foi um ponto a ser refinado, já que os programas mais utilizados para análise de expressão diferencial como o *DeSeq2* e o *EdgeR* (ROBINSON; MCCARTHY; SMYTH, 2009) não foram pensados a priori para uma análise a nível de isoforma. Os resultados obtidos pelo Deseq2 e pelo *swish*, que é um pacote mais recente, foram então comparados para verificarmos a sensibilidade do programa na detecção de isoformas de lncRNA diferencialmente expressas utilizando os mesmos valores de corte. Nas séries cujas bibliotecas apresentavam maior profundidade o Deseq2 e o *swish* performaram de maneira muito similar com um número irrisório de transcritos sendo observado mais ou menos por cada

programa, no entanto nas séries com menor profundidade o *swish* se mostrou muito mais sensível na detecção.

Por este resultado, as expressões diferenciais calculadas a partir da execução do *swish* foram selecionados para as análises posteriores. Metformina e controle foi o grupo contraste utilizado em cada comparação o que resultou em seis tabelas independentes de expressão diferencial. Todas as isoformas que estavam presentes em, ao menos, quatro dos seis experimentos com a direção do *log2FoldChange* igual, super ou sub, foram selecionados para as análises posteriores.

Os valores de corte para a análise de correlação entre isoformas de lncRNAs e possíveis transcritos codificadores de proteínas alvos também foram acertados. Vários valores diferentes de corte de valor de correlação e de p valor foram testados para o melhor resultado com confiabilidade estatística, de forma a não perder um número muito grande de correlações (**Figura 10**). O valor de corte de correlação estabelecido foi de 0,8 (80%) de valor de correlação e 0,01 de p valor.

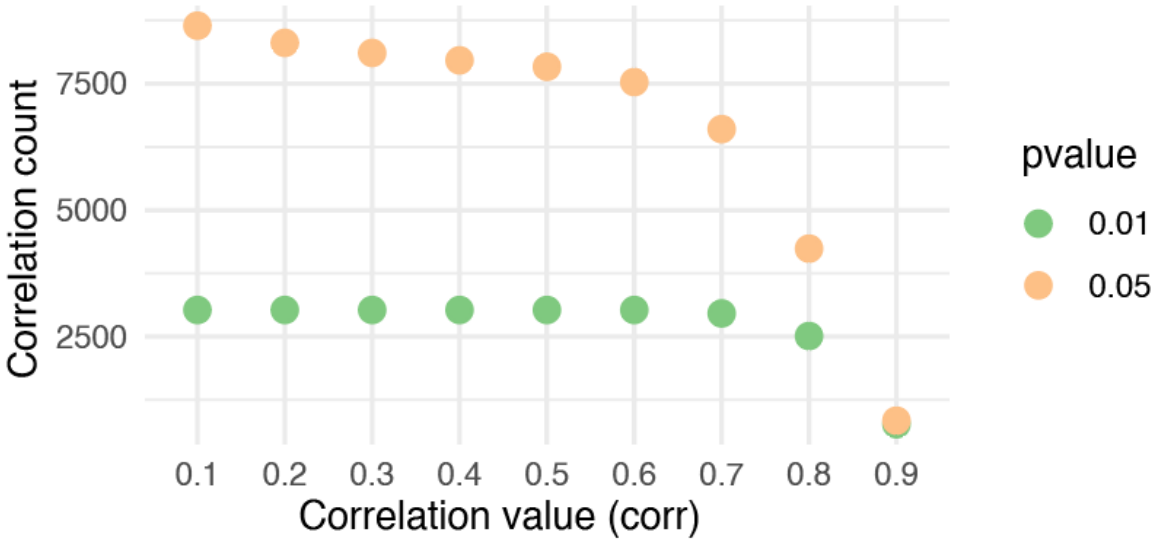


***Figura 10:*** *Gráfico de comparação entre o número total de correlações obtidas e os valores de corte utilizados. No eixo x o valor de correlação e no eixo y as contagens do correlações que passam naquele valor de corte. Em verde número de correlações que são significativas de acordo com o pvalor, em verde pvalor de 0,01, em laranja pvalor de 0,05.*

Para a verificação de possíveis isoformas agindo na mesma região genômica em que são transcritas, vários tamanhos médios de TAD foram considerados e analisados de forma a

se contar a quantidade de transcritos presente no tamanho médio (300 kilobases, 500 kilobases e 1 megabase; **Figura 11** A, B e C, respectivamente) para cada lado ao redor de cada um dos 36 lncRNAs selecionados para as análises. Como foi descrito nos métodos, a distância final selecionada foi de 1 megabase para cada lado, que fazia sentido tanto associada aos resultados da literatura quanto à **Figura 10**.

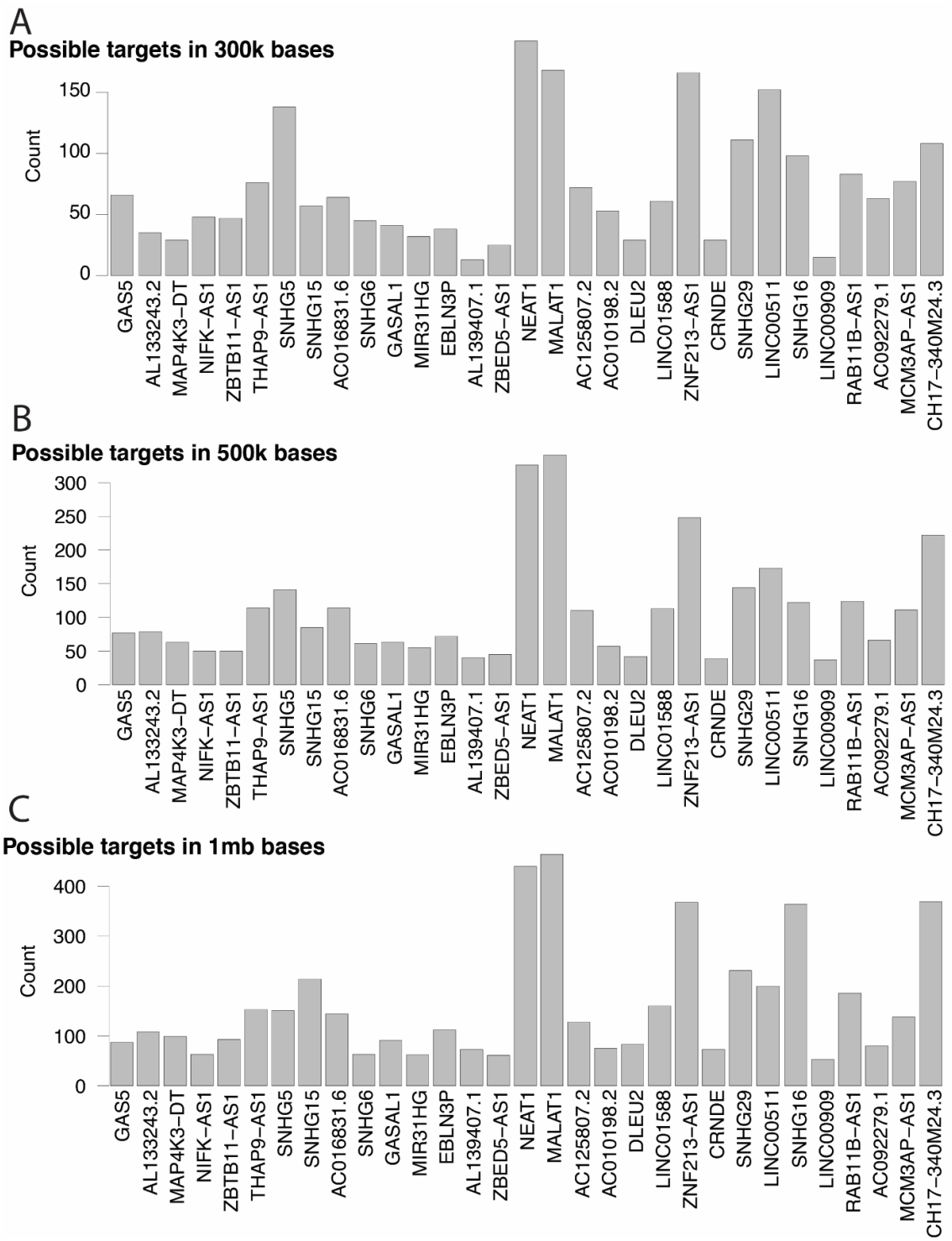


***Figura 11:*** *Contagem de transcritos ao redor das regiões de cada um dos lncRNAs cujas*

*isoformas foram selecionadas para análises posteriores. No eixo x os genes associados às isoformas de lncRNA e no eixo y a contagem de transcritos ao redor. Na letra **A** foi contado o número de transcritos em 300kilobases para cada lado dos genes, na letra **B** 500 kilobases e na letra **C** 1 megabase.*

### 4.3 Transcritos diferencialmente expressos

O número de isoformas de lncRNA diferencialmente expressas varia drasticamente entre as séries (**Figura 12**). Em todas as séries foi utilizado o valor de corte de p-valor < 0,05 e *log2FoldChange* absoluto > 0,5 para determinação de isoformas diferencialmente expressas.

Esse valor de corte utilizado se baseou na literatura associada tanto à expressão diferencial de isoformas e quanto à lncRNAs. Os lncRNAs possuem um nível de expressão mais baixo se comparado a genes codificadores de proteínas (XU *et al.*, 2017), eles são mais sensíveis a alterações no meio e pequena diferença em sua expressão provoca mudanças significativas no ambiente celular (MATTICK *et al.*, 2023). O cálculo de expressão de gene a partir de um pseudo-alinhamento no transcriptoma faz uma somatória dos valores de contagens por milhão de todos os transcritos originários daquele gene como a contagem total do gene. Dessa forma, genes tem um valor de expressão muito superior à de seus transcritos, o que resulta em valores maiores de *log2FoldChange*, quando é calculada a expressão diferencial. Por essas razões, utilizamos um valor de corte de *log2FoldChange* absoluto maior que 0.5, e não 1 como é comum na literatura da área.

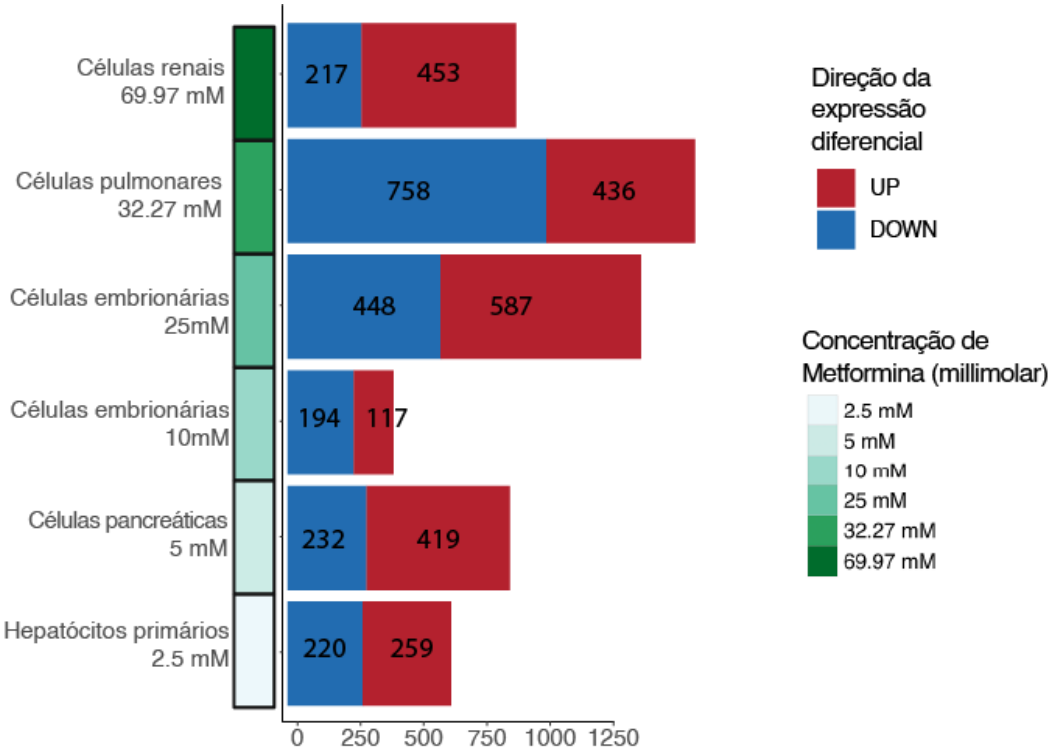


***Figura 12:*** *Contagem de isoformas de lncRNA diferencialmente expressas nas séries organizado por concentração de metformina utilizada no tratamento, utilizando um valor de corte de pvalor < 0,05 e log2FoldChange> 0,5. As séries estão ordenadas de acordo com concentração de metformina descendente representada em tons de verde a esquerda do gráfico.*

Existe uma leve tendência do maior número de isoformas estar associado à maior concentração do fármaco, com séries como células pulmonares, renais e embrionárias com tratamento de 25 milimolar (mM) tendo o maior número de transcritos no geral se comparadas às demais (**Figura 12**). No entanto, a profundidade de sequenciamento e tecnologia utilizada também afetou o número total de transcritos diferencialmente expressos por série. Séries com maior profundidade e com tecnologias mais avançadas, como o Nova-seq, também obtiveram um número de isoformas de lncRNA diferencialmente expressas maior.

Seguindo a ordem crescente de tratamento por metformina: Hepatócitos primários possuem um número similar de isoformas super e subexpressas, respectivamente 259 e 220. Células pancreáticas possuem um número de supereguladas quase que duas vezes maior se comparado com o número de isoformas subexpressas, 419 a 232. Células embrionárias tratadas com 10 mM de metformina possuem o menor número total de transcritos diferencialmente expressos, 117 super 194 subexpressos. Células embrionárias tratadas com 25 mM de

metformina possuem 587 super e 448 isoformas subexpressas. Células pulmonares possuem o maior número total de isoformas de lncRNA diferencialmente expressas, 436 superexpressas e 758 subexpressas. Células renais possuem a maior concentração total de metformina e sua análise resultou em 453 isoformas superexpressas e 217 isoformas subexpressas (**Figura 12**).

### 4.4 Intercessão entre as séries

Todas as isoformas de lncRNA diferencialmente expressas em cada série foram então comparadas com àquelas diferencialmente expressas nas demais séries por análises de intercessão (**Figura 13 A e B**).

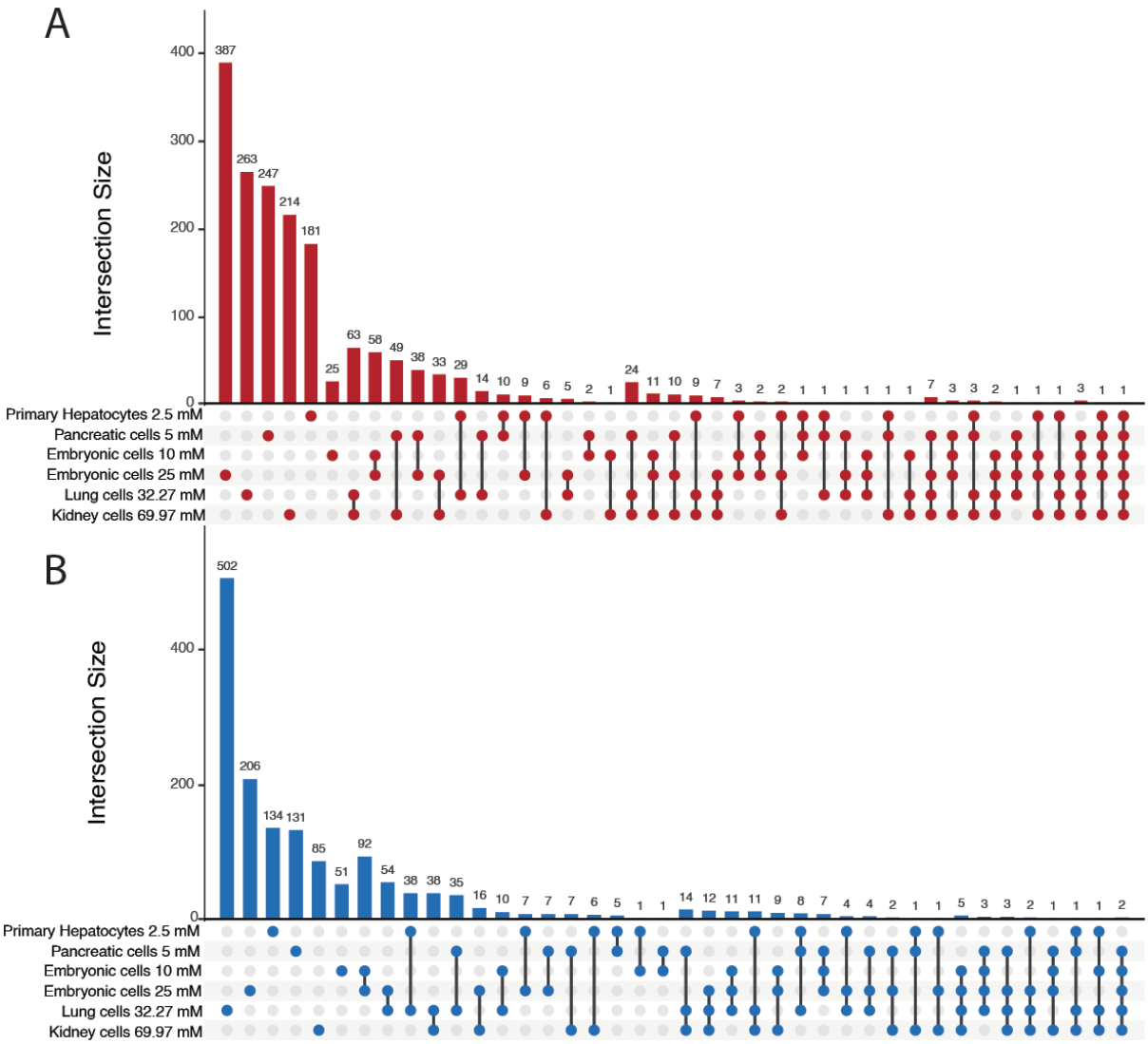


***Figura 13**: Intercessão entre as isoformas diferencialmente expressas presentes nas seis séries representada em formato de upsetplots. Cada ponto preenchido mostra que aquele grupo de isoformas está presente naquela comparação e cada linha preta mostra intercessões com*

*outros grupos, o gráfico de barra no topo mostra a quantidade de intercessões entre cada grupo. Em **A** transcritos superexpressos e em **B** transcritos subexpressos.*

Como esperado, a grande maioria das isoformas diferencialmente expressas era série específica. De todas as isoformas encontradas (1731 superexpressas e 1528 subreguladas) 1317 superexpressas e 1109 subexpressas estavam presentes em apenas uma das séries e 414 superreguladas e 419 subreguladas se repetiam em pelo menos duas séries.

Em relação as isoformas superexpressas os pares de série com maior sobreposição, 63 isoformas entre pulmonar e renal e 58 isoformas entre ambas as células embrionárias, também são as séries derivadas do mesmo experimento original, como era esperado. Também existe uma alta sobreposição entre células de pâncreas e renais (49), células de pâncreas e embrionárias tratadas com 25 mM do fármaco (38), células embrionárias tratadas com 25 mM e células renais (33) e hepatócitos primários e células pulmonares (29).

