DANIELA KAJIHARA

Caracterização e investigação funcional de *splicings* alternativos do gene *P4HB* da dissulfeto isomerase proteica (PDI)

Tese apresentada à Faculdade de Medicina da
Universidade de São Paulo para a obtenção do
título de Doutor em Ciências

Programa de Ciências Médicas

Área de concentração: Processos Inflamatórios
e Alérgicos

Orientador: Prof. Dr. Francisco Rafael Martins
Laurindo

(Versão corrigida. Resolução CoPGr 6018/11, de 13 de outubro de 2011. A versão original está disponível na Biblioteca da FMUSP)

São Paulo

2019

DANIELA KAJIHARA

Caracterização e investigação funcional de *splicings* alternativos do gene *P4HB* da dissulfeto isomerase proteica (PDI)

Tese apresentada à Faculdade de Medicina da
Universidade de São Paulo para a obtenção do
título de Doutor em Ciências

Programa de Ciências Médicas

Área de concentração: Processos Inflamatórios
e Alérgicos

Orientador: Prof. Dr. Francisco Rafael Martins
Laurindo

São Paulo

2019

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

©reprodução autorizada pelo autor

Kajihara, Daniela

Caracterização e investigação funcional de
splittings alternativos do gene P4HB da dissulfeto
isomerase proteica (PDI) / Daniela Kajihara. -- São
Paulo, 2019.

Tese(doutorado)--Faculdade de Medicina da
Universidade de São Paulo.

Programa de Ciências Médicas.Área de
Concentração: Processos Inflamatórios e Alérgicos.
Orientador: Francisco Rafael Martins Laurindo.

Descritores: 1.Isomerasas de dissulfetos de
proteínas 2.Células de músculo liso 3.Processamento
alternativo 4.Isoformas de RNA 5.Processamento de
RNA 6.Músculo liso vascular 7.Retículo
endoplasmático

USP/FM/DBD-219/19

Responsável: Erinalva da Conceição Batista, CRB-8 6755

Á memória de meu pai, Luiz.

Por acreditar nas minhas ideias e sempre me apoiar

Agradecimentos

Ao Professor Francisco por orientar esse trabalho e ser um exemplo de pesquisador e professor. Agradeço pela confiança, paciência e tranquilidade em orientar esse trabalho. Pela oportunidade de desbravar fronteiras e tentar olhar para uma área diferente que o lab normalmente segue.

To Professor Piero Carninci for opening his lab, allowing me to grow as a PhD student and encouraging my research.

To Professor Kosuke Hashimoto thank you for mentoring me and for all the scientific discussions about this work.

To Chung Chau Hon, thank you for your support and great ideas to this work. Showing me that write scripts is not that hard.

Ao Professor Diego Bonatto, Joice Poloni e Kendi do Laboratório de Biologia Computacional da UFRGS, muito obrigada por abrirem mais caminhos e perspectivas no mundo da bioinformática.

A Ana Moretti, João Wosniak e Victor pelo suporte nos experimentos e discussões científicas.

A Prof. Ayumi Miyakawa e João Carlos do Laboratório de Genética e Cardiologia- InCor.

A Leonora, José Arruda e Maria Bertolini, pelo apoio e as histórias animadas do dia a dia.

A todos do Laboratorio de Biologia Vascular e labs ao redor: Leonardo, Laura, Angela, Percília, Patrícia, Vitor, Luciana, Denise, Tiffany, Carolina, Jessyca, Thaís, Fernando, Rafael, Ana Cecília, Elaine, Telma, Andreia, Marici e Olga.

For all members of Laboratory for Transcriptome Technology-RIKEN to extend the horizons in research. Aimi, Ferdous, Michael, Hazuki, Harshita, Kaori, Oguchi san, Nitta san, Shruti, Zarma, Erma, Prashanti, Fernando, Hin, Kurosawa san, Nana san, Furukawa san and Nagai san.