Em relação às isoformas reguladas, o grupo com maior sobreposição foi entre as séries de células embrionárias (92). No entanto, as segundas maiores sobreposição foram entre células embrionárias e pulmonares (38) e células pulmonares e renais (38), também houve uma alta sobreposição entre células pancreáticas e pulmonares (35). Um número muito pequeno de isoformas, no total 74 superexpressas e 84 subexpressas estão presentes em ao menos três séries. Sendo as maiores sobreposições tanto nos super (24) quanto nos subexpressos (14) sendo entre células pancreáticas, pulmonares e renais. Um número ainda menor de isoformas, 20 superexpressas e 16 subexpressas, estão presentes em ao menos quatro séries. Em cinco séries aparecem como intercessão apenas quatro isoformas superexpressas e duas subexpressas. Apenas uma isoforma está presente como diferencialmente expressa em todas as séries e está superexpressa.

Como o objetivo do trabalho era a análise das isoformas de lncRNA que possuem expressão diferencial equivalente entre as séries, o foco principal de nossas análises foram as 36 isoformas super e subexpressas em, ao menos, quatro séries e, em especial, as seis isoformas super e subexpressas que aparecem em, ao menos, cinco séries.

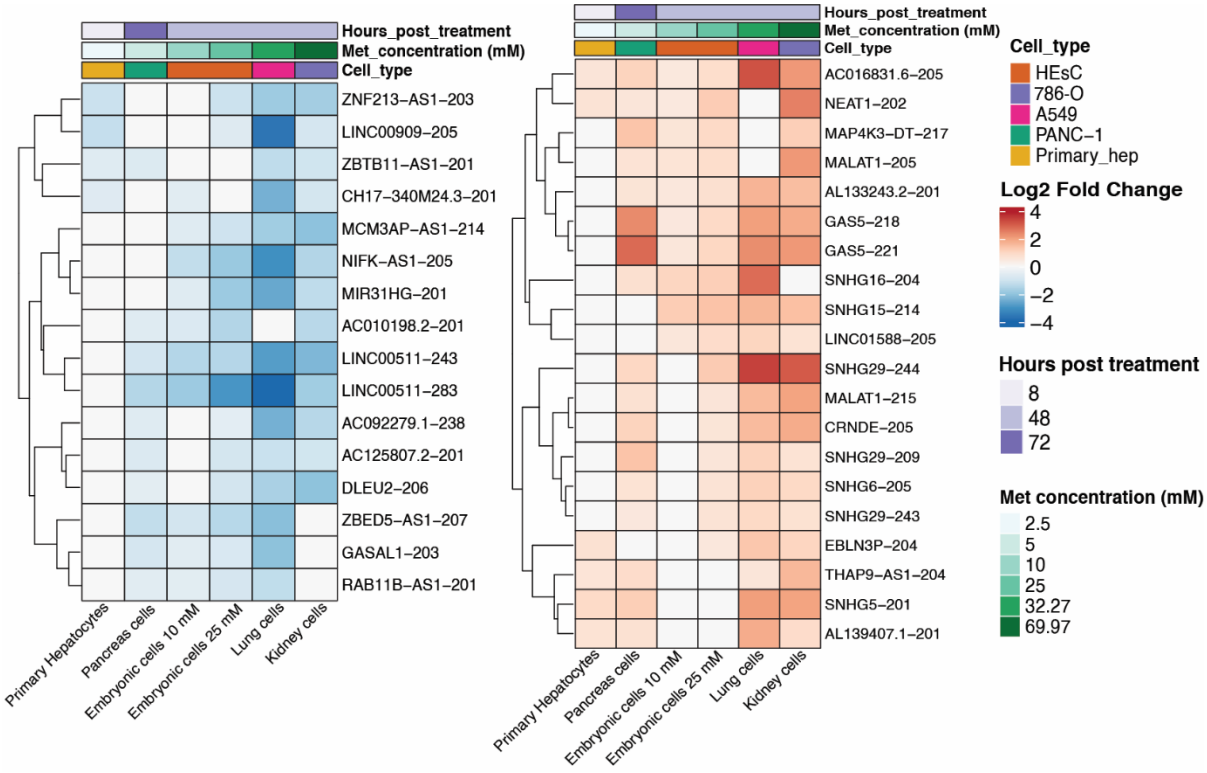


***Figura 14**: Heatmap das 36 isoformas de lncRNA diferencialmente expressas em ao menos quatro séries com seus metadados. As séries estão ordenadas pela concentração de metformina representada na legenda. No eixo x as séries e no eixo y as isoformas de lncRNA. Na legenda pode ser visto as horas após o tratamento, o tipo celular de cada série, o valor de FoldChange dos transcritos e a concentração de metformina.*

Quando estas 36 isoformas diferencialmente expressas são visualizadas na forma de um *heatmap* (**Figura 14**) é possível perceber que as séries com maior concentração de metformina utilizada no tratamento também possuem o maior número de transcritos compartilhados com as demais séries (célula pulmonar e célula renal).

A nomenclatura principal utilizada tanto no *heatmap* quanto nas análises posteriores foram os nomes de transcrito de cada isoforma, estes que normalmente se caracterizam pelo nome do gene a qual está associado seguido de um traço e um número, que representa a ordem de descobrimento da isoforma. Esse nome de gene para os lncRNA normalmente está associado à função proposta pelos autores que o descreveram. Por exemplo, *MALAT1* é uma sigla para *Metastasis Associated Lung Adenocarcinoma Transcript 1*, ou transcrito associado a metástase em adenocarcinoma pulmonar 1. No entanto, muitas outras possíveis funções já foram atribuídas ao *MALAT1* desde sua anotação em 2003 que diferem do nome dado ao transcrito. O problema de nomenclatura de lncRNAs e o porquê da escolha de se referir às isoformas por

seu nome e pelo seu número de identificação do ENSEMBL será abordado com mais profundidade na discussão.

No heatmap existem algumas isoformas cujo nome é uma sequência de números. Essas isoformas são recém anotadas, cuja função ainda não foi explorada na literatura, então possuem apenas seu nome de anotação. Entre as superexpressas AC016831.6-205 (ENST00000643779.1), AL133243.2-201 (ENST00000610331.1) e AL139407.1-201 (ENST00000613826.1); e entre as subexpressas CH17-340M24.3-201 (ENST00000360656.2), AC010198.2-201 (ENST00000650193.1), AC092279.1-238 (ENST00000670058.1) e AC125807.2-201 (ENST00000513358.3). O tipo de análise exploratória que foi realizada nos dados é uma boa forma de atribuir possíveis funções a estes transcritos recentemente anotados de função ainda desconhecida.

Isoformas que possuem o -AS1- no nome são isoformas antisenso ao gene do qual são nomeadas, ou seja, possuem sequência complementar e oposta parcial ou totalmente aquele gene. Em nossos resultados foram encontradas cinco isoformas subexpressas; ZNF213-AS1-203 (ENST00000571963.5), ZBTB11-AS1-201 (ENST00000609682.1), MCM3AP-AS1-214 (ENST00000669476.1), ZBED5-AS1-207 (ENST00000664276.1) e RAB11B-AS1-201 (ENST00000593581.5) e uma superexpressa, THAP9-AS1-204 (ENST00000504792.6) que são caracterizadas como antisenso. Essa característica de antisenso normalmente implica na possibilidade desses transcritos regularem a expressão de seu gene senso, inclusive com possibilidades biotecnológicas dessa aplicação para controle de expressão (CUI; ZHAN; LIU, 2021; KRAPPINGER *et al.*, 2021) . Outra classe de lncRNA que aparece no *heatmap* são os lncRNA com sufixo -DT- no nome. Esse sufixo caracteriza o lncRNA como um transcrito divergente (*Divergent Transcript* ou DT), que são transcritos similares ao gene original, mas que apresentam diferentes sítios de iniciação, promotores ou término (CROUSE *et al.*, 1985), essa classificação é bem antiga e atualmente as novas anotações tem anotado lncRNA DT como isoformas diferentes do mesmo lncRNA original.

Entre os lncRNA nomeados aparecem muitos indivíduos conhecidos da literatura de lncRNAs. NEAT1-202 (*Nuclear Paraspeckle Assembly Transcript 1*, ENST00000501122.2) é um dos lncRNA mais conhecidos desde sua descoberta no início dos anos 2000 (BOND; FOX, 2009; CLEMSON *et al.*, 2009), sua função primária está associada a formação de *paraspeckles*, corpos nucleares encontrados na região próxima a eucromatina que são responsáveis por controle da expressão de certos genes e manutenção da forma do núcleo (IP; NAKAGAWA,

2012). *NEAT1* e seus transcritos associados são exclusivos de vertebrados e tem sido encontrados em resposta à diversas condições humanas, em especial o câncer (ZHANG, MIAO *et al.*, 2022). *NEAT1* já foi encontrado previamente em bibliotecas de células tratadas com metformina (QIN et al., 2021; SCHULTEN; BAKHASHAB, 2019). No entanto como em sua ação no câncer, se este aparece super ou subexpresso e a função associada a este lncRNA varia entre estudos. Uma meta-análise recente de estudos de microarranjos tratados com metformina em células tumorais (SCHULTEN; BAKHASHAB, 2019), encontrou o lncRNA *NEAT1* diferencialmente regulado a depender do tipo celular e condição de tratamento, sendo que, na maioria das vezes, ele aparecia subexpresso em células tratadas com metformina.

O gene NEAT1 possui 15 isoformas anotadas na última versão disponível da anotação do genoma humano (GENCODE40), porém ele não é comumente analisado em nível de isoforma e esta é a primeira vez na literatura em que *NEAT1* isoforma-específico é visto em sua regulação relacionada ao tratamento por metformina. As análises que já foram feitas em nível de isoforma, normalmente dividem *NEAT1* em apenas duas isoformas: NEAT1_1 e NEAT1_2, sendo a primeira de tamanho menor, que é expressa de forma ubíqua e a segunda bem maior, e mais diretamente relacionada à formação de *paraspeckles* e controle da expressão gênica. Esses mesmos estudos acharam a isoforma NEAT1_2 superexpressa em alguns tipos de canceres de pulmão (KNUTSEN *et al.*, 2020; KNUTSEN; HARRIS; PERANDER, 2022) .

Em nossos resultados, a isoforma NEAT1-202 está superexpressa em todas as séries, exceto nas células pulmonares. Essa isoforma possui uma grande região exonica à jusante se comparada às demais isoformas de NEAT1 e hoje é a considerada canônica **(Figura 15)**.

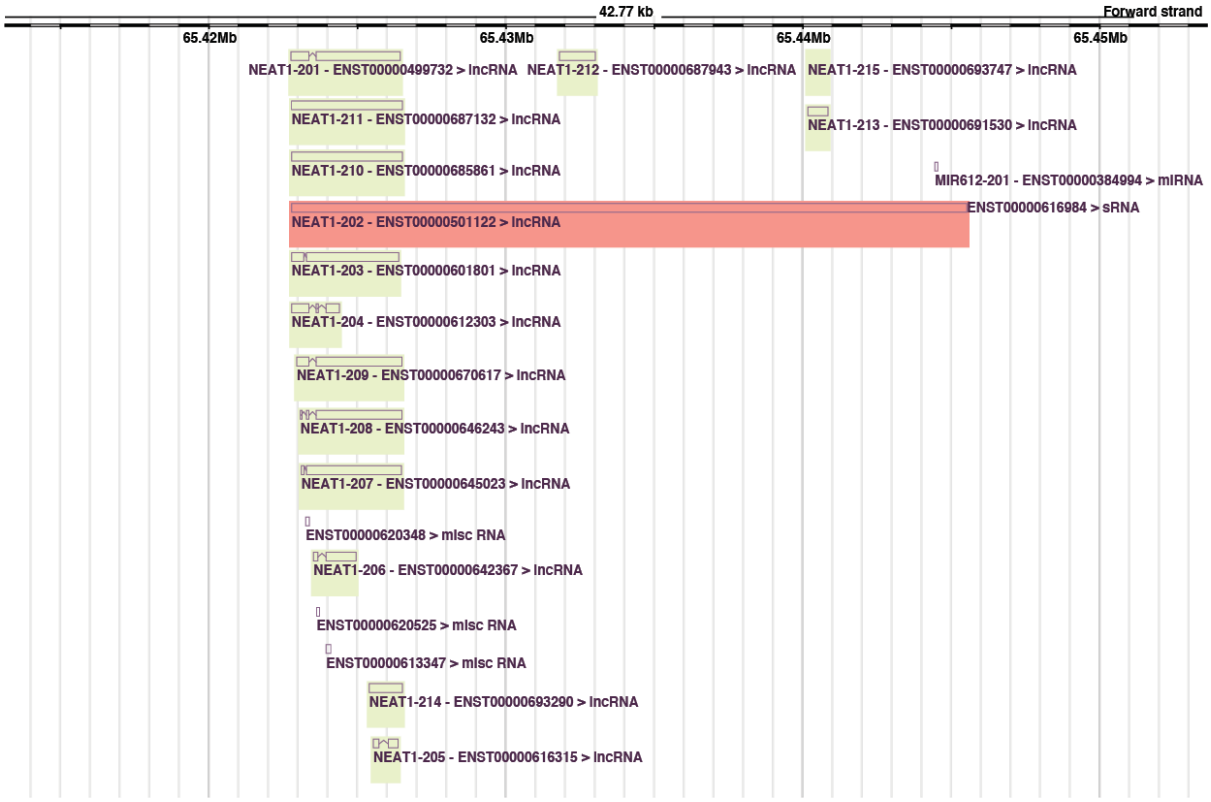


***Figura 15:*** *Região dos transcritos de NEAT1 do Ensembl Genome Browser. Em vermelho a isoforma considerada canônica de NEAT1 que também foi achada em nossos resultados, em verde as demais isoformas de NEAT1.*

O *LINC00511* (*Long Intergenic Non-Protein Coding RNA* 511) também foi observado em cinco das seis séries analisadas, dessa vez com duas isoformas distintas diferencialmente expressas (subreguladas): LINC00511-243 (ENST00000648623.2) e -283 (ENST00000650131.2), que aparecem em todas as séries exceto nos hepatócitos primários. *LINC00511* foi primeiro reportado em 2016, em um estudo em que aparece muito superexpresso em câncer de pulmão enriquecido por *HER2* quando comparado à tecido normal (YANG, F. *et al.*, 2016). Desde então, diversos estudos têm reportado sua superregulação como associada à piores prognósticos e cânceres mais nocivos tanto em estudos *in vitro* como em pacientes (CHENG; WANG; MU, 2021; DING *et al.*, 2020) . Em alguns casos, sua superregulação é crucial para o desenvolvimento do tumor, com reportado em uma meta-análise recente, sugerindo seu uso como biomarcador de prognóstico para certos tumores (AGBANA *et al.*, 2020). *LINC00511* também já foi associado à diversas funções oncogênicas, como promover proliferação celular, aceleração de metástase e influência em comportamento celular invasivo de células de tumor (LU *et al.*, 2018).

Nenhum estudo até o presente momento fez análise de *LINC00511* em nível de isoforma, sendo que ele possui mais de 500 isoformas anotadas, segundo a última anotação do transcriptoma humano. As isoformas encontradas em nossa análise são menores que a isoforma canônica (LINC00511-244), possuindo apenas metade de seu tamanho. Elas são muito similares, no entanto, se diferem pela presença de um exon extra na segunda posição da isoforma -243, se comparada com a -283 **(Figura 16)**.

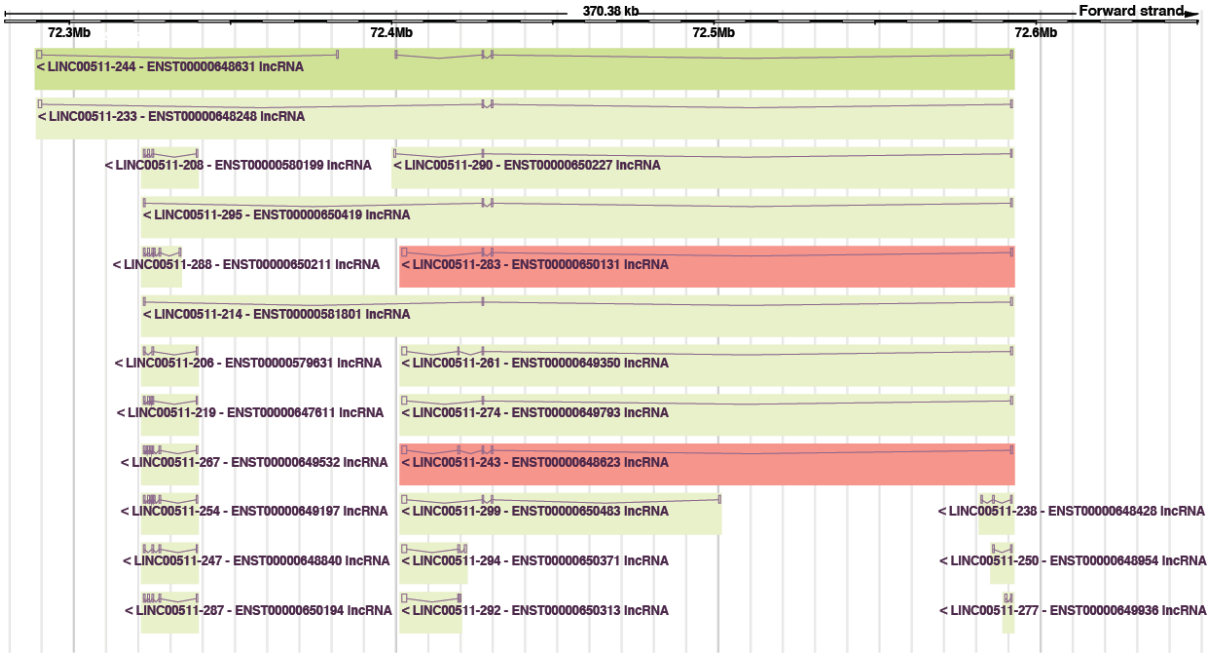


***Figura 16:*** *Região de alguns dos transcritos de LINC00511 do Ensembl Genome Browser. Em vermelho os transcritos encontrados em nossa análise, em verde escuro a isoforma canônica de LINC00511 e em verde claro as demais isoformas deste gene.*

*GAS5* (*Growth Arrest Specific 5*) é outro lncRNA que foi observado em cinco das seis séries analisadas, também com duas isoformas como superexpressas GAS5-218 (ENST00000448718.5) e GAS5-221 (ENST00000450589.5), exceto em hepatócitos primários. *GAS5* foi primeiramente identificado e caracterizado em 1992 em camundongo, como ubiquamente expresso durante o desenvolvimento tecidual. Também foi demonstrado que o acúmulo desse mRNA leva à parada de crescimento celular (COCCIA *et al.*, 1992). O GAS5 já foi associado à múltiplos fenótipos na literatura como diabetes (CHU *et al.*, 2022), doenças cardiovasculares (CAO *et al.*, 2022) e câncer (GHAFOURI-FARD *et al.*, 2021). Uma recente revisão da literatura sobre GAS5 mostra seu potencial como codificador de pequenos peptídeos, como esponja de *small nucleolar RNAs* (snoRNAs), assim como sua conservação em mamíferos (GOUSTIN *et al.*, 2019) .

Até o presente momento, não encontramos na literatura nenhum artigo em que *GAS5* tivesse sua expressão analisada em nível de isoforma, mesmo possuindo mais de 70 isoformas anotadas na última versão do transcriptoma humano. As isoformas que achamos em nosso estudo, -218 e -221, possuem um sítio de início e término alternativo de transcrição se comparadas à isoforma canônica, GAS5-231. As isoformas -218 e -221 compartilham todos os mesmos exons, no entanto, o tamanho de seu sétimo exon difere **(Figura 17)**.

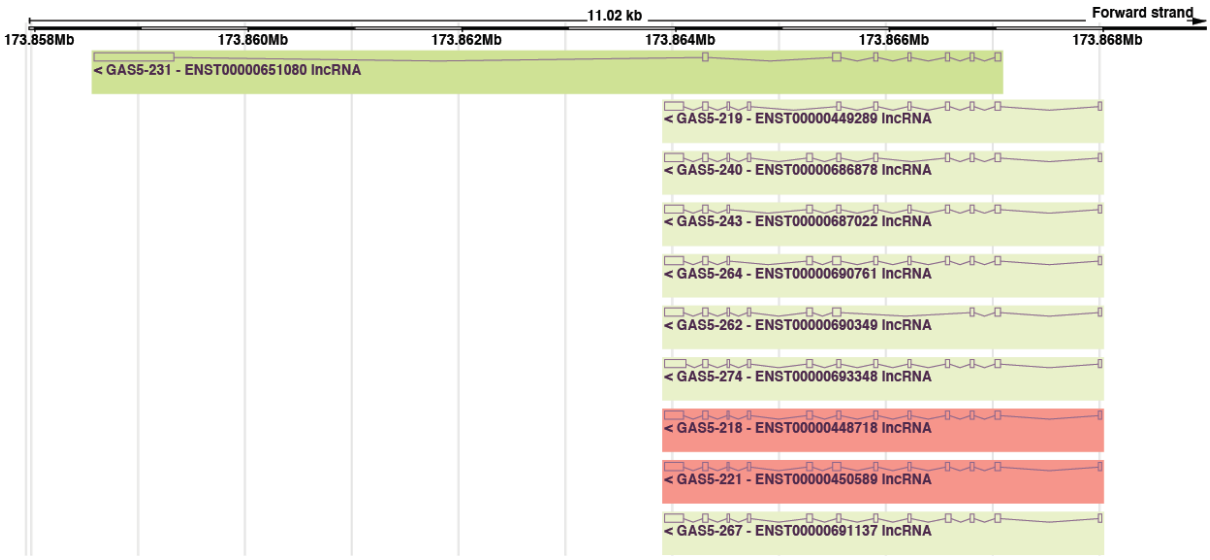


***Figura 17**: Região de alguns dos transcritos de GAS5 do Ensembl Genome Browser. Em vermelho os transcritos encontrados em nossa análise, em verde escuro a isoforma canônica de GAS5 e em verde claro as demais isoformas dos transcritos deste gene.*

AL133243.2-201 (ENST00000610331.1) é outra isoforma que aparece superexpressa em cinco das seis séries do nosso estudo, correspondendo a um transcrito humano recentemente descoberto na região de um dos introns do gene *BIRC6* (*Baculoviral IAP Repeat Containing 6*) **(Figura 18)**. Tanto essa isoforma quanto o gene *BIRC6* nunca foram mencionados na literatura como relacionados ao tratamento com metformina.

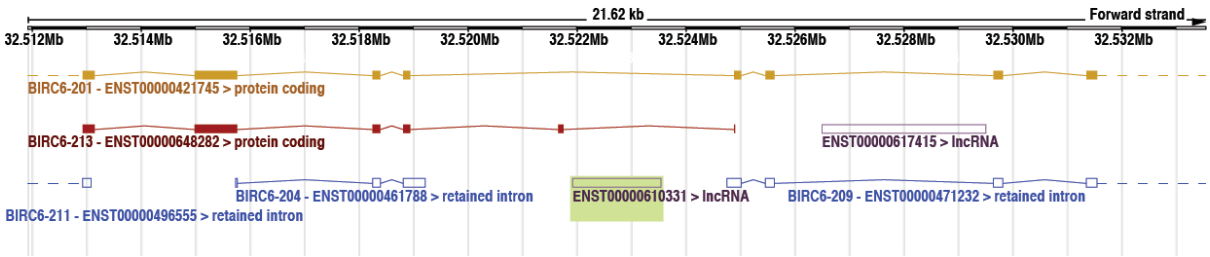


***Figura 18:** Região do transcrito AL133243.2-201 no Ensembl Genome Browser. O transcrito está marcado em verde. As cores das isoformas se referem ao consorcio que as anotou pela*

*primeira vez: em amarelo isoformas anotadas pelo consórcio HAVANNA, em vermelho isoformas anotação pelo GENCODE e em azul isoformas anotadas pelo refseq.*

Notadamente, AC016831.6-205 (ENST00000643779.1) é a única isoforma de lncRNA que está diferencialmente expressa em todas as séries analisadas, sendo este um transcrito recentemente anotado na fita *forward*. A única menção prévia na literatura ao gene de quem esse transcrito é proveniente vem de um estudo de 2018 que o identificou como superexpresso em um transcriptoma global de pacientes com osteoartrite (YANG, G. *et al.*, 2021). Quando a região ao redor do AC016831.6-205 é analisada **(Figura 19 A)** podemos ver que ele se sobrepõe quase totalmente com transcritos de outro lncRNA, o LINC-PINT (*Long Intergenic Non-Protein Coding RNA, P53 Induced Transcript*). Esse lncRNA é muito explorado desde 2013, inclusive havendo uma revisão sobre ele publicada recentemente (Bukhari, Khan et al. 2022), que destaca seu papel como possível regulador de diversas funções como interações RNA-proteína, como esponja de microRNAs e modulação epigenética. Em mais detalhe, um estudo de 2018 demonstrou que uma das isoformas do *LINC-PINT*, o LINC-PINT-208 pode codificar um peptídeo a partir de seu segundo exon por *back-splicing* seguido de tradução a partir de IRES (*internal ribossome entry sites*) e que sua superexpressão foi considerada um marcador de melhor prognóstico em certos tipos de câncer (ZHANG, MAOLEI *et al.*, 2018). Interessantemente, a isoforma AC016831.6-205 que está superexpressão em todas as séries também compartilha o mesmo exon dois do LINC-PINT-208, exon que não é compartilhado pelas demais isoformas de LINC-PINT **(Figura 19A e B)**.

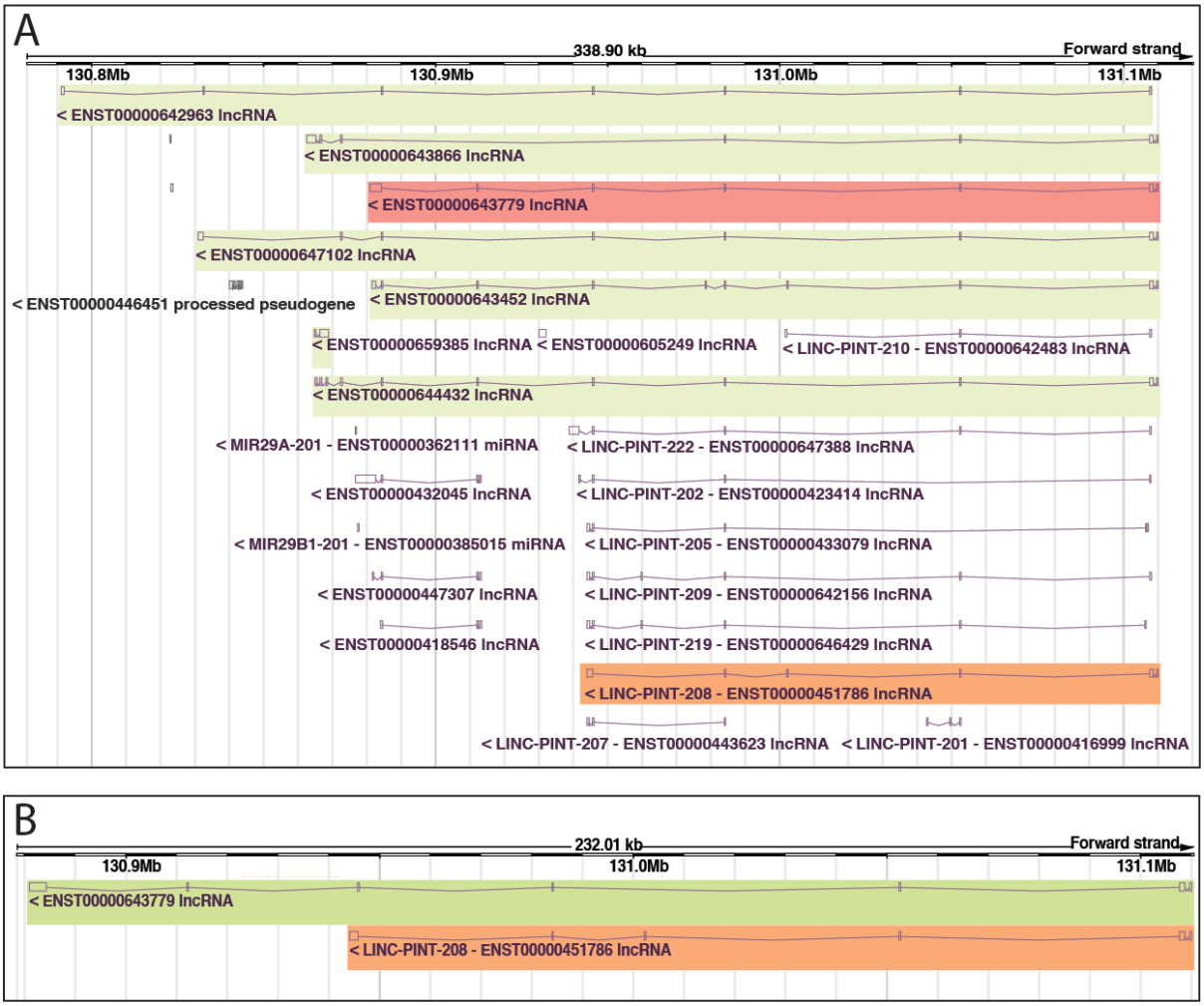


***Figura 19:*** *Região ao redor dos transcritos de AC016831.6-205 e sua sobreposição com LINC-PINT. Em* ***A*** *região maior de alguns dos transcritos de LINC-PINT e em* ***B*** *uma aproximação da sobreposição de LINC-PINT-208 e AC016831.6-205. Em vermelho a isoforma encontrada de AC016831.6 em nossos resultados, em laranja a isoforma 208 de LINC-PINT, em verde na A as demais isoformas de AC016831.6 e em verde na B a isoforma encontrada em nossos resultados do mesmo.*

Além destes seis transcritos que aparecem diferencialmente expressos em quatro ou mais séries, procuramos caracterizar os demais 36 como isoformas de lncRNA com possibilidades de agir em *cis*, na mesma região em que são transcritos, ou *trans*, em regiões mais distantes daquelas a partir de onde foram transcritos, ou mesmo após terem sido traduzidos em pequenos peptídeos.

### 4.5 Isoformas com possível ação em *cis*

Para identificarmos isoformas que poderiam agir em *cis*, primeiro buscamos outros transcritos, não lncRNA, que estivessem sendo diferencialmente expressos junto com os 36

transcritos de lncRNA encontrados na nossa análise e, em seguida, utilizamos um filtro de região genômica para verificar se estes pares correlacionados estavam próximos ou não das isoformas de lncRNA. Para tanto, realizamos análises de correlação utilizando os valores de TPM para verificar outros transcritos que estivessem correlacionados positivamente ou negativamente com 36 lncRNAs. Posteriormente filtramos os pares de correlação presentes na região genômica de 1 megabase para cada lado do transcrito. Essa região máxima foi selecionada de acordo com os valores máximos quanto com o tamanho médio máximo de TADs que são encontrados na literatura (WINICK-NG *et al.*, 2021).

No total foram 24.493 pares de correlações entre todos os transcritos caracterizados como codificadores de proteínas que estão diferencialmente expressos (pvalor < 0,05 & log2FC absoluto > 0.5) nas seis séries e todas as isoformas de lncRNA diferencialmente expressas. Destes 24.493, 2.510 passaram no corte de pvalor e valor de correlação maior que 0,01 e 0,8, respectivamente **(Figura 10)** e 1.930 destes são pares que contém algumas das 36 isoformas de lncRNA selecionadas para análise posterior. Das 1.930, 11 foram correlações inversas e 1.919 correlações diretas.

Quando separamos as 1.930 correlações de acordo com as isoformas que as incluem **(Figura 20)** é possível perceber que um número pequeno de isoformas possui correlação significativa com muitos transcritos, com os LINC00511-243 e o LINC00511-283 possuindo o maior número de correlações, seguindo pelo AC125807.2-201 e o AC092279.1-238. Sete das 36 isoformas não possuem nenhuma correlação significativa, assim foram excluídas do gráfico, sendo essas MAP4K3-DT-217, ZBTB11-AS1-201, SNHG15-214, NEAT1-202, MALAT1-205, AC010198.2-201 e SNHG16-204.

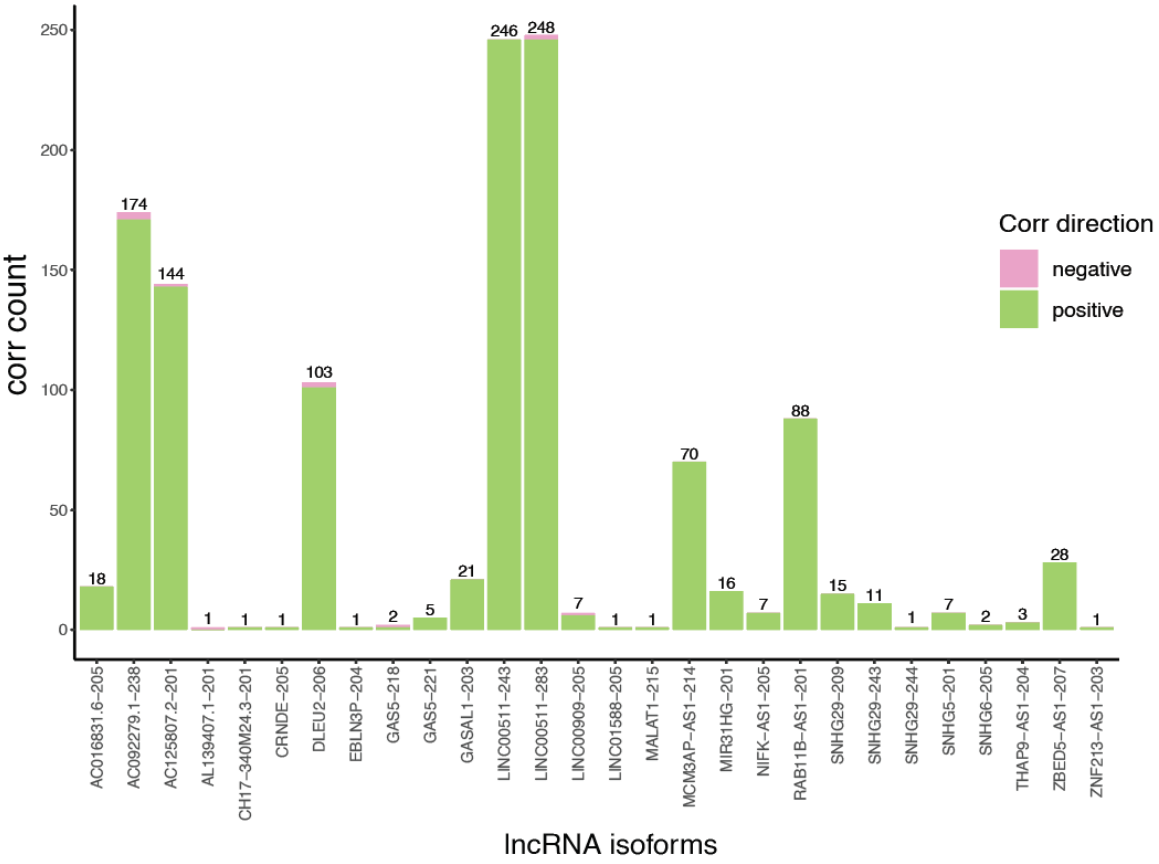


***Figura 20:*** *Contagem de pares de correlação que passam nos valores de corte por isoforma de lncRNA selecionada. Em verde as correlações positivas e em rosa correlações negativas. Na figura estão presentes isoformas, entre as 36 selecionadas, que possuíam correlações significativas (pvalor < 0.01 e valor de correlação absoluto > 0.8) em ao menos quatro séries.*

Independentemente foi feita a análise de número de transcritos que ocupavam posição genômica ao redor do gene de cada uma das 36 isoformas **(Figura 21)**. Isoformas de transcritos diferentes de um mesmo gene não aparecem na figura já que, por ocuparem a mesma região genômica teriam a mesma contagem de transcritos ao redor. É possível se notar que alguns genes estão em regiões muito populadas de transcritos ao redor, como o *NEAT1* e o *MALAT1*, que estão em uma região com mais de 400 transcritos ao redor e são os dois lncRNAs mais comentados na literatura, com 1967 publicações que citam *MALAT1* no PubMed e 1084 que citam *NEAT1* em uma pesquisa feita em 29 de junho de 2022.

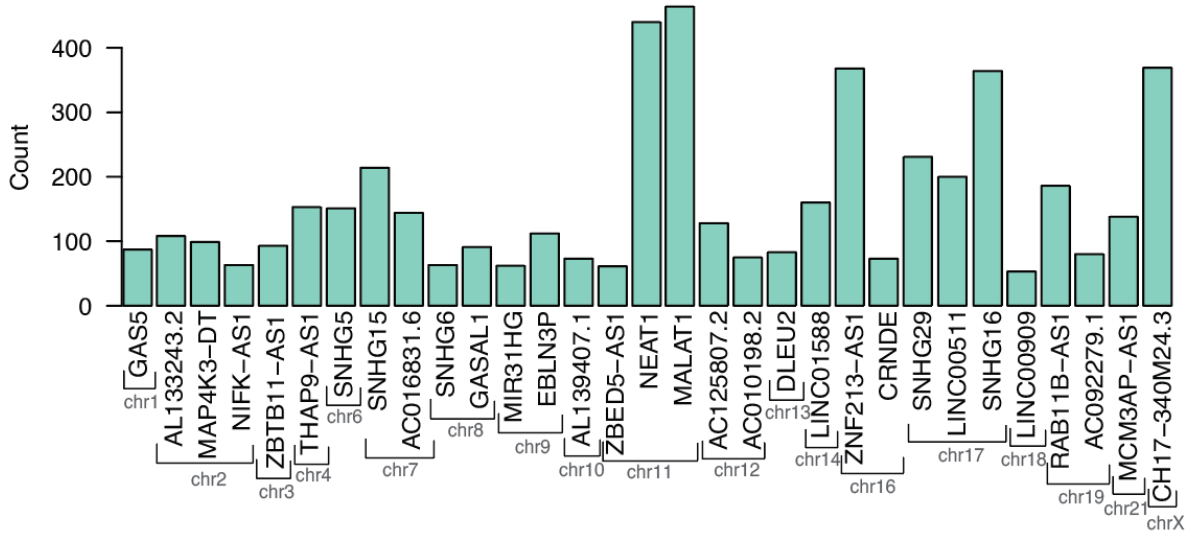


***Figura 21:*** *Número de transcritos presentes na região de 2Mb, 1 Mb para cada lado, ao redor de cada lncRNA analisado, ordenado por localização cromossômica. Na figura estão presentes os genes originários das 36 isoformas de lncRNA selecionadas para análise posterior.*

Mesmo lncRNAs nos maiores cromossomos, chr1 e chr2, estão em regiões pouco populadas como o GAS5, o AL133243.2, o MAP4K3-DT e o NIFK-AS1, todos com um valor próximo de 100 transcritos ao redor. Nos chama a atenção que ainda existam genes não nomeados no genoma humano, como o CH17-340M24.3, que estão em regiões extremamente populadas no cromossomo X.

Os resultados da análise de correlação foram então filtrados utilizando os resultados da análise de distância genômica. Foi feita uma avaliação se cada transcrito que aparecia como um par de correlação significativa estava na região genômica ao redor do gene cuja isoforma era seu par, em busca de possíveis isoformas de lncRNAs agindo em *cis*.