A Secretária da Pós-Graduação de Ciências Médicas FM-USP, principalmente a Angélica pela atenção e apoio nas atividades da pós-graduação.

A todos os meus amigos e amigas de longa data, por todo apoio, risadas e conversas.

A minha mãe, por todo amor, carinho e dedicação. Ao meu irmão e minha cunhada pela amizade e incentivo. Aos meus familiares por todo o suporte e carinho.

Este trabalho foi desenvolvido pelo apoio da Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

This work was also funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) to the RIKEN Center for Life Science Technologies.

ÍNDICE DE FIGURAS

Figura 1. Distribuição dos domínios da *P4HB* humana

Figura 2. Representação esquemática da composição de 21 proteínas na família gênica da PDI.

Figura 3. Estrutura gênica do gene *P4HB* humano e suas variantes de *splicing* predita

Figura 4. Organização dos domínios da proteína PDI e as isoformas de proteínas preditas

Figura 5. Alinhamento das sequências de aminoácidos de PDI e suas isoformas

Figure 6. Amostras representativas de um subgrupo do (A) FANTOM5 (FANTOM CAGE) e amostras do (B) ENCODE CAGE

Figura 7. Distribuição da expressão das variantes de *splicing* do gene *P4HB* nas amostras de RNA-seq do FANTOM5.

Figura 8. Perfil de expressão das variantes de *splicing* *P4HB*-02 e *P4HB*-021

Figura 9. Gráficos representativos quatro amostras de RNA-seq de FANTOM5

Figura 10. Distribuição da expressão das variantes de *splicing* do gene *P4HB* em amostras de RNA-seq do ENCODE

Figura 11. Perfil de expressão das variantes de *splicing* *P4HB* em 27 amostras do banco de dados do ENCODE

Figura 12. (A) SJ-TPM das variantes de *splicing* do gene *P4HB* em dois tipos celulares: células endoteliais (HUVEC e HAoEC) e fibroblastos adventícios da aorta (HaoAF)

Figura 13. Quantificação das variantes de *splicing* *P4HB*-02, *P4HB*-021 e *P4HB*-027 no GTEx.

Figura 14. Quantificação de variantes de *splicing* do gene *P4HB* em amostras de vasos sanguíneos e coração.

Figura 15. Quantificação de variantes de *splicing* do gene *P4HB* em amostras de sangue total.

Figura 16. Análise de dados de RNA-seq de um estudo (Yu *et al.*, 2018) usando células VSMC (células de músculo liso vasculares) mimetizando as condições patológicas (stiff) e fisiológicas (soft).

Figura 17. Amplificação por PCR da junção específica das variantes de *splicing*: *P4HB*-02, *P4HB*-021 e *P4HB*-027.

Figura 18. qPCR de células de músculo liso de artéria mamária para o gene *P4HB*, três variantes de *splicing* do gene *P4HB* e 4 genes relacionados ao estresse do retículo endoplasmático submetidas a carenciamento de soro por 16h e 24h.

Figura 19. qPCR de células de músculo liso de artéria mamária tratado com CoCl_2 durante 24h.

Figura 20. qPCR de células HEK-293 para o gene *P4HB* e as três variantes de *splicing* escolhidas tratadas com tunicamicina por 10h e tratadas com tunicamicina 40h

Figura 21. qPCR de células HEK-293 tratados com tunicamicina por 16 e 24h para a análise de expressão de sete genes relacionados ao estresse do retículo endoplasmático.

Figura 22. qPCR de células HEK-293 para a análise de expressão do gene *P4HB* e três variantes de *splicing*, *P4HB-02*, *P4HB-021* e *P4HB-027*. HEK-293 tratada com CoCl_2 durante 10h e 24h

Figura 23. Modelagem proteica predita da isoforma *P4HB-021*. (A) Organização do domínio proteico do PDIA1 com a indicação de quatro domínios: *a*, *b*, *b'*, *a'*.