Após a aplicação desse filtro, apenas dois pares que também estavam na região genômica foram encontrados: AC125807.2-201 e FOXM1-202 (ENST00000359843); e ZBED5-AS1-207 (ENST00000664276.1) e EIF4G2-204 (ENST00000530211.6). Ambos os pares possuem correlação positiva e os transcritos estão subexpressos na análise de expressão diferencial **(Figura 22)**. AC125807.2-201 é um lncRNA recentemente anotado no cromossomo 12 que, por enquanto, possui apenas uma isoforma anotada, a -201. Ele aparece como subexpressos em todas as séries, mas somente passa nos valores de corte em cinco das seis séries **(Figura 22, painel superior)**. Essa isoforma está coexpressa e na mesma região

genômica da FOXM1-202 (*Forkhead box M1* isoforma 202) um transcrito codificador de proteína de um fator de transcrição que está comumente superexpresso na mitose e na progressão de diversos cânceres (KOPANJA *et al.*, 2022; LA *et al.*, 2022; ZHANG, X. *et al.*, 2022). Já o segundo par de correlação **(Figura 22, painel inferior)**. ZBED5-AS1-207 (*Zinc finger BED-containing 5*) é um lncRNA recém anotado no cromossomo 11 antisenso ao gene *ZBED5*, os genes da família ZBED são genes derivados de um transposon que sofreu duplicações ao longo da evolução (HAYWARD *et al.*, 2013); esse lncRNA não possue literatura associada e o ZBED5 em específico não foi abordado em resultados possivelmente relacionados a metformina até o momento. EIF4G2 (*Eukaryotic translation initiaion fator 4 gamma 2)* é um membro da família de fatores de iniciação da tradução (LEWIS *et al.*, 2008), essa família costuma estar superexpressa em condições de estresse, em especial no câncer (LI, S. *et al.*, 2021), no entanto eles também ainda não foram associados à metformina.

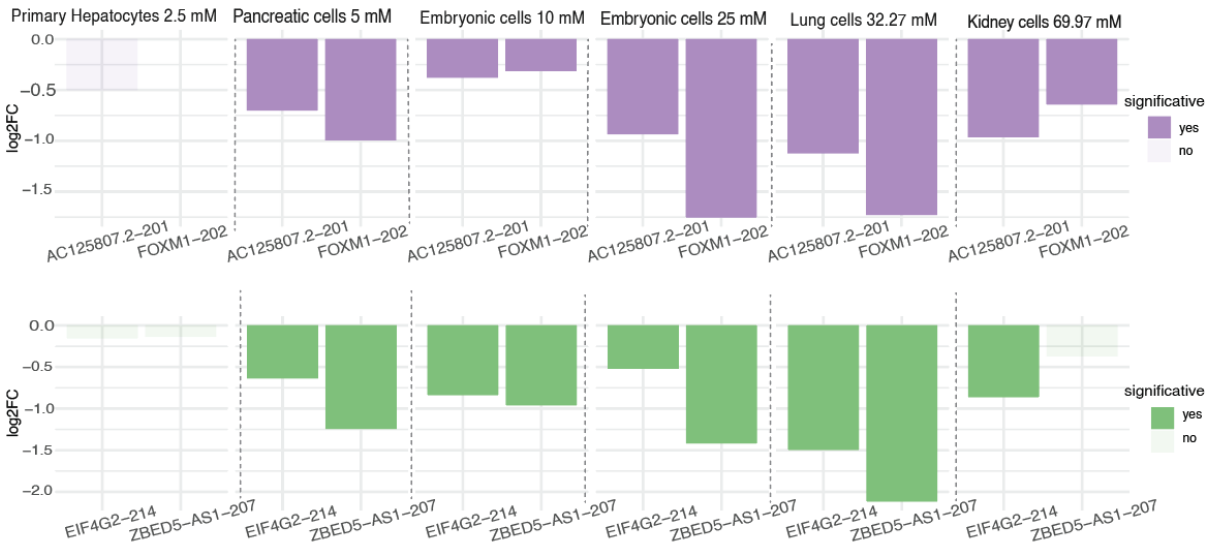


***Figura 22****: log2FC em cada série dos pares de isoforma de lncRNA-mRNA que possivelmente agem em cis com o valor de significância como a transparência. No topo em roxo os valores de log2FoldChange do par AC125807.2-201 e FOXM1-202 por série e abaixo em verde os valores de log2FoldChange do par EIF4G2-214 e ZBED5-AS1-207.*

Observando a **figura 22** é possível notar que em todas as séries os transcritos têm tendencia a subexpressos nos dois pares, mesmo nas quais eles não possuem valor de significância para serem considerados diferencialmente expressos, como os hepatócitos primários ou nas células renais. Ambos os pares de transcritos possuem a maior diferença de expressão em relação ao controle nas células epiteliais pulmonares, que também é uma das

duas séries com maior número de reads. A partir dessas informações nós analisamos profundamente a região genômica que contêm esses pares para verificar se informações advindas do uso de outros tipos de técnicas suportavam a teoria de que essas isoformas de lncRNA (AC125807.2-201 e ZBED-AS1-207) poderiam estar regulando em *cis* os seus pares (respectivamente, FOXM1-202 e EIF4G2-204). Utilizamos visualizações no UCSC e no *Ensembl Genome Browser* para tal.

Quando observamos a região genômica em que está o par de transcritos AC125807.2-201/FOXM1-202 podemos ver que eles estão a cerca de 200 kilobases um do outro **(Figura 23)**. Na *track* do UCSC *Genome Browser* de captura de conformação cromossômica (HI-C) é possível perceber uma clara região triangular vermelho escura mostrando os prováveis mapas de contato de cromatina entre o AC125807.2-201 e o FOXM1-202, a borda de um TAD. Além disso, na *track* de elementos regulatórios é possível perceber que o AC125807.2-201 está, não só em cima, mas seu único exon engloba toda uma região de marcação de CTCF (*CCCTC-Binding Factor*). Essa presença do lncRNA na borda de um TAD e em uma marcação de CTCF pode indicar a possível ação do AC125807.2-201 como um RNA que é ponto de ancoragem transcricional (*topological anchor point RNA*) os tap-RNAs, uma nova classe proposta de RNA que estão localizados em pontos de ancoragem da cromatina para a formação de alças e nas bordas de TADs(AMARAL *et al.*, 2018).

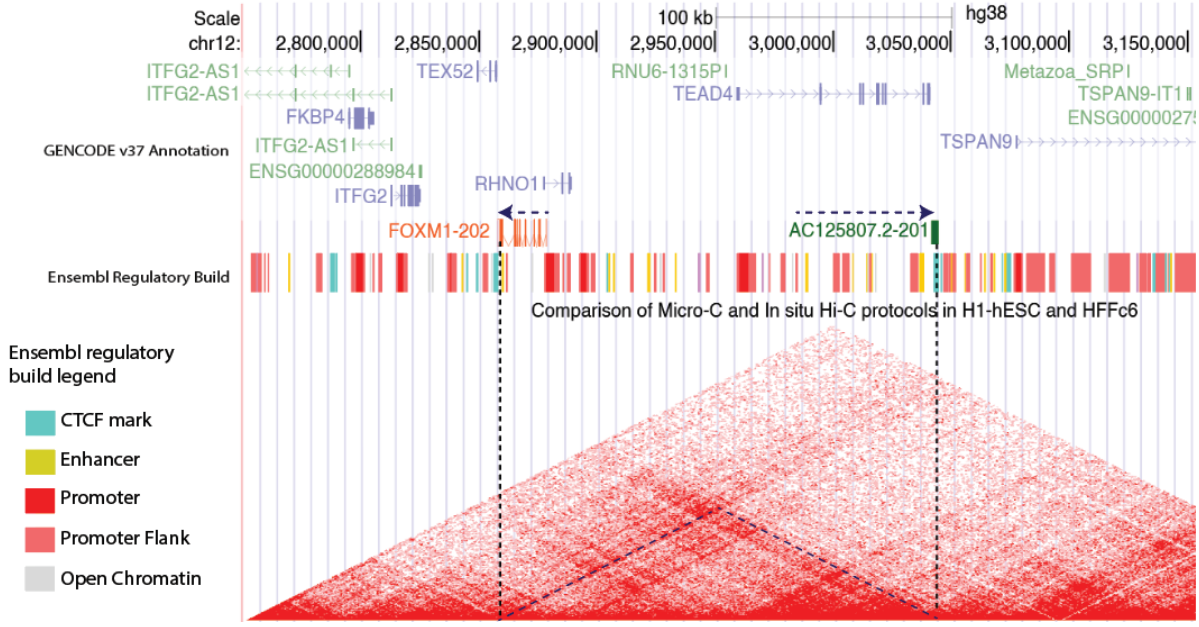


***Figura 23:*** *Figura combinada do UCSC Genome Browser e do Ensembl Genome Browser da região ao redor do AC125807.2-201. De cima para baixo*: a track de distância genômica, que mostra em que cromossomo e a que distância os transcritos estão localizados, a anotação na

versão 37 do GENCODE onde introns são representados como linhas e exons como blocos, a track regulatória do Ensembl com a legenda na lateral, e a track de dados de HI-C. Em azul escuro linhas tracejadas mostrando o TAD entre os transcritos e uma seta representando a direção da expressão de cada.

A isoforma FOXM1-202 já foi achada como subexpressa no tratamento com metformina, o que protegeria contra a proliferação de fibroblastos em um modelo de fibrose pulmonar, dependentemente da AMPK (GU *et al.*, 2021). Nossos resultados apontam para a ação de AC125807.2-201 como um possível regulador em *cis* da expressão de FOXM1-202 e ambos sendo subexpresso pelo tratamento com metformina.

O outro par de isoforma de lncRNA-mRNA que obtivemos em nossos resultados como possível ação em *cis* é o ZBED-AS1-207 EIF4G2-214. Quando observamos a região genômica dos dois **(Figura 24)**, eles estão bem próximos sendo separados por menos de 70 kilobases. Na *track* de HI-C do UCSC *Genome Browser* também é possível verificar um claro TAD em cujas bordas os transcritos estão. Mais interessante ainda é a *track* de Interações gênicas do GeneHancer, onde aparecem interações claras dos três exons do ZBED-AS1-207 e do primeiro exon do EIF4G2-214. Na track de elementos regulatórios do Ensembl, o EIF4G2-241 se encontra, quase que inteiramente, em uma região de promotor, o primeiro exon do ZBED-AS1-207 em uma marcação de CTCF e seu segundo exon em uma possível região promotora.

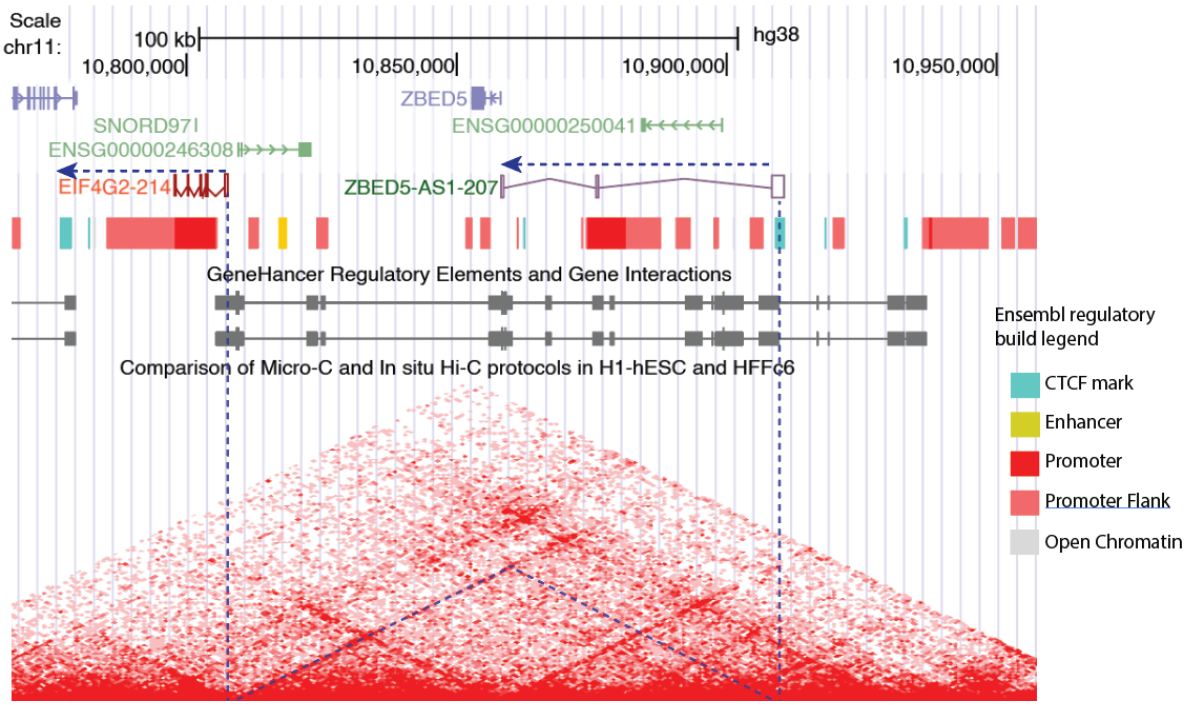


***Figura 24:*** *Figura combinada do UCSC Genome Browser e do Ensembl Genome Browser da*

*região ao redor do ZBED-AS1-207. De cima para baixo*: a track de distância genômica, que mostra em que cromossomo e a que distância os transcritos estão localizados, a anotação na versão 37 do GENCODE onde introns são representados como linhas e exons como blocos, a track regulatória do Ensembl com a legenda na lateral, a track de interações cromossomais de curta distância do geneHancer, e a track de dados de HI-C. Em azul escuro linhas tracejadas mostrando o TAD entre os transcritos e uma seta representando a direção da expressão de cada.

A expressão diferencial de EIF4G2 ainda não foi associada com metformina na literatura, no entanto, sua expressão diferencial já foi associada a dois fenótipos que a metformina é proposta como regulador: o remodelamento de matriz extracelular e a proliferação de células de tumor por microRNAs que possivelmente regulam EIF4G2. No primeiro caso, encontraram o microRNA miR-10-3p controlando a positivamente a expressão de EIF4G2. Quando o microRNA ou a EIF4G2 são suprimidos, a expressão do outro também reduz, melhorando a efetividade do tratamento com células ósseas mesenquimais (BMSCs, *Bone mesenchymal stem cell*) (WANG, Z. *et al.*, 2021). Já no segundo caso, a superexpressão do microRNA miR-144 inibe a expressão de EIF4G2, sendo que a superexpressão desse gene é verificada em carcinoma hepatocelular (LI, S. *et al.*, 2021).

Em ambos os genes FOXM1 e EIF4G2 não foram analisados na literatura em nível de transcrito, até o momento. A isoforma FOXM1-202 é a isoforma canônica anotada que codifica a proteína canônica associada ao gene FOXM1 **(Figura 25)**. Este não é o caso do EIF4G2-214, que é uma isoforma com o mesmo promotor e exon 1 da canônica, mas que codifica um transcrito muito menor se comparado a esta **(Figura 26)**.

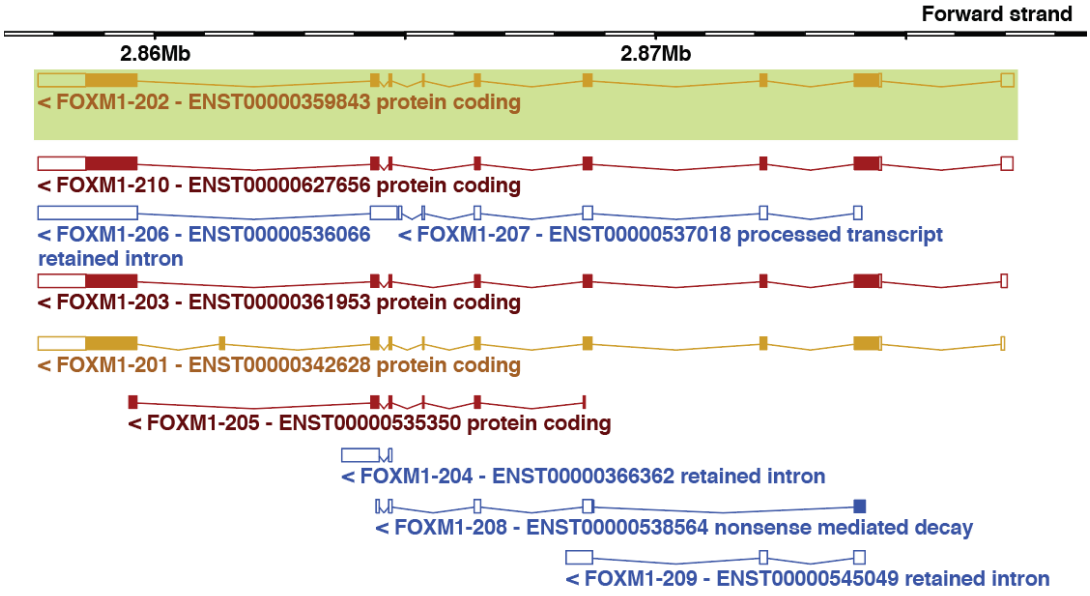


***Figura 25:*** *Isoformas do gene FOXM1 com seus éxons e íntrons. As cores das isoformas se referem ao consorcio que as anotou pela primeira vez: em amarelo isoformas anotadas pelo consórcio HAVANNA, em vermelho isoformas anotação pelo GENCODE e em azul isoformas anotadas pelo refseq. A marcação em verde se refere à isoforma que encontramos nesse estudo e que também é a isoforma canônica. Figura retirada da região associada ao FOXM1 no Ensembl Genome Browser.*

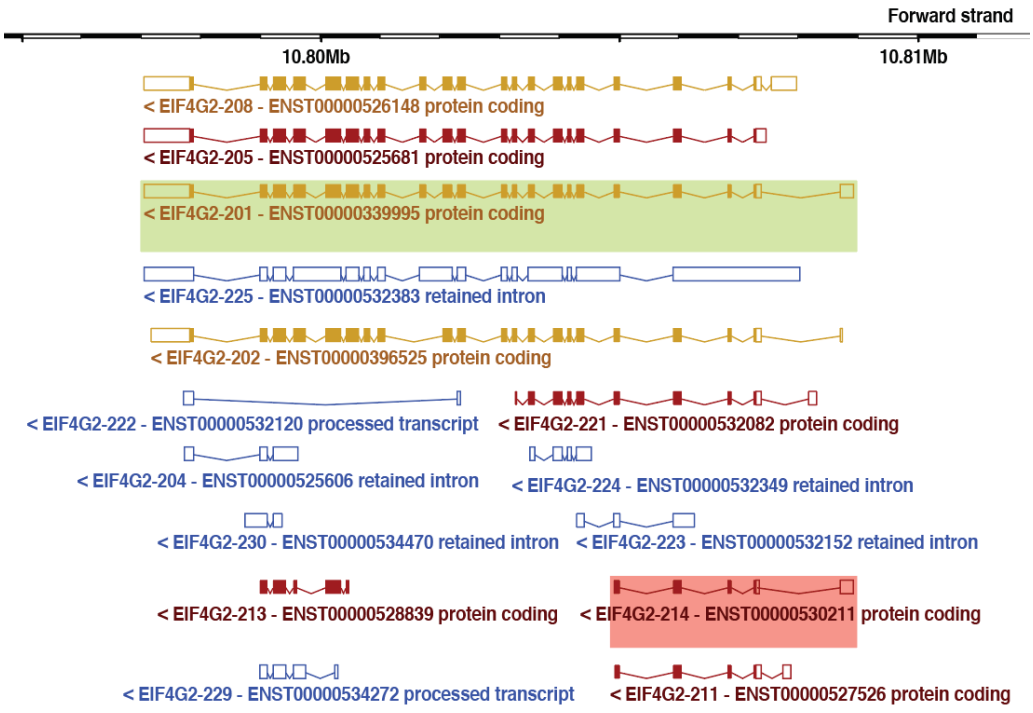


***Figura 26:*** *Algumas isoformas de transcritos do gene EIFG2 com seus éxons e introns. As cores das isoformas se referem ao consorcio que as anotou pela primeira vez: em amarelo isoformas anotadas pelo consórcio HAVANNA, em vermelho isoformas anotação pelo*

*GENCODE e em azul isoformas anotadas pelo refseq. A marcação em vermelho é a isoforma que encontramos nesse estudo, sendo que a em verde é a canônica.*

### 4.6 Isoformas com possível ação em *trans*

Os lncRNAs também podem regular a expressão gênica afetando transcritos em regiões muito distantes do seu loco cromossômico e até mesmo em outros cromossomos (PENG; KOIRALA; MO, 2017). Eles podem agir tanto na forma de RNA maduro, interagindo com proteínas ou como outros componentes da maquinaria celular, como também sendo traduzidos em micropeptídeos pela presença de sORFs (*small ORFs*) ou uORFs (*upstream ORFs*) nas suas sequencias (LÜSCHER-DIAS *et al.*, 2021).