Figura 24. (A) Estrutura da tiorredoxina (12 kDa) e (g) enovelamento da tiorredoxina

ÍNDICE DE TABELAS

Tabela 1. Lista de oligonucleotídeos usados nesse estudo.

Tabela 2. Variantes de *splicing* do gene *P4HB* em Humanos

LISTA DE ABREVIATURAS

Akt ½: Inibidor de proteína quinase B

BAM: binary alignment map;

BRCA: Breast câncer 1 gene

CAGE: Cap analysis gene expression;

CHOP: C/EBP homologous protein

CoCl₂: Cloreto de cobalto

ER: retículo endoplasmático

EDTA: Ácido etilenodiamino tetra-acético

FANTOM5: Functional Annotation of the Mouse/Mammalian Genome

GAPDH: Gliceraldeído-3-fosfato desidrogenase

Gm12878: Human lymphoblastoid cell line

GTEX: Genotype Tissue Expression

H₂O₂: Peróxido de hidrogênio

HaoEC: Human aortic endothelial cells

HaoAF: Human aortic adventitial fibroblasts

HCT-116: Human colon cancer cells

Hmsc: Human mesenchymal stem cell line

HPA: Human Protein Atlas

HPRT1: Hypoxanthine phosphoribosyltransferase 1

Hsmm: Human skeletal muscle myoblast cell line

HUVEC: Human Umbilical Vein Endothelial Cells

KDEL: sequência-sinal de aminoácidos

KRAS: Kirsten rat sarcoma viral oncogene homolog

MDM2: Proto-oncogene MDM2

NADPH: Fosfato de dinucleotídeo de adenina e nicotinamida

NCBI: National Center for Biotechnology Information

PBS: Tampão fosfato salino

PDI: Proteína disulfeto isomerase

PCR: Reação em cadeia da polimerase

PDGF: Platelet-derived growth factor

RNA-seq: RNA sequencing

RE: Retículo endoplasmático

SAM: Sequence alignment map

SK-N-SH: Neuroblastoma cell line

SETD2: SET Domain Containing 2 gene

TBE: Tampão Tris/Borato/EDTA

TGF- β : Fator de transformação do crescimento beta

TRX: Tiorredoxina

TPM: Tags per million

VSMC: Células musculares lisas vasculares

UPR: Resposta a proteínas mal enoveladas

UTR: Região não traduzida

Resumo

Kajihara D. *Caracterização e investigação funcional de splicings alternativos do gene P4HB da dissulfeto isomerase proteica (PDI) [tese]* São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2019.