Para verificar a possibilidade de algum de nossos transcritos diferencialmente expressos em cinco ou mais séries do nosso estudo estarem agindo em *trans* como reguladores de mRNAs-alvo, utilizamos uma abordagem de correlação de valores de expressão seguida de enriquecimento funcional de processos biológicos associadas ao fármaco.

A isoforma em que primeiro executamos essa análise foi o AC016831.6-205 por este ser o único transcrito a aparecer como diferencialmente expresso em todas as seis séries. Assim como foi mencionado no primeiro tópico dos resultados, nossas evidências apontam que este transcrito é capaz de codificar um micropeptídeo anti-oncogênico por circularização do seu segundo exon, evento que ocorre com o LINC-PINT-208, isoforma que compartilha este exon com o AC016831.6-205 (ZHANG, MAOLEI *et al.*, 2018).

O resultado do enriquecimento funcional a partir dos dados de correlação do AC016831.6-205 com seus supostos transcritos-alvo codificados de proteína pode ser visto na **Figura 27**. Essa figura foi gerada após a exclusão de processos biológicos não relacionados à ação da metformina.

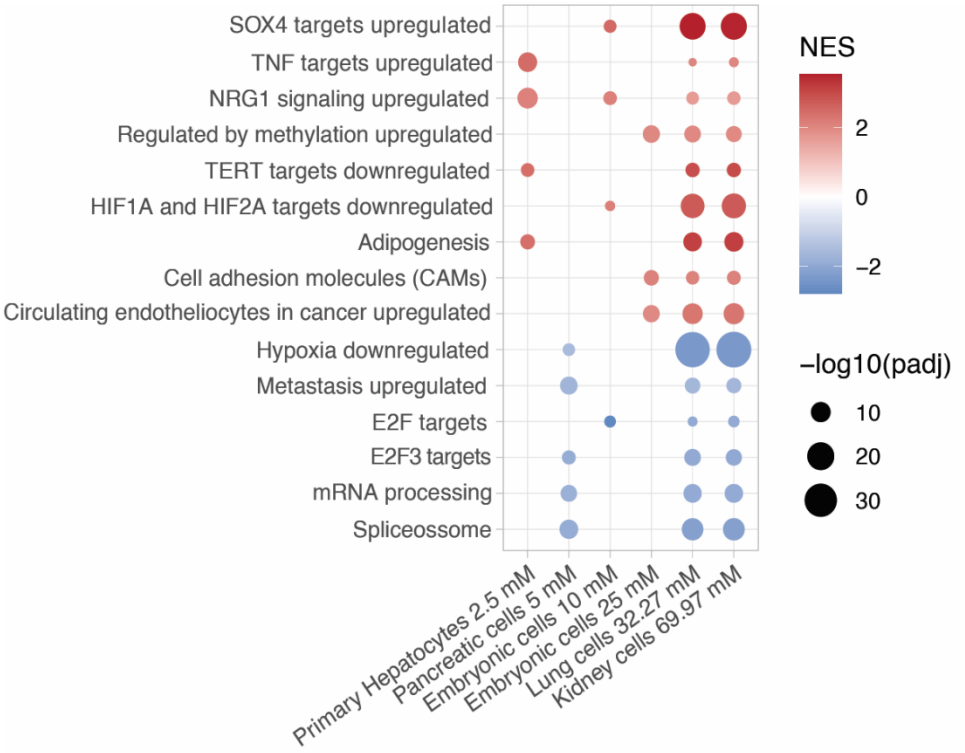


***Figura 27:*** *Enriquecimento funcional de transcritos codificadores de proteínas alvos da isoforma de lncRNA AC016831.6-205. NES é o valor de enriquecimento normalizado associado ao valor de correlação das isoformas de lncRNA com os genes alvo presentes no processo biológico. O tamanho do círuculo se refere a -log10 do p ajustado da análise. As ´series estão presentes no eixo x e os processos biológicos no eixo y.*

Entre os processos biológicos encontrados nessa análise, é possível verificar aqueles associados à hipoxia, sendo que uma das ações mais conhecidas do fármaco é a redução dos efeitos da hipóxia intracelular. As duas vias de hipóxia que aparecem são as de "Alvos de HIF1A e HIF2A subexpressos", que aparece com o NES positivo e "Hipóxia subexpresso" que aparece com o NES negativo. Essas duas vias possuem genes centrais que são alvos da ação da metformina. A primeira dessas vias possui alvos como *EGFR*, *SERPINE1* e *EIF1* e engloba cerca de 150 genes que são alvos de HIFAs, já a segunda via é mais geral e engloba mais de 300 genes que aparecem subexpressos em hipoxia e outros processos celulares como *HMBS* e *PSMG1*.

Outro resultado queque vale a pena ressaltar entre os processos biológicos enriquecidos é o ciclo celular. As vias "Alvos de SOX4 superexpressos", "Alvos de TNF superexpressos", "Alvos de NRG1 superexpressos", que aparecem com NES positivo e as vias "Alvos de E2F" e "Alvos de E2F3", que aparecem com o NES negativo, são todas vias associadas à regulação

positiva do ciclo celular e às porções do ciclo celular que regulam a proliferação exacerbada e a morte de tumores. A via "Metástase superexpressos" também aparece com o NES negativo e está diretamente associada ao fenômeno de morte de células tumorais e regulação do ciclo celular.

O enriquecimento do processo biológico *splicing* nas vias de "processamento de mRNA" e "spliceossomo", ambas com NES negativo, e da regulação epigenética "superexpressos por metilação", também reforça a possível ação desse lncRNA nos processos biológicos conhecidamente afetados por metformina.

Quando fazemos a análise de todos os transcritos que mostram correlação com o AC016831.6-205 e estão diferencialmente expressos em, pelo menos, quatro das seis séries **(Figura 28)** podemos notar vários transcritos cujos genes já foram extensivamente explorados na literatura sobre metformina.

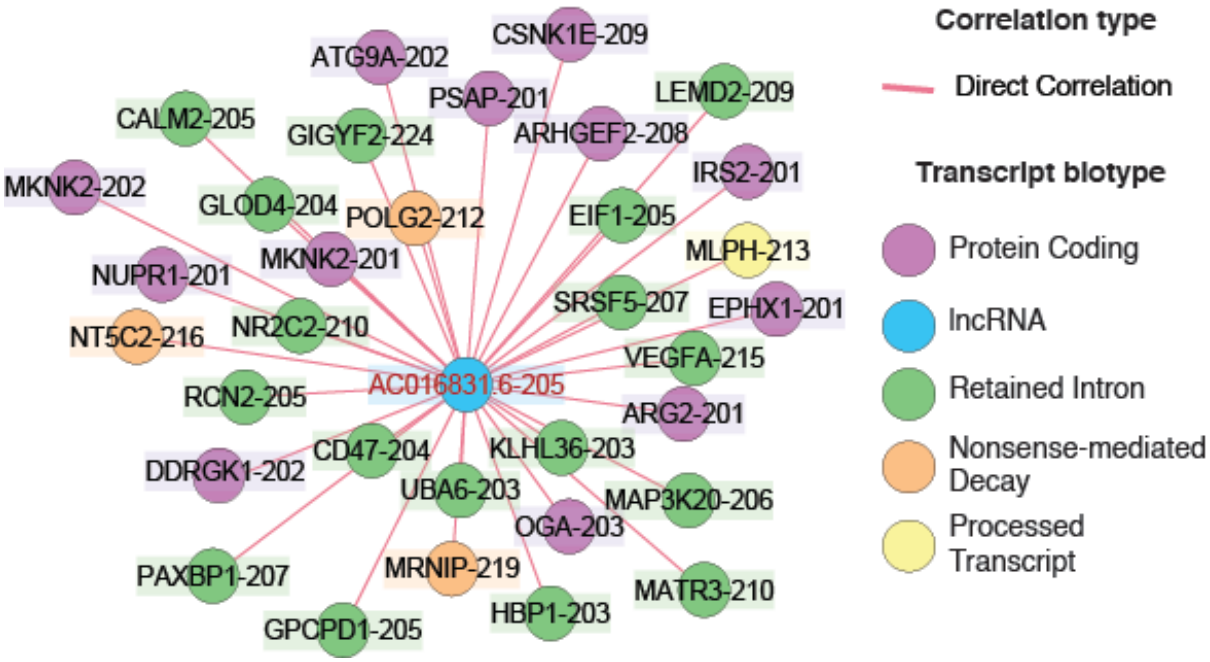


***Figura 27:*** *Rede de interação de supostos transcritos alvo de AC016831.6-205 que também aparecem diferencialmente expressos em mais de quatro séries nesse estudo. Na figura se apresentam transcritos derivados de splicing sendo os codificadores de proteína (em roxo) os produtivos e os intron retido (em verde), decaimento mediado por nonsense (em laranja) e transcrito processado (em amarelo) os não produtivos.*

VEGFA-215 é um dos transcritos não canônicos do gene *VEGFA* (fator de crescimento endotelial vascular A) e a metformina já foi demonstrada como um regulador de VEGFA ativando a expressão de suas isoformas não codificadoras (não produtivas) e, portanto, influenciando no *splicing* destas (YI *et al.*, 2016). Aqui sugerimos que este efeito pode estar relacionado à superexpressão de AC016831.6-205. Outros dois transcritos cujas isoformas não-codificantes estão com correlação positiva com AC016831.6-205, CALM2-205 e MAP3K20-206, também são muito associadas na literatura ao fármaco, no entanto, seus efeitos ainda não foram estudados com enfoque em suas diversas isoformas.

Essa mesma análise de enriquecimento funcional da correlação entre lncRNAs e mRNAs foi executada para as demais isoformas de lncRNAs presentes em cinco ou mais séries, no entanto, o enriquecimento dos alvos de LINC00511-243, LINC00511-283 e AL133243.2-201 não resultou em nenhum processo biológico que passasse no valor de corte de p-ajustado inferior a 0,01. Já os transcritos GAS5-218 e GAS-221 resultaram em apenas um processo biológico enriquecido, para seus supostos alvos (“metabolismo de selenoaminoácidos”) que passava no valor de corte, em quatro das cinco séries em que eles aparecem como diferencialmente expressos. Esse metabolismo consiste no metabolismo de aminoácidos onde o enxofre (S) foi substituído por selênio (Se). Esse metabolismo ainda não foi associado na literatura ao lncRNA GAS5, no entanto, esses selenoaminoácidos já foram analisados como preventivos dos efeitos deletérios de quimioterapia (WEEKLEY *et al.*, 2011), o que pode estar associado aos efeitos já descritos do lncRNA. Os selenoaminoácidos mais comuns são a selenocisteína e a selenometionina que produzem benefícios fisiológicos agindo como agentes antioxidantes (RAHMANTO; DAVIES, 2012).

A outra isoforma presente em cinco ou mais séries cuja análise de enriquecimento produziu resultados que passassem nos valores de corte foi o NEAT1-202. Como mencionado anteriormente, a isoforma que encontramos superexpresso é a isoforma canônica de NEAT1 e também é a maior isoforma de NEAT1 anotada (Figura 14).

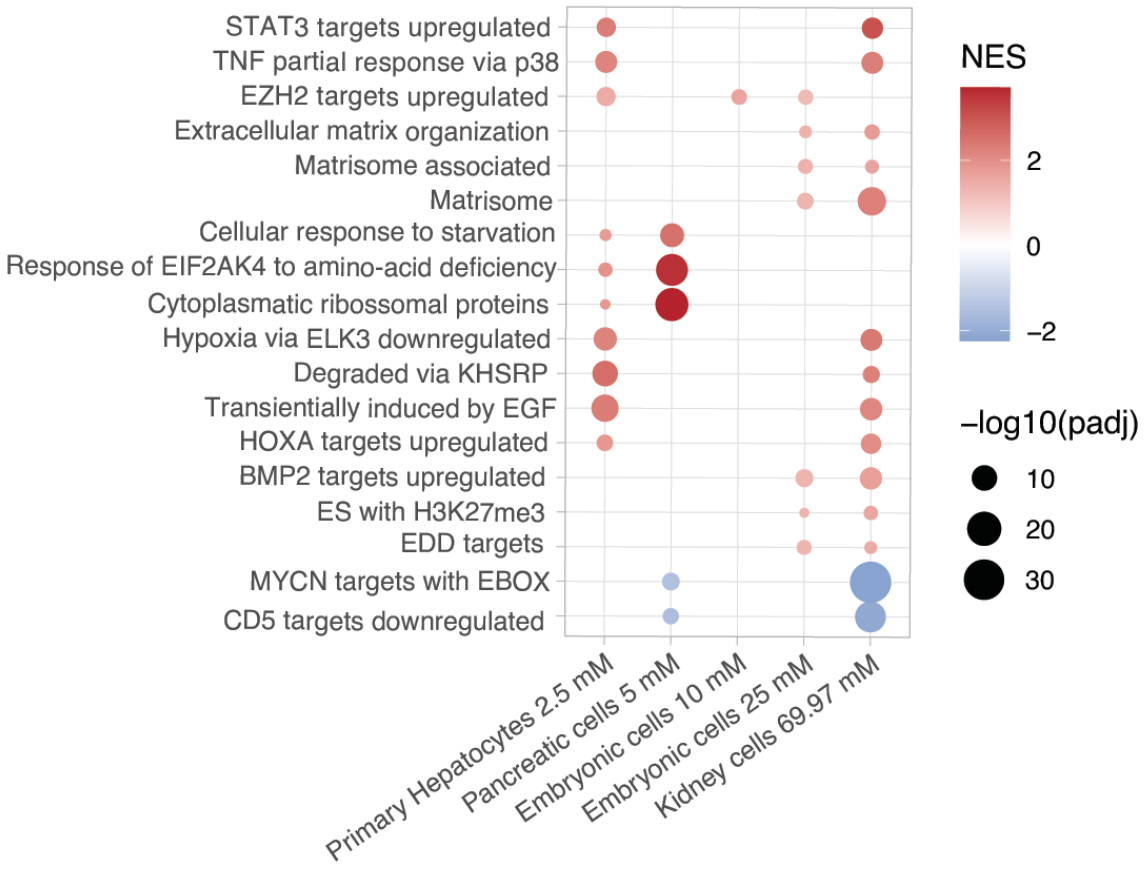


***Figura 29:*** *Enriquecimento funcional de transcritos codificadores de proteínas alvos da isoforma de lncRNA NEAT1-202. NES é o valor de enriquecimento normalizado associado ao valor de correlação das isoformas de lncRNA com os genes alvo presentes no processo biológico. O tamanho do círculo se refere a -log10 do p ajustado da análise. As ´series estão presentes no eixo x e os processos biológicos no eixo y.*

Entre os processos biológicos encontradas enriquecidos na análise com o NEAT1-202 **(Figura 29)**, podemos ver diversos processos ativados por esta isoforma de lncRNA. Os dois únicos processos encontrados negativamente regulados foram “Alvos de MYCN com E-BOX” e “Alvos de CD5 subexpressos”. A primeira dessas vias contém genes da família MYC de oncogenes, que são genes com papel importante no controle de divisão celular e apoptose. Já a segunda via possui diversas citocinas e quimiocinas como alvo, sendo que a regulação negativa dessas citocinas leva à menor sobrevivência das células (GARY-GOUY *et al.*, 2007).

Processos biológicos que complementam a ação do fármaco e somente apareceram no enriquecimento do NEAT1-202 foram as vias relacionadas ao remodelamento de matriz extracelular: “Matrisoma”, “Associado ao matrisoma” e “Organização de matriz extracelular”;

todas as três aparecendo como superexpressas no enriquecimento. Diversos ativadores de AMPK, como a metformina, atuam no remodelamento anormal da matriz extracelular, diminuindo efeitos de fibrose (WU *et al.*, 2021), com a redução da deposição de colágeno (LUO *et al.*, 2016).

O gene NEAT1 já foi previamente associado com remodelamento de matriz com efeitos conflitantes a depender do método de estudo, com diversos artigos relatando a progressão a fibrose via NEAT1 e com outros relatando diminuição da fibrose e proliferação via NEAT1 (LIU, F. *et al.*, 2020; TODOROVSKI; FOX; CHOI, 2020; WANG, X. *et al.*, 2019; YANG, J. *et al.*, 2021). No entanto, como no caso anterior, estudos específicos de degradação de matriz não analisaram NEAT1 em nível de isoforma.

Outros processos biológicos encontrados também se relacionam à proliferação celular, como: "Alvos de EZH2 superexpressos" e "Transativado por EGF", que são os dois mais específicos e ambos atuando na manutenção do estado padrão de expressão gênica em células durante diferentes gerações. Processos biológicos associados à morte, hipóxia e reparo também consistem desse mesmo caso e já foram muito abordados na literatura sobre o fármaco e o lncRNA (CHEN *et al.*, 2018; SIVALINGAM *et al.*, 2020; ZHANG, P. *et al.*, 2019; ZHANG, Q. *et al.*, 2020), mas nunca em nível de transcrito. Finalmente, NEAT1-202 não apresentou mRNAs correlacionados em comum entre, ao menos, quatro das seis séries. Assim, sua análise como potencial fator em *trans* foi feita somente a partir do enriquecimento funcional.

**5. Discussão**

RNAs longos não-codificadores já foram estudados diversas vezes como reguladores finos da expressão gênica e são algumas das moléculas mais abordadas no contexto da biologia molecular médica. Já foram associados com quase todos os processos celulares existentes e diversas revisões mostram que, muitas vezes, quando o mesmo lncRNA é analisado na mesma condição e com técnicas experimentais similares, os resultados divergem (LI, Z.-X. *et al.*, 2018; RAJAGOPAL *et al.*, 2020; XING, C. *et al.*, 2021). Mesmo com toda a literatura de estudos primários (originais) disponível, existem pouco mais de 200 revisões sistemáticas seguidas ou não de meta-análise sobre efeitos de um lncRNA específico em uma condição, ou do estudo de uma condição em busca de lncRNAs como biomarcadores no PubMed e esse número é bem menor em outros agregadores de artigos. No PROSPERO, um agregador de protocolos de revisão sistemática, apenas 118 protocolos abordam o tema e grande parte (82) dos artigos ainda não foi publicada (Tabela 3). No maior agregador de estudos clínicos

existente, o Clinicaltrials.gov, existem apenas 70 estudos clínicos iniciados, em processo ou completados que abordam lncRNAs, sendo que destes, nenhum procura associar expressão de lncRNA com resposta à fármacos.

*Tabela 3: Resultados da busca por revisões sistemáticas seguidas ou não de meta-análise e lncRNAs em vários bancos de dados*

| **Agregador de artigos** | **Número de itens encontrados** |
|---|---|
| PubMed/ncbi | 211 |
| Scopus | 58 |
| Lilacs | 6 |
| PROSPERO | 118 |

Esse vazio na literatura de estudos secundários e terciários se devem à problemas encontrados na literatura dos lncRNA e nas técnicas utilizadas para detecção de sua ação, já abordados em outros artigos (ALI; GROTE, 2020). Um ponto importante para reflexão é de que a nomenclatura de lncRNAs não é padronizada (BRUFORD *et al.*, 2020) e existem diversos nomes para um mesmo gene codificador de lncRNA, sem nenhum padrão. Três exemplos são os lncRNA nomeados *HULC* (SHARMA; TRIPATHI; DAS, 2019), *PAAN* (WANG, J. *et al.*, 2018) e *HEAL* (CHAO *et al.*, 2019), que na literatura já são nomeados de outras formas, o que faz com que se seja muito difícil, ou mesmo impossível, identificar o nome original do lncRNA e, efetivamente, o que há de informação disponível sobre ele (LÜSCHER-DIAS *et al.*, 2021). Nessa dissertação e no artigo associado, no intuito de padronizar a leitura e referência posterior, o nome de cada isoforma foi padronizado ao redor de seu nome de transcrito proveniente da anotação do GENCODE e entre parênteses, após o nome em sua primeira menção no texto, também existe o código ENSEMBL associado à isoforma, o que permite uma busca isoforma específica no banco de dados.

A forma com que cada isoforma de lncRNAs está associada à um gene também é um ponto para ser levado em consideração. Com novas versões da anotação do transcriptoma humano saindo, em média, a cada 6 meses no GENCODE, as mudanças mais constantemente observadas são a anotação de novas isoformas e a associação destas a genes (FRANKISH *et al.*, 2019). A descrição de isoformas na literatura também é bastante confusa, com os poucos lncRNAs cujas isoformas são abordadas independentemente, como o NEAT, sendo descritas apenas como isoforma 1 e isoforma 2 (NEAT1_1 e NEAT1_2) (ISOBE *et al.*, 2020; SUN *et*

*al.*, 2021) e sem o arquivo de sequência associado, o que prejudica a referenciação a longo prazo e a identificação da isoforma específica a qual os artigos estão se referindo.

O caso do AC016831.6-205 foi um caso em que entrei em contato com a equipe de anotação do GENCODE porque, pelos critérios de anotação adotados, ele deveria ser anotado como uma isoforma de LINC-PINT com uma região extra à jusante do gene, conhecido como DOG, do inglês *downstream of gene*. No momento que entrei em contato, o transcrito AC016831.6-205 na anotação mais recente que havia saído há poucas semanas (GENCODE 40) havia mudado de nome para RP1136B6.2. Eu obtive resposta da equipe do GENCODE e fui informada de que nas próximas versões este transcrito será anotado como isoforma de LINC-PINT (Apêndice 2).

A falta de análises em nível de transcrito, em geral, também é um grande fator de confusão em resultados de expressão de lncRNA, pois diferentes isoformas possuem papeis diferentes dentro da célula (DUMBOVIĆ *et al.*, 2021; ZIEGLER; KRETZ, 2017), o que afeta toda a análise e interpretação posterior dos resultados de expressão (YIP *et al.*, 2021). Análises de expressão em nível de isoforma ainda são a minoria, até mesmo para genes codificadores de proteínas e ainda não existe um padrão ouro de técnicas de análise bioinformáticas em nível de isoformas, com os programas existentes tendo muitas falhas. Alguns exigem transcriptomas desnecessariamente profundos para serem efetivamente executados (SHEN *et al.*, 2014) e outros possuem métricas rígidas e discutíveis do que se caracteriza como um efeito de *splicing* alternativo (GUO *et al.*, 2021; VITTING-SEERUP; SANDELIN; BERGER, 2019). Entretanto, o maior problema desses programas é não dialogar com a anotação de isoformas curada do GENCODE, ou NCBI, ou entre si, fazendo com que análises com um dos programas não tenham o mesmo resultado que os outros e que não seja possível utilizar a anotação referência como comparativo.

Existe apenas uma revisão na literatura que aborda os efeitos do *splicing* na ação de lncRNA e que estes sofrem um *splicing* mais permissivo, o que pode explicar o grande número de isoformas que possuem (STANĚK, 2021), o que é muito pouco se comparado à toda a literatura disponível dos efeitos biológicos da expressão de lncRNA.

Esse problema de falta de ferramentas e de literatura sobre isoforma de transcritos em geral e sobre lncRNA está relacionado ao fato de que a cinética por trás do processamento de RNA ainda não foi completamente elucidada com artigos recentes ainda abordando o tempo que um transcrito leva para sofrer *splicing*, ser capeado e sair do núcleo apresentando resultados

sobre meia-vida de RNAs, como eles agem na célula e do tempo levado para o processo de *splicing* (PAI *et al.*, 2017, 2018).

Outro processo ainda não completamente elucidado que foi tratado tanto nos resultados de ações em *cis* de lncRNA, pela possível influência da expressão de ZBED5-AS1-207 na expressão de EIF4G2 **(Figura 22)**, quanto de ações em *trans* com a codificação por IRES de um peptídeo pela circularização de um exon de AC016831.6-205 **(Figura 30)** é o processo de regulação da tradução. O EIF4G2, cujo nome de gene oficial anterior é NAT1, é um homólogo do fator canônico de iniciação da tradução EIF4G1 e, mesmo não fazendo parte do processo canônico de tradução, sua deleção total inviabiliza a embriogênese (YAMANAKA, 2000) e deleção parcial reduz a atividade das vias de Akt e Erk (SUGIYAMA *et al.*, 2017). Estudos recentes associam a presença de EIF4G2 ao escaneamento alternativo do início do transcrito pela subunidade 40S ribossomal e resgate da tradução relacionada às ORFs à montante (*upstream* ORFs) e à tradução sob estresse (LEE, S. H.; MCCORMICK, 2006; SMIRNOVA *et al.*, 2022). Como já mencionado, a expressão diferencial de EIF4G2 ainda não foi associada à metformina, no entanto, estudos já abordaram a influência do fármaco no processo de tradução, mesmo que não de forma aprofundada (BUIST; FUSS; RASTEGAR, 2021; HAN *et al.*, 2019).

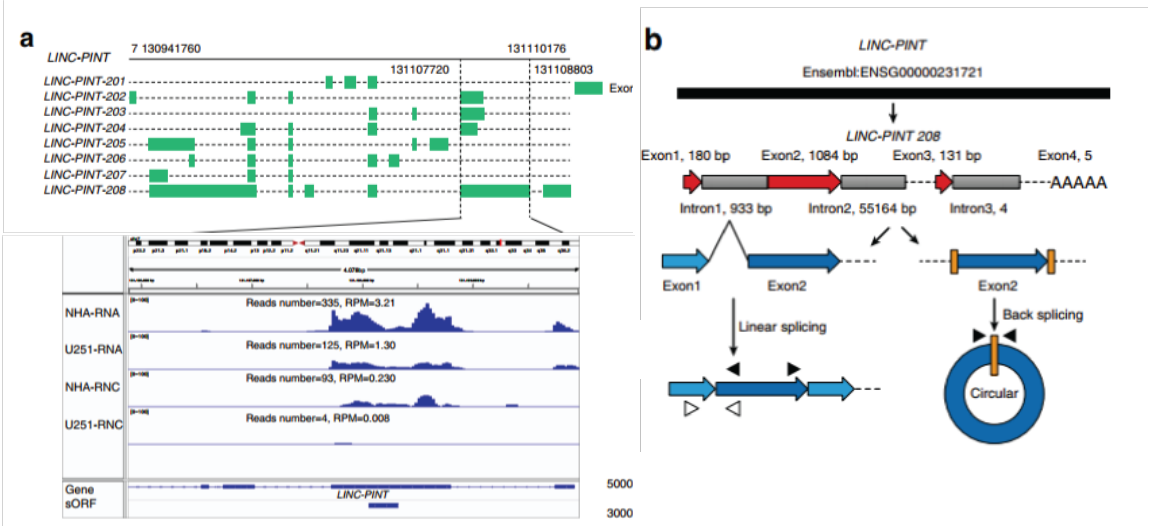


***Figura 30:*** *Região genômica do lncRNA LINC-PINT (A) e circularização de seu segundo exon por backsplicing (B). Em A as diferentes isoformas de LIC-PINT comparadas e no detalhe a visão do segundo exon em cada uma delas. Em B um modelo esquemático de como a circularização do segundo exon de LINC-PINT ocorre. Editado. Fonte: Zhang (2018)*

Os processos alternativos de tradução foram evidenciados nesse estudo de duas formas: (1) dependente de IRES, com a associação direta de um lncRNA potencial codificador de peptídeo por IRES, transcrito esse que se encontra superexpresso em todas as séries e (2) de forma indireta por meio do EIF4G1 e dos processos biológicos enriquecidos, o que sugere um possível efeito do fármaco no controle desses processos nas células.

O papel de IRES é muito conhecido durante infecções virais, mas só recentemente tem sido explorado sua relação com células cancerígenas e outras condições de estresse (SRIRAM; BOHLEN; TELEMAN, 2018) e cerca de 10% de todo o genoma eucarioto, fora de ORFs, possui sequências de IRES (GODET *et al.*, 2019; KWAN; THOMPSON, 2019). O papel de como as células de câncer exploram a tradução por IRES foi explorada em uma revisão de 2018(SRIRAM; BOHLEN; TELEMAN, 2018), mas outras condições de estresse conhecidas ainda não foram associadas ao uso de IRES, como o tratamento por diferentes fármacos. É de se esperar que fármacos muito utilizados, ou sugeridos para o tratamento de câncer, como a cisplatina ou a própria metformina, interfiram no processo de tradução por IRES, mas a revisão de literatura feita para esta dissertação não encontrou nada relacionado. Diversos bancos de dados se propõem a identificar sequências associadas à IRES e ORFs alternativas (ZHAO *et al.*, 2020), no entanto, como no caso do *splicing*, a falta de conhecimento teórico *a priori* da área e a pouca literatura disponível faz com que ainda não exista nenhum método referência para sua identificação e, tampouco, revisões narrativas ou sistemáticas específicas sobre os métodos existentes. O mais próximo encontrado é uma revisão de 2021 (KUTE *et al.*, 2022) que aborda detecção de sORFs, mas não o processo de IRES em específico(KUTE *et al.*, 2022).

Peptídeos codificados a partir de lncRNAs são uma área em expansão, sendo os mais conhecidos também traduzidos por IRES (XING, J. *et al.*, 2021). Além de possível codificador de peptídeos as funções de lncRNAs que ultrapassam sua ação como RNA tem sido muito discutidas recentemente (ALI; GROTE, 2020). Em especial, temos sua ação na regulação da transcrição de regiões codificadoras de outros peptídeos por sua interação com a cromatina, ou como RNAs acentuadores, podendo regular a ligação de fatores de transcrição no promotor desse genes, assim como favorecer a interação do spliceossomo com o pré-mRNA (ARNOLD; WELLS; LI, 2020). Mesmo com tantas funções já descritas, a anotação funcional de lncRNA ainda gera muita discussão na literatura e, como no caso de microRNAs, estudos que não contemplam a deleção ou redução da expressão desses transcritos não são capazes de esclarecer suas possíveis ações e, muitas vezes, se resumem listar o enriquecimento funcional dos

transcritos correlacionados à expressão com o lncRNA alvo, sem uma interpretação fina dos resultados.

O que tentamos adicionar no presente estudo foram informações extras de coordenadas genômicas, resultados de outros tipos de experimento, tais como abertura e contato de cromatina, análise da região da isoforma focada em sua sequência, análise do potencial codificador e uma análise de enriquecimento funcional de vias metabólicas pautada e filtrada nos efeitos já propostos do fármaco, em busca de ter resultados mais confiáveis.

A análise de regiões genômicas ao redor dos lncRNAs é algo que já é feito, mas normalmente só é observado os transcritos a montante e a jusante do gene, e não de forma a se utilizar um tamanho médio de TADs em busca de ações em *cis*. Esse tamanho médio de TAD é extremamente discutível e optamos por utilizar uma distância correspondente aos maiores tamanhos médios preditos em estudos de GAM e HI-C (CHIARIELLO *et al.*, 2022; WINICK-NG *et al.*, 2021). Quando a contagem de isoformas presentes nas regiões genômicas ao redor dos genes das 36 isoformas de lncRNA selecionados foram analisadas e ordenadas por cromossomo **(Figura 11 e Figura 21)** um padrão interessante também emergiu: os dois lncRNAs com região genômica ao redor com mais alvos, NEAT1 e MALAT1, também são dois dos lncRNAs com maior literatura sobre e cuja anotação foi mais antiga. Uma pergunta que fica é se estes lncRNA são tão discutidos exatamente por estarem em uma região genômica muito ativa e, assim, diversos alvos para eles são identificados com maior facilidade, ou se este dado é apenas coincidência. Pela nossa revisão, nenhum estudo chegou a comparar número de artigos publicados com a quantidade de transcritos na região ao redor de certos genes.

O potencial codificador das isoformas foi avaliado pela literatura existente sobre o transcrito, ou de outros transcritos com o qual este compartilhava exons e introns, como o caso do *LINC-PINT*. Softwares de predição de potencial codificador não foram utilizados pelo problema discutido anteriormente de não se existir revisão robusta na área comparando esses softwares, o que leva a diversos falsos positivos. O peptídeo codificado pela circularização do exon 2 de LINC-PINT foi procurado em resultados de análises de cromatografia líquida seguida de espectrometria de massa (LC-MS/MS) de presença diferencial de proteínas em células tratadas com metformina e controles (GAO *et al.*, 2018; LANGE *et al.*, 2021; MORELLI *et al.*, 2021), no entanto, este não foi encontrado. Isso provavelmente se deve à como as análises de proteômica quantitativa são realizadas, onde se utiliza uma anotação *a proiri* do tempo de voo, ou relação massa por carga de peptídeos, como as presentes no *Human*

*Protein Atlas* (UHLÉN *et al.*, 2015) e também de que os espectros de MS de peptídeos são cortados no pré-processamento, a menos que se esteja procurando especificamente por peptídeos pequenos. Duas revisões com repositórios específicos para micropeptídeos codificados por ORFs alternativas foram publicadas recentemente (UHLÉN *et al.*, 2015; YUANYUAN; XINQIANG, 2022), então seria necessária uma reanálise dos dados brutos dos proteomas para uma procura efetiva pela presença deste peptídeo. No entanto, não encontramos nenhuma biblioteca de proteoma quantitativo ou translatoma de células humanas tratadas com metformina também pareada com análise de RNA-Seq.

O enriquecimento funcional é um ponto de discussão recorrente nos trabalhos de transcriptomica (WIJESOORIYA *et al.*, 2022) e diversos softwares se propõem a reduzir seus problemas mais comentados, tais como sobreposição de vias, resultados não significativos, introdução de vieses, dentre outros (HUANG *et al.*, 2013; SUBHASH; KANDURI, 2016; TIPNEY; HUNTER, 2010). No intuito de reduzir estes problemas em nossas análises utilizamos diversas estratégias. Utilizamos um gene set manualmente curado do Msigdb (LIBERZON *et al.*, 2011), após testarmos diversos outros gene sets incluindo GO (ASHBURNER *et al.*, 2000) e REACTOME (GILLESPIE *et al.*, 2022). O gene set do MsigDB se mostrou bem menos redundante e, portanto, apresenta menos interferência manual para retirada de processos biológicos não associados. Ainda não existe nenhuma forma de enriquecimento que englobe isoformas de transcritos. Assim, utilizamos os valores de expressão apenas dos transcritos codificadores de proteínas, aumentando o gene set para aceitar os nomes de transcritos como *input*. A outra estratégia que usamos foi utilizar os valores de correlação entre as isoformas de lncRNA e os mRNAs como valor de ranqueamento no fgsea. Esse valor de ranqueamento normalmente permite que o *FoldChange* da diferença de expressão seja levado em conta na escala do enriquecimento e a direção dessa expressão. Utilizar o valor de correlação como ranqueamento para análises de enriquecimento de expressão diferencial, em nosso conhecimento, é inédito na literatura. As vias selecionadas foram filtradas manualmente de acordo com interpretação biológica das vias conhecidas pela ação do fármaco e por um p ajustado inferior a 0,0001.

A influência genética para o efeito de fármacos hoje é muito focada nas variações de DNA para se individualizar tratamento (MORGANTI *et al.*, 2019; SÁ; SADEE; JOHNSON, 2018; WAKE *et al.*, 2019). Entretanto, o estudo de um fármaco a partir da análise de transcriptoma de células tratadas é bem menos abordado, mesmo com as grandes possibilidades

trazidas para a exploração de efeitos moleculares específicos de cada fármaco, que ainda não são bem conhecidos, nem em um fármaco tão antigo como a metformina.

As bibliotecas utilizadas nesse estudo são de células diferentes tratadas com quantidades diferentes do fármaco e com um tempo de tratamento também distinto **(Tabela 1)**, o que mostra uma altíssima heterogeneidade biológica, mesmo com as mesmas técnicas computacionais de análise sendo utilizadas. Os resultados mostram essa heterogeneidade, tanto com a contagem de lncRNA diferencialmente expressos **(Figura 12)** quanto sua intercessão em várias séries **(Figura 13)**, demonstrando que a maioria dos lncRNA é série-específico e poucos se sobrepõe entre as séries. Pela hipótese do trabalho, já esperávamos toda essa heterogeneidade, já que buscamos os efeitos do fármaco comuns entre os tipos celulares.

É interessante ressaltar que todas as doses farmacológicas utilizadas nos tratamentos estão muito acima das doses farmacológicas prescritas, ou encontradas no sangue dos pacientes, o que já foi discutido extensivamente na literatura, com estudos com doses baixas sendo incapazes de resultar nos efeitos propostos da metformina no complexo I da mitocôndria, e sim observando ações distintas do fármaco (LAMOIA *et al.*, 2023). No entanto, para estudos exploratórios iniciais já foi observado que o modelo *in vitro* é o melhor começo, e é necessário para maximizar a resposta, já que estamos tratando de células em cultura ou primárias e não de pacientes (LIU, Y. *et al.*, 2020; POLLI, 2008). Em nossos resultados é possível verificar (Figura 6), que dosagens muito baixas causam muito pouca perturbação no transcriptoma e que a quantidade de transcritos diferencialmente expressos que passam nos valores de corte, aumenta com a dosagem até certo ponto. Inclusive, análises exploratórias iniciais com outras ferramentas não tão eficientes quanto o swish com todas as bibliotecas encontradas na revisão da literatura **(Figura 31)** mostram isso claramente. É possível notar que os tratamentos com doses menores apresentam nenhum ou quase nenhum lncRNA diferencialmente expresso em comum com as demais. Devido a isso e a outros critérios mencionados na sessão de métodos, para análises posteriores mantivemos as seis séries selecionadas.


***Figura 31:*** *Heatmap de transcritos presentes em quatro ou mais bibliotecas incluindo todas as bibliotecas com número mínimo de replicatas necessário encontradas na análise da literatura. No eixo x as séries analisadas e no eixo y as isoformas de lncRNA que se sobrepõe*

*em mais de 4 séries. Na leganda estão presentes: horas pós-tratamento, concentração de metformina, tipo celular, tamanho da read do sequenciamento e tipo de biblioteca utilizado. O log2FoldChange de cada tarsncrito está representado nas cores roxo amarelo e verde, sendo roxo os transcritos subexpressos, amarelo os superexpressos e verde os que não apresentam valor de expressão diferencial significativa.*

Quanto à intercessão de isoformas diferencialmente expressas entre as seis séries selecionadas **(Figura 13)**, uma possibilidade era de que a origem embriológica das células, o tecido embrionário da qual elas derivam, pudesse explicar as sobreposições maiores entre: Células pancreáticas e renais, células pancreáticas e embrionárias 25 mM, e células embrionárias 25 mM e renais entre as isoformas de lncRNA superexpressas e entre hepatócitos primários e células pulmonares, células embrionárias 25 mM e células pulmonares e células pancreáticas e células pulmonares. Os hepatócitos, as células pancreáticas e o epitélio pulmonar, são derivados de endoderma, já renais são de mesoderma intermediário e as células embrionárias são blastocistos pré-diferenciação. Todas as células derivadas de tumor são linhagens manipuladas para evitar senescência. Entretanto, isso não explica os resultados encontrados, que pode ter havido maior sobreposição por motivos aleatórios derivados de algumas bibliotecas apresentarem muito mais isoformas de lncRNAs diferencialmente expressos do que outras.

Dos quatro artigos que geraram essas seis séries, o primeiro explorava efeitos de metformina em células superexpressoras de Glicerol-3-fosfato desidrogenase (GPD1), células controle e células tratadas somente com metformina. A principal hipótese desse artigo era de que as células superexpressoras de GDP1 respondiam melhor ao efeito antiproliferativo do fármaco do que células controle (XIE *et al.*, 2020). Aqui não analisamos as células superexpressoras e, no artigo, eles não mencionam expressão diferencial de lncRNAs, apenas disponibilizam uma tabela suplementar em que se encontram alguns alvos diferencialmente expressos após o tratamento, inclusive o MALAT1(XIE *et al.*, 2020) .

O segundo artigo, das células pancreáticas, observou os efeitos da metformina sozinhos e em conjunto com efeitos de aspirina no crescimento de tumores de pâncreas. Os autores foram capazes de obter significância estatística de que o tratamento com metformina conjuntamente com aspirina diminuía o crescimento dos tumores *in vitro*. No entanto, não realizaram análises bioinformáticas, ou mencionaram nenhum lncRNA, tanto no artigo quanto nos materiais suplementares (YUE *et al.*, 2015).

O terceiro artigo, das células embrionárias tratadas com diferentes concentrações de metformina, também possuíam outras amostras de células embrionárias com uma mutação na sequência do gene que provoca distrofia miotônica tipo I (DM1), uma doença multi-sistêmica caracterizada por defeitos no *splicing*. No artigo eles chegam à conclusão que o fármaco promove efeitos corretivos no *splicing* e no *splicing* alternativo da célula, entretanto, sua análise foi focada quase que exclusivamente nos genes codificadores das proteínas do spliceossomo e o artigo não cita lncRNAs associados previamente ao *splicing,* ou faz análises bioinformáticas robustas. Esse foi o primeiro artigo a mencionar os efeitos do fármaco como corretivo do *splicing* aberrante (LAUSTRIAT *et al.*, 2015).

O último artigo, cuja célula de estudo são os hepatócitos primários, também possuía dados das mesmas células tratadas com metformina com bloqueadores de AMPK, em busca de explorar os efeitos do fármaco independentes desta via e bibliotecas de ChIP-seq do *Activating Transcription Factor 3* (ATF3), um fator de transcrição central na regulação de inflamação e proliferação. Dos artigos analisados, este era o único que possuía uma análise bioinformática robusta, mesmo assim, seu principal foco foi avaliar os efeitos do fármaco na expressão de fatores de transcrição e, especialmente, em genes superexpressos pela metformina (LUIZON *et al.*, 2016).

Pela nossa revisão, na presente literatura, nenhum estudo se propôs a fazer análise global nem de transcritos diferencialmente expressos pelo fármaco, nem das isoformas de lncRNAs afetadas. Em nossos resultados encontramos tanto isoformas de lncRNAs específicas, que podem ser centrais nos efeitos celulares do fármaco quanto processos biológicos em que estas podem estar agindo. Também sugerimos funções para estas isoformas não antes anotadas e distinguimos possíveis ações moleculares de isoformas de um mesmo lncRNA. Adicionalmente, demonstramos que a heterogeneidade decorrente do uso de diferentes concentrações do fármaco em diferentes tipos celulares pode gerar distintas assinaturas moleculares que corroboram a literatura mais recente.

## 6. Conclusão

- Uma análise robusta de caracterização isoformas diferencialmente expressas de lncRNA em resposta a metformina foi realizada utilizando-se apenas ferramentas *in sílico* englobando região genômica, informações epigenômicas, enriquecimento funcional e revisão da literatura.
- A isoforma de lncRNA AC016831.6-205 está superexpressa nas seis séries em resposta ao tratamento por metformina.
- Duas isoformas de lncRNA, ZBED5-AS1-207 e AC125807.2-201, estão em regiões genômicas que contém transcritos codificadores de proteínas cuja expressão está correlacionada com a expressão destas isoformas. Estes transcritos codificadores de proteína já foram previamente associados à efeitos moleculares propostos da metformina.
- Isoformas presentes em cinco ou mais séries regulam vias metabólicas chave em que o fármaco age.

**Referências**

ADRIAENS, C. *et al.* The long noncoding RNA NEAT1_1 is seemingly dispensable for normal tissue homeostasis and cancer cell growth. 2019. Disponível em: <http://www.rnajournal.org/cgi/doi/10.1261/rna.>.

AFFYMETRIX ENCODE TRANSCRIPTOME PROJECT; COLD SPRING HARBOR LABORATORY ENCODE TRANSCRIPTOME PROJECT. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature*, v. 457, n. 7232, p. 1028–1032, 25 fev. 2009.

AGBANA, Y. L. *et al. LINC00511 as a prognostic biomarker for human cancers: A systematic review and meta-analysis*. *BMC Cancer*. [S.l.]: BioMed Central. , 22 jul. 2020

ALI, T.; GROTE, P. Beyond the RNA-dependent function of LncRNA genes. *eLife*, v. 9, p. 1–14, 1 out. 2020.

AMARAL, P. P. *et al.* Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biology*, v. 19, n. 1, 15 mar. 2018.

ARNOLD, P. R.; WELLS, A. D.; LI, X. C. *Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate*. *Frontiers in Cell and Developmental Biology*. [S.l.]: Frontiers Media S.A. , 14 jan. 2020

ASHBURNER, M. *et al. Gene Ontology: tool for the unification of biology The Gene Ontology Consortium**. . [S.l: s.n.], 2000. Disponível em: <http://www.flybase.bio.indiana.edu>.

BAILEY, C. J. *Metformin: historical overview*. *Diabetologia*. [S.l.]: Springer Verlag. , 1 set. 2017

BEJERANO, G. *et al. Ultraconserved Elements in the Human Genome*. *Adv. Immunol*. [S.l: s.n.], 2004. Disponível em: <www.sciencemag.org/cgi/content/full/304/5675/1318/www.sciencemag.org>.

BOND, C. S.; FOX, A. H. *Paraspeckles: Nuclear bodies built on long noncoding RNA*. *Journal of Cell Biology*. [S.l: s.n.]. , 7 set. 2009

BRUFORD, E. A. *et al. Guidelines for human gene nomenclature*. *Nature Genetics*. [S.l.]: Nature Research. , 1 ago. 2020

BUIST, M.; FUSS, D.; RASTEGAR, M. Transcriptional Regulation of MECP2E1-E2 Isoforms and BDNF by Metformin and Simvastatin through Analyzing Nascent RNA Synthesis in a Human Brain Cell Line. *Biomolecules*, v. 11, n. 8, p. 1253, 22 ago. 2021.

CÁCERES, C. J. *et al.* Non-canonical translation initiation of the spliced mRNA encoding the human T-cell leukemia virus type 1 basic leucine zipper protein. *Nucleic Acids Research*, v. 46, n. 20, p. 11030–11047, 16 nov. 2018.

CAO, M. *et al. Research advances on circulating long noncoding RNAs as biomarkers of cardiovascular diseases*. *International Journal of Cardiology*. [S.l.]: Elsevier Ireland Ltd. , 15 abr. 2022

CHAO, T.-C. *et al.* The Long Noncoding RNA *HEAL* Regulates HIV-1 Replication through Epigenetic Regulation of the HIV-1 Promoter. *mBio*, v. 10, n. 5, 29 out. 2019.

CHEN, X. *et al.* Activation of AMPK inhibits inflammatory response during hypoxia and reoxygenation through modulating JNK-mediated NF-κB pathway. *Metabolism: Clinical and Experimental*, v. 83, p. 256–270, 1 jun. 2018.

CHENG, Y.; WANG, S.; MU, X. Long non-coding RNA LINC00511 promotes proliferation, invasion, and migration of non-small cell lung cancer cells by targeting miR-625-5p/GSPT1. *Translational Cancer Research*, v. 10, n. 12, p. 5159–5173, dez. 2021. Disponível em: <https://tcr.amegroups.com/article/view/58854/html>.

CHIARIELLO, A. M. *et al. Physical mechanisms of chromatin spatial organization*. *FEBS Journal*. [S.l.]: John Wiley and Sons Inc. , 1 mar. 2022

CHU, P.-M. *et al.* Regulation of Oxidative Stress by Long Non-Coding RNAs in Vascular Complications of Diabetes. *Life*, v. 12, n. 2, p. 274, 12 fev. 2022.

CLEMSON, C. M. *et al.* An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Molecular Cell*, v. 33, n. 6, p. 717–726, mar. 2009.

COCCIA, E. M. *et al.* Regulation and Expression of a Growth Arrest-Specific Gene *(gas5)* during Growth, Differentiation, and Development. *Molecular and Cellular Biology*, v. 12, n. 8, p. 3514–3521, 1 ago. 1992.

COOK, M. N. *et al.* Initial monotherapy with either metformin or sulphonylureas often fails to achieve or maintain current glycaemic goals in patients with Type 2 diabetes in UK primary care. *Diabetic Medicine*, v. 24, n. 4, p. 350–358, abr. 2007.

CROUSE, G. F. *et al.* Analysis of the Mouse *dhfr* Promoter Region: Existence of a Divergently Transcribed Gene. *Molecular and Cellular Biology*, v. 5, n. 8, p. 1847–1858, 1 ago. 1985.

CUI, X. Y.; ZHAN, J. K.; LIU, Y. S. *Roles and functions of antisense lncRNA in vascular aging*. *Ageing Research Reviews*. [S.l.]: Elsevier Ireland Ltd. , 1 dez. 2021

CUYÀS, E. *et al. Metformin targets histone acetylation in cancer-prone epithelial cells*. *Cell Cycle*. [S.l.]: Taylor and Francis Inc. , 16 dez. 2016

DING, J. *et al. The role of long intergenic noncoding RNA 00511 in malignant tumors: a meta-analysis, database validation and review*. *Bioengineered*. [S.l.]: Taylor and Francis Inc. , 1 jan. 2020

DUMBOVIĆ, G. *et al.* Nuclear compartmentalization of TERT mRNA and TUG1 lncRNA is driven by intron retention. *Nature Communications*, v. 12, n. 1, 1 dez. 2021.

ENGREITZ, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, v. 539, n. 7629, p. 452–455, 2016.

FRANKISH, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, v. 47, n. D1, p. D766–D773, 8 jan. 2019.

GAO, Y. *et al.* Identification and characterization of metformin on peptidomic profiling in human visceral adipocytes. *Journal of Cellular Biochemistry*, v. 119, n. 2, p. 1866–1878, 7 fev. 2018.

GARY-GOUY, H. *et al.* Natural Phosphorylation of CD5 in Chronic Lymphocytic Leukemia B Cells and Analysis of CD5-Regulated Genes in a B Cell Line Suggest a Role for CD5 in Malignant Phenotype. *The Journal of Immunology*, v. 179, n. 7, p. 4335–4344, 1 out. 2007.

GEBAUER, F.; HENTZE, M. W. *IRES unplugged*. *Science*. [S.l.]: American Association for the Advancement of Science. , 15 jan. 2016

GHAFOURI-FARD, S. *et al.* LncRNAs: Novel Biomarkers for Pancreatic Cancer. *Biomolecules*, v. 11, n. 11, p. 1665, 10 nov. 2021.

GILLESPIE, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, v. 50, n. D1, p. D687–D692, 7 jan. 2022.

GODET, A. C. *et al. IRES trans-acting factors, key actors of the stress response*. *International Journal of Molecular Sciences*. [S.l.]: MDPI AG. , 2 fev. 2019

GOUSTIN, A. *et al.* The Growth-Arrest-Specific (GAS)-5 Long Non-Coding RNA: A Fascinating lncRNA Widely Expressed in Cancers. *Non-Coding RNA*, v. 5, n. 3, p. 46, 17 set. 2019.

GU, X. *et al.* Activated AMPK by metformin protects against fibroblast proliferation during pulmonary fibrosis by suppressing FOXM1. *Pharmacological Research*, v. 173, 1 nov. 2021.

GUO, W. *et al.* 3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *RNA Biology*, v. 18, n. 11, p. 1574–1587, 2 nov. 2021. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/15476286.2020.1858253>.

HAN, Y. *et al.* Post-translational regulation of lipogenesis via AMPK-dependent phosphorylation of insulin-induced gene. *Nature Communications*, v. 10, n. 1, 1 dez. 2019.

HARROW, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, v. 22, n. 9, p. 1760–1774, 5 set. 2012. Disponível em: <http://genome.cshlp.org/lookup/doi/10.1101/gr.135350.111>.

HAYWARD, A. *et al.* ZBED Evolution: Repeated Utilization of DNA Transposons as Regulators of Diverse Host Functions. *PLoS ONE*, v. 8, n. 3, 22 mar. 2013.

HECKMAN-STODDARD, B. M. *et al. Repurposing metformin for the prevention of cancer and cancer recurrence*. *Diabetologia*. [S.l.]: Springer Verlag. , 1 set. 2017

HOWE, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research*, v. 49, n. D1, p. D884–D891, 8 jan. 2021.

HUANG, Q. *et al.* GOMA: Functional enrichment analysis tool based on GO modules. *Chinese Journal of Cancer*, v. 32, n. 4, p. 195–204, 2013.

IP, J. Y.; NAKAGAWA, S. *Long non-coding RNAs in nuclear bodies*. *Development Growth and Differentiation*. [S.l: s.n.]. , jan. 2012

ISOBE, M. *et al.* Forced isoform switching of Neat1_1 to Neat1_2 leads to the loss of Neat1_1 and the hyperformation of paraspeckles but does not affect the development and growth of mice. 2020. Disponível em: <http://www.rnajournal.org/cgi/doi/10.1261/rna.>.

JARROUX, J.; MORILLON, A.; PINSKAYA, M. History, Discovery, and Classification of lncRNAs. [S.l: s.n.], 2017. p. 1–46.

KASTNER, B. *et al.* Structural Insights into Nuclear pre-mRNA Splicing in Higher Eukaryotes. *Cold Spring Harbor Perspectives in Biology*, v. 11, n. 11, p. a032417, nov. 2019. Disponível em: <http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a032417>.

KNUTSEN, E. *et al.* The expression of the long NEAT1_2 isoform is associated with human epidermal growth factor receptor 2-positive breast cancers. *Scientific Reports*, v. 10, n. 1, 1 dez. 2020.

KNUTSEN, E.; HARRIS, A. L.; PERANDER, M. *Expression and functions of long non-coding RNA NEAT1 and isoforms in breast cancer*. *British Journal of Cancer*. [S.l.]: Springer Nature. , 9 mar. 2022

KOPANJA, D. *et al.* Transcriptional Repression by FoxM1 Suppresses Tumor Differentiation and Promotes Metastasis of Breast Cancer. *Cancer Research*, v. 82, n. 13, p. 2458–2471, 1 jul. 2022.

KRAPPINGER, J. C. *et al.* Non-coding Natural Antisense Transcripts: Analysis and Application. *Journal of Biotechnology*, v. 340, p. 75–101, 10 nov. 2021.

KUTE, P. M. *et al. Small Open Reading Frames, How to Find Them and Determine Their Function*. *Frontiers in Genetics*. [S.l.]: Frontiers Media S.A. , 28 jan. 2022

KWAN, T.; THOMPSON, S. R. Noncanonical translation initiation in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, v. 11, n. 4, 2019.

LA, H. M. *et al.* Distinctive molecular features of regenerative stem cells in the damaged male germline. *Nature Communications*, v. 13, n. 1, 1 dez. 2022.

LAMOIA, T. E. *et al.* Metformin, phenformin, and galegine inhibit complex IV activity and reduce glycerol-derived gluconeogenesis. 2023.

LANGE, C. *et al.* Changes in protein expression due to metformin treatment and hyperinsulinemia in a human endometrial cancer cell line. *PLOS ONE*, v. 16, n. 3, p. e0248103, 9 mar. 2021.

LAUSTRIAT, D. *et al.* In vitro and in vivo modulation of alternative splicing by the biguanide metformin. *Molecular Therapy - Nucleic Acids*, v. 4, n. 11, p. e262, 1 nov. 2015.

LEE, A. S. Y. *et al.* EIF3d is an mRNA cap-binding protein that is required for specialized translation initiation. *Nature*, v. 536, n. 7614, p. 96–99, 27 jul. 2016.

LEE, S. H.; MCCORMICK, F. p97/DAP5 is a ribosome-associated factor that facilitates protein synthesis and cell proliferation by modulating the synthesis of cell cycle proteins. *EMBO Journal*, v. 25, n. 17, p. 4008–4019, 6 set. 2006.

LEWIS, S. M. *et al.* The eIF4G homolog DAP5/p97 supports the translation of select mRNAs during endoplasmic reticulum stress. *Nucleic Acids Research*, v. 36, n. 1, p. 168–178, jan. 2008.

LI, S. *et al.* MiR-144-3p-mediated dysregulation of EIF4G2 contributes to the development of hepatocellular carcinoma through the ERK pathway. *Journal of Experimental and Clinical Cancer Research*, v. 40, n. 1, 1 dez. 2021.

LI, Z.-X. *et al.* MALAT1: a potential biomarker in cancer. *Cancer Management and Research*, v. Volume 10, p. 6757–6768, dez. 2018. Disponível em: <https://www.dovepress.com/malat1-a-potential-biomarker-in-cancer-peer-reviewed-article-CMAR>.

LIBERZON, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, v. 27, n. 12, p. 1739–1740, jun. 2011.

LIU, F. *et al.* NEAT1/miR-193a-3p/SOX5 axis regulates cartilage matrix degradation in human osteoarthritis. *Cell Biology International*, v. 44, n. 4, p. 947–957, 1 abr. 2020.

LIU, Y. *et al.* Narrowing the Gap Between In Vitro and In Vivo Genetic Profiles by Deconvoluting Toxicogenomic Data In Silico. *Frontiers in Pharmacology*, v. 10, 8 jan. 2020.

LOGSDON, G. A.; VOLLGER, M. R.; EICHLER, E. E. *Long-read human genome sequencing and its applications*. *Nature Reviews Genetics*. [S.l.]: Nature Research. , 1 out. 2020

LOVE, M. I. *et al.* Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLoS Computational Biology*, v. 16, n. 2, 2020.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, v. 15, n. 12, 5 dez. 2014.

LU, G. *et al.* Long noncoding RNA LINC00511 contributes to breast cancer tumourigenesis and stemness by inducing the miR-185-3p/E2F1/Nanog axis. *Journal of Experimental & Clinical Cancer Research*, v. 37, n. 1, p. 289, 27 dez. 2018.

LUIZON, M. R. *et al.* Genomic Characterization of Metformin Hepatic Response. *PLoS Genetics*, v. 12, n. 11, 1 nov. 2016.

LUO, T. *et al.* AMPK Activation by Metformin Suppresses Abnormal Extracellular Matrix Remodeling in Adipose Tissue and Ameliorates Insulin Resistance in Obesity. *Diabetes*, v. 65, n. 8, p. 2295–2310, 1 ago. 2016.

LÜSCHER-DIAS, T. *et al.* Long non-coding RNAs associated with infection and vaccine-induced immunity. *Essays in Biochemistry*, v. 65, n. 4, p. 657–669, 27 out. 2021. Disponível em: <https://portlandpress.com/essaysbiochem/article/65/4/657/229791/Long-non-coding-RNAs-associated-with-infection-and>.

MA, R. *et al. Metformin and cancer immunity*. *Acta Pharmacologica Sinica*. [S.l.]: Springer Nature. , 1 nov. 2020

MADIRAJU, A. K. *et al.* Metformin inhibits gluconeogenesis via a redox-dependent mechanism in vivo. *Nature Medicine*, v. 24, n. 9, p. 1384–1394, 1 set. 2018.

MANTERE, T.; KERSTEN, S.; HOISCHEN, A. *Long-read sequencing emerging in medical genetics*. *Frontiers in Genetics*. [S.l.]: Frontiers Media S.A. , 2019

MARTINEZ-SALAS, E. *et al. Insights into structural and mechanistic features of viral IRES elements*. *Frontiers in Microbiology*. [S.l.]: Frontiers Media S.A. , 4 jan. 2018

MATTICK, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*, 3 jan. 2023.

MENENDEZ, J. A. Metformin: Sentinel of the Epigenetic Landscapes That Underlie Cell Fate and Identity. *Biomolecules*, v. 10, n. 5, p. 780, 18 maio 2020.

MIRAS, M. *et al.* Non-canonical Translation in Plant RNA Viruses. *Frontiers in Plant Science*, v. 8, 6 abr. 2017.

MORELLI, A. *et al.* Metformin impairs cisplatin resistance effects in A549 lung cancer cells through mTOR signaling and other metabolic pathways. *International Journal of Oncology*, v. 58, n. 6, p. 28, 8 abr. 2021.

MORGANTI, S. *et al.* Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer. [S.l: s.n.], 2019. p. 9–30.

PAI, A. A. *et al.* Numerous recursive sites contribute to accuracy of splicing in long introns in flies. *PLoS Genetics*, v. 14, n. 8, 1 ago. 2018.

PAI, A. A. *et al.* The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. 2017. Disponível em: <https://doi.org/10.7554/eLife.32537.001>.

PATRO, R. *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, v. 14, n. 4, p. 417–419, 2017.

PENG, W.-X.; KOIRALA, P.; MO, Y.-Y. LncRNA-mediated regulation of cell signaling in cancer. *Oncogene*, v. 36, n. 41, p. 5661–5667, 12 out. 2017.

PENZIAS, A. *et al.* Role of metformin for ovulation induction in infertile patients with polycystic ovary syndrome (PCOS): a guideline. *Fertility and Sterility*, v. 108, n. 3, p. 426–441, 1 set. 2017.

PERNICOVA, I.; KORBONITS, M. *Metformin-Mode of action and clinical implications for diabetes and cancer*. *Nature Reviews Endocrinology*. [S.l.]: Nature Publishing Group. , 2014

PLASCHKA, C.; NEWMAN, A. J.; NAGAI, K. Structural Basis of Nuclear pre-mRNA Splicing: Lessons from Yeast. *Cold Spring Harbor Perspectives in Biology*, v. 11, n. 5, p. a032391, maio 2019.

POLLI, J. E. In vitro studies are sometimes better than conventional human pharmacokinetic in vivo studies in assessing bioequivalence of immediate-release solid oral dosage forms. *AAPS Journal*, v. 10, n. 2, p. 289–299, jun. 2008.

QIN, W. *et al.* Combination of dendrobium mixture and metformin curbs the development and progression of diabetic cardiomyopathy by targeting the lncrna neat1. *Clinics*, v. 76, 2021.

RAHMANTO, A. S.; DAVIES, M. J. *Selenium-containing amino acids as direct and indirect antioxidants*. *IUBMB Life*. [S.l: s.n.]. , nov. 2012

RAJAGOPAL, T. *et al. HOTAIR LncRNA: A novel oncogenic propellant in human cancer*. *Clinica Chimica Acta*. [S.l.]: Elsevier B.V. , 1 abr. 2020

RAN, R. *et al. Mechanisms and functions of long noncoding RNAs in intervertebral disc degeneration*. *Pathology Research and Practice*. [S.l.]: Elsevier GmbH. , 1 jul. 2022

RENA, G.; HARDIE, D. G.; PEARSON, E. R. *The mechanisms of action of metformin*. *Diabetologia*. [S.l.]: Springer Verlag. , 1 set. 2017

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, v. 26, n. 1, p. 139–140, 11 nov. 2009.

ROMERO, R. *et al.* Metformin, the aspirin of the 21st century: its role in gestational diabetes mellitus, prevention of preeclampsia and cancer, and the promotion of longevity. *American Journal of Obstetrics and Gynecology*, v. 217, n. 3, p. 282–302, set. 2017.

SÁ, A. C. C.; SADEE, W.; JOHNSON, J. A. Whole Transcriptome Profiling: An RNA-Seq Primer and Implications for Pharmacogenomics Research. *Clinical and Translational Science*, v. 11, n. 2, p. 153–161, 1 mar. 2018.

SCHULTEN, H.-J.; BAKHASHAB, S. Meta-Analysis of Microarray Expression Studies on Metformin in Cancer Cell Lines. *International Journal of Molecular Sciences*, v. 20, n. 13, p. 3173, 28 jun. 2019.

SHARMA, G.; TRIPATHI, S. K.; DAS, S. lncRNA HULC facilitates efficient loading of HCV-core protein onto lipid droplets and subsequent virus-particle release. *Cellular Microbiology*, v. 21, n. 10, 1 out. 2019.

SHEN, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, v. 111, n. 51, p. E5593–E5601, 23 dez. 2014.

SIEBER, P.; PLATZER, M.; SCHUSTER, S. The Definition of Open Reading Frame Revisited. *Trends in Genetics*, v. 34, n. 3, p. 167–170, mar. 2018.

SIVALINGAM, V. N. *et al.* Hypoxia and hyperglycaemia determine why some endometrial tumours fail to respond to metformin. *British Journal of Cancer*, v. 122, n. 1, p. 62–71, 7 jan. 2020.

SMIRNOVA, V. V. *et al.* Ribosomal leaky scanning through a translated uORF requires eIF4G2. *Nucleic Acids Research*, v. 50, n. 2, p. 1111–1127, 25 jan. 2022.

SOROKIN, I. I. *et al. Non-Canonical Translation Initiation Mechanisms Employed by Eukaryotic Viral mRNAs*. *Biochemistry (Moscow)*. [S.l.]: Pleiades journals. , 1 set. 2021

SRIRAM, A.; BOHLEN, J.; TELEMAN, A. A. Translation acrobatics: how cancer cells exploit alternate modes of translational initiation. *EMBO reports*, v. 19, n. 10, out. 2018.

STANĚK, D. Long non-coding RNAs and splicing. *Essays in Biochemistry*, v. 65, n. 4, p. 723–729, 27 out. 2021.

STARK, R.; GRZELAK, M.; HADFIELD, J. *RNA sequencing: the teenage years*. *Nature Reviews Genetics*. [S.l.]: Nature Publishing Group. , 1 nov. 2019

SUBHASH, S.; KANDURI, C. GeneSCF: A real-time based functional enrichment tool with support for multiple organisms. *BMC Bioinformatics*, v. 17, n. 1, 13 set. 2016.

SUGIYAMA, H. *et al.* Nat1 promotes translation of specific proteins that induce differentiation of mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, v. 114, n. 2, p. 340–345, 10 jan. 2017.

SUN, W. *et al.* The NEAT1_2/miR-491 Axis Modulates Papillary Thyroid Cancer Invasion and Metastasis Through TGM2/NFκb/FN1 Signaling. *Frontiers in Oncology*, v. 11, 2 mar. 2021.

TAN, Y. T. *et al. LncRNA-mediated posttranslational modifications and reprogramming of energy metabolism in cancer*. *Cancer Communications*. [S.l.]: John Wiley and Sons Inc. , 1 fev. 2021

THE ENCODE PROJECT CONSORTIUM. An integrated encyclopedia of DNA elements in the human genome. *Nature*, v. 489, n. 7414, p. 57–74, 5 set. 2012.

TIPNEY, H.; HUNTER, L. An introduction to effective use of enrichment analysis software. *Human Genomics*, v. 4, n. 3, p. 202, 2010. Disponível em: <http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-4-3-202>.

TODD, J. N.; FLOREZ, J. C. An update on the pharmacogenomics of metformin: Progress, problems and potential. *Pharmacogenomics*, v. 15, n. 4, p. 529–539, 2014.

TODOROVSKI, V.; FOX, A. H.; CHOI, Y. S. Matrix stiffness-sensitive long noncoding RNA NEAT1 seeded paraspeckles in cancer cells. *Molecular Biology of the Cell*, v. 31, n. 16, p. 1654–1662, 21 jul. 2020.

TRIGGLE, C. R. *et al. Metformin: Is it a drug for all reasons and diseases? Metabolism: Clinical and Experimental*. [S.l.]: W.B. Saunders. , 1 ago. 2022

UHLÉN, M. *et al.* Tissue-based map of the human proteome. *Science*, v. 347, n. 6220, 23 jan. 2015.

ULE, J.; BLENCOWE, B. J. *Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution*. *Molecular Cell*. [S.l.]: Cell Press. , 17 out. 2019

VITTING-SEERUP, K.; SANDELIN, A.; BERGER, B. IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, v. 35, n. 21, p. 4469–4471, 1 nov. 2019.

WAKE, D. T. *et al. Pharmacogenomics: Prescribing Precisely*. *Medical Clinics of North America*. [S.l.]: W.B. Saunders. , 1 nov. 2019

WALTERS, B.; THOMPSON, S. R. *Cap-independent translational control of carcinogenesis*. *Frontiers in Oncology*. [S.l.]: Frontiers Media S.A. , 2016

WANG, E.; AIFANTIS, I. *RNA Splicing and Cancer*. *Trends in Cancer*. [S.l.]: Cell Press. , 1 ago. 2020

WANG, J. *et al.* Host Long Noncoding RNA lncRNA-PAAN Regulates the Replication of Influenza A Virus. *Viruses*, v. 10, n. 6, p. 330, 16 jun. 2018.

WANG, X. *et al.* LncRNA NEAT1 promotes extracellular matrix accumulation and epithelial-to-mesenchymal transition by targeting miR-27b-3p and ZEB1 in diabetic nephropathy. *Journal of Cellular Physiology*, v. 234, n. 8, p. 12926–12933, 1 ago. 2019.

WANG, Z. *et al.* Bone Mesenchymal Stem Cells Promote Extracellular Matrix Remodeling of Degenerated Nucleus Pulposus Cells via the miR-101-3p/EIF4G2 Axis. *Frontiers in Bioengineering and Biotechnology*, v. 9, 27 ago. 2021.

WEEKLEY, C. M. *et al.* Uptake, distribution, and speciation of selenoamino acids by human cancer cells: X-ray absorption and fluorescence methods. *Biochemistry*, v. 50, n. 10, p. 1641–1650, 15 mar. 2011.

WEI, L. *et al. Noncoding RNAs in gastric cancer: Implications for drug resistance*. *Molecular Cancer*. [S.l.]: BioMed Central Ltd. , 19 mar. 2020

WIJESOORIYA, K. *et al.* Urgent need for consistent standards in functional enrichment analysis. *PLoS Computational Biology*, v. 18, n. 3, 1 mar. 2022.

WINICK-NG, W. *et al.* Cell-type specialization is encoded by specific chromatin topologies. *Nature*, v. 599, n. 7886, p. 684–691, 25 nov. 2021.

WU, M. *et al. Metformin and Fibrosis: A Review of Existing Evidence and Mechanisms*. *Journal of Diabetes Research*. [S.l.]: Hindawi Limited. , 2021

XIE, J. *et al.* GPD1 enhances the anticancer effects of metformin by synergistically increasing total cellular glycerol-3-phosphate. *Cancer Research*, v. 80, n. 11, p. 2150–2162, 1 jun. 2020.

XING, C. *et al. Role of lncRNA LUCAT1 in cancer*. *Biomedicine and Pharmacotherapy*. [S.l.]: Elsevier Masson s.r.l. , 1 fev. 2021

XING, J. *et al. LncRNA-Encoded Peptide: Functions and Predicting Methods*. *Frontiers in Oncology*. [S.l.]: Frontiers Media S.A. , 14 jan. 2021

XU, Q. *et al.* Systematic comparison of lncRNAs with protein coding mRNAs in population expression and their response to environmental change. *BMC Plant Biology*, v. 17, n. 1, 13 fev. 2017.

YAMANAKA, S. Essential role of NAT1/p97/DAP5 in embryonic differentiation and the retinoic acid pathway. *The EMBO Journal*, v. 19, n. 20, p. 5533–5541, 16 out. 2000. Disponível em: <http://emboj.embopress.org/cgi/doi/10.1093/emboj/19.20.5533>.

YANG, F. *et al.* Expression profile analysis of long noncoding RNA in HER-2-enriched subtype breast cancer by next-generation sequencing and bioinformatics. *OncoTargets and Therapy*, v. 9, p. 761–772, 12 fev. 2016.

YANG, G. *et al.* Identification of Critical Genes and lncRNAs in Osteolysis after Total Hip Arthroplasty and Osteoarthritis by RNA Sequencing. *BioMed Research International*, v. 2021, p. 1–13, 13 mar. 2021. Disponível em: <https://www.hindawi.com/journals/bmri/2021/6681925/>.

YANG, J. *et al.* NEAT1 Knockdown Inhibits Keloid Fibroblast Progression by miR-196b-5p/FGF2 Axis. *Journal of Surgical Research*, v. 259, p. 261–270, 1 mar. 2021.

YI, Q.-Y. *et al. Metformin inhibits development of diabetic retinopathy through inducing alternative splicing of VEGF-A*. *Am J Transl Res*. [S.l: s.n.], 2016. Disponível em: <www.ajtr.org/ISSN:1943-8141/AJTR0033278>.

YIP, C. W. *et al. Functional annotation of lncRNA in high-throughput screening*. *Essays in Biochemistry*. [S.l.]: Portland Press Ltd. , 1 out. 2021

YUANYUAN, J.; XINQIANG, Y. *Micropeptides Identified from Human Genomes*. *Journal of Proteome Research*. [S.l.]: American Chemical Society. , 1 abr. 2022

YUE, W. *et al.* Metformin combined with aspirin significantly inhibit pancreatic cancer cell growth *in vitro* and *in vivo* by suppressing anti-apoptotic proteins Mcl-1 and Bcl-2. *Oncotarget*, v. 6, n. 25, p. 21208–21224, 28 ago. 2015. Disponível em: <https://www.oncotarget.com/lookup/doi/10.18632/oncotarget.4126>.

ZEA, D. J. *et al.* Assessing conservation of alternative splicing with evolutionary splicing graphs. *Genome Research*, v. 31, n. 8, p. 1462–1473, ago. 2021.

ZHANG, MAOLEI *et al.* A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nature Communications*, v. 9, n. 1, 1 dez. 2018.

ZHANG, MIAO *et al.* The Role of Long Non-coding RNA, Nuclear Enriched Abundant Transcript 1 (NEAT1) in Cancer and Other Pathologies. *Biochemical Genetics*, v. 60, n. 3, p. 843–867, 24 jun. 2022.

ZHANG, P. *et al.* The lncRNA Neat1 promotes activation of inflammasomes in macrophages. *Nature Communications*, v. 10, n. 1, 1 dez. 2019.

ZHANG, Q. *et al.* Hypoxia-Induced lncRNA-NEAT1 Sustains the Growth of Hepatocellular Carcinoma via Regulation of miR-199a-3p/UCK2. *Frontiers in Oncology*, v. 10, 24 jun. 2020.

ZHANG, X. *et al.* FOXM1-mediated activation of phospholipase D1 promotes lipid droplet accumulation and reduces ROS to support paclitaxel resistance in metastatic cancer cells. *Free Radical Biology and Medicine*, v. 179, p. 213–228, 1 fev. 2022.

ZHAO, J. *et al.* IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. *Genomics, Proteomics & Bioinformatics*, v. 18, n. 2, p. 129–139, abr. 2020. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1672022920300577>. Acesso em: 19 abr. 2023.

ZHOU, K. *et al.* Heritability of variation in glycaemic response to metformin: a genome-wide complex trait analysis. *The Lancet Diabetes & Endocrinology*, v. 2, n. 6, p. 481–487, jun. 2014.

ZHU, A. *et al.* Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Research*, v. 47, n. 18, p. E105, 10 out. 2019.

ZIEGLER, C.; KRETZ, M. *The more the Merrier-Complexity in long non-coding RNA loci*. *Frontiers in Endocrinology*. [S.l.]: Frontiers Research Foundation. , 25 abr. 2017

Gmail Izabela Mamede <iza.mamede@gmail.com>

# Help with the annotation of transcript RP1136B6.2

2 messages

**Izabela Mamede** <iza.mamede@gmail.com> Fri, Nov 19, 2021 at 3:17 PM
To: gencode-help@ebi.ac.uk

Dear GENCODE team,

My name is Izabela Mamede Conceição, I am a master's student from the Federal University of MInas Gerais in Brazil, and the group I work with studies alternative splicing of lncRNA in human diseases. When I analyze human RNAseq datasets I use GENCODE annotation to determine which lncRNA isoforms I am seeing in my results so your work has been extremely helpful to me.

My question is about the annotation of transcript RP1136B6.2, more than 70% of its sequence is the same as lncRNA LINC-PINT, but it was annotated as RP1136B6.2 on GENCODE38.

I would like to know which criteria you used to annotate it as its own transcript and not as a LINC-PINT isoform. I work with cancer datasets and much of the biological proposed effects of LINC-PINT reside as protective in cancer. When I look at multiple TCGA datasets of non-invasive cancer I see also RP1136B6.2 upregulated.

I hope this email finds its recipient in good health.

Best regards,

Izabela Mamede C. A. Conceição

Master Student at the Laboratory of Genetics Biochemistry
Federal University of Minas Gerais (UFMG)

Lattes: http://lattes.cnpq.br/7854459951544002
ORCID: https://orcid.org/0000-0002-0707-5588

---

**Toby Hunt via RT** <gencode-help@ebi.ac.uk> Wed, Nov 24, 2021 at 10:38 AM
Reply-To: gencode-help@ebi.ac.uk
To: iza.mamede@gmail.com

<URL: http://helpdesk.ebi.ac.uk/Ticket/Display.html?id=547626 >

Hi Izabela,

many thanks for your email. It would seem that the answer to your question lies in the history of our annotation in this region. Our original annotation (based on just EST and cDNA evidence) consisted of the LINC-PINT (ENSG00000231721) gene and a further lincRNA just downstream (ENSG00000226380), which did not overlap each other.

Later on, when we came to look at this region again, we had some additional long read PacBio evidence available to us that overlapped both these genes and this is what was used to support the annotation of RP1136B6.2

(ENSG00000285106). Because RP1136B6.2 overlapped two pre-existing genes we made it as what we call a "readthrough locus", as its transcripts readthrough between two known genes.

However, upon examining this region yet again (due to your query), and with our latest annotation guidelines and yet more long read sequences now available, it would seem more appropriate to merge all three of these loci together as part of a single gene (LINC-PINT - ENSG00000231721). So moving forward, these are all now going to be part of one locus - although I should point out that, due to the length of time our release cycle takes these changes probably won't appear in Ensembl/GENCODE for around six months or so.

For more details about our readthrough loci annotation please see this blogpost:

https://www.ensembl.info/2019/02/11/
annotating-readthrough-transcription-in-ensembl/

there's a section specifically about lncRNA loci towards the bottom.

I hope this answers your question.

Cheers,

Toby.
[Quoted text hidden]