O *splicing* alternativo (AS) é um processo que produz diferentes mRNAs de um transcrito primário, resultando na produção de isoformas de proteínas que podem exibir propriedades funcionais distintas e localização subcelular. Mais de 90% dos genes humanos sofrem AS. O gene *P4HB* codifica a proteína PDIA1 (proteína dissulfeto isomerase-A1), uma chaperona redox composta basicamente por quatro domínios tiorredoxina. A PDIA1 exerce função essencial na proteostase mediante funções como chaperona e isomerase ou oxidoreductase de grupos tiol em cisteínas, catalisando a inserção adequada de pontes dissulfeto em proteínas nascentes no retículo endoplasmático. A atividade de isomerase de PDI requer a presença de todos os seus domínios redox catalíticos (*a* e *a'*) e os domínios não redox (*b* e *b'*), enquanto as isoformas poderiam possuir estruturas distintas e atuam como oxidoreductases, ou apenas chaperonas. Ainda, a PDIA1, amplamente expressa, pode ter efeitos adicionais na homeostase celular. Nosso grupo tem caracterizado uma série de funções da PDIA1 na sinalização redox, homeostase e respostas biomecânicas, em alguns casos envolvendo locais de proteína subcelular adicionais, como na superfície celular ou no meio extracelular. No entanto, há muito pouca informação sobre a regulação, expressão e localização de transcritos derivados de processamento alternativo da PDI e a localização de diferentes produtos proteicos. Avanços recentes nos dados de NGS (Next Generation Sequencing) e proteômicos forneceram informações mais acuradas de variantes alternativamente processadas. A disponibilidade de dados de RNA-seq de diferentes fontes tornou possível a investigação de variantes de *splicing* usando uma quantidade razoável de dados, contemplando diferentes tipos e tecidos celulares. Usando três diferentes bases de dados RNA-seq, investigamos o perfil de isoformas alternativamente processadas da PDIA1. A análise das bases FANTOM5 (Anotação Funcional de Genoma de Mamíferos, que reuniu informações biológicas de 70 amostras de RNA-seq), Projeto ENCODE/Caltech e CSHL RNA-seq e Projeto GTEx (Genotype-Tissue Expression, um enorme banco de dados com mais de 10.000 amostras), identificamos a expressão das variantes de *splicing* do gene *P4HB* em diferentes tipos de células e linhagens celulares cancerígenas. A expressão das variantes de *splicing* do gene *P4HB* compreende 10 variantes que possuem a sequência de codificação de proteína. Desenvolvemos abordagens específicas para quantificar as variantes desse gene. Para validar a junção específica para cada variante de *splicing*, nós desenhamos oligonucleotídeos para amplificar a junção, clonamos e sequenciamos a região específica de cada uma das três variantes. Os dados indicam que a expressão das isoformas do gene *P4HB* é altamente variável entre os diferentes tecidos e células e mesmo entre diferentes tipos da mesma célula, por exemplo, células musculares lisas de aorta vs. coronária. A análise da sequência das isoformas mostra que apenas uma, a *P4HB*-021, tem os quatro domínios tiorredoxina, apenas duas conservam a sequência de retenção no retículo endoplasmático e várias tem truncamentos dos distintos domínios, sugerindo que os produtos proteicos derivados não devam apresentar atividade isomerase canônica. Analisamos a expressão gênica de três variantes de *splicing* em células HEK293 e células musculares lisas primárias submetidas à carencia de soro, tratamento com tunicamicina e cloreto de cobalto para mimetizar a hipóxia. O gene da *P4HB* e da isoforma *P4HB*-021 foram responsivos ao carencia de soro em células musculares lisas humanas. A importância da isoforma *P4HB*-021 neste tipo

celular foi também confirmada por análise de outros bancos de dados e experimentos da literatura. Interessante, a estrutura prevista para *P4HB-021* indica uma proteína com maior mobilidade que a proteína canônica. Estes dados indicam uma base para o entendimento da regulação gênica do gene da PDIA1 e podem contribuir para o entendimento das diversas funções previamente descritas por nosso grupo para a PDIA1 neste tipo celular.

Descritores: 1. Isomerasas de dissulfetos de proteínas 2. Células de músculo liso 3. Processamento alternativo 4. Isoformas de RNA 5 Processamento de RNA 6. Músculo liso vascular 7. Retículo endoplasmático

Summary

Kajihara D. *Characterization and functional investigation of P4HB splicing variants [thesis]*. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2019.

More than 90% of human genes undergo AS. The *P4HB* gene encodes the PDIA1 protein (disulfide protein isomerase-A1), a redox chaperone composed primarily of four domains of thioredoxin. PDIA1 exerts essential function in proteostasis through functions such as chaperone and isomerase or oxido-reductase of thiol groups in cysteines, catalyzing the proper insertion of disulfide bridges into nascent proteins in the endoplasmic reticulum. The isomerase activity of PDI requires the presence of all its catalytic redox domains (*a* and *a'*) and the non-redox domains (*b* and *b'*), while distinctly structured isoforms could act as oxidoreductases or just as chaperones. Furthermore, PDIA1, widely expressed, may have additional effects on cellular homeostasis. Our group has characterized a number of PDIA1 functions in redox signaling, homeostasis and biomechanical responses, in some cases involving additional subcellular localization, such as on the cell surface or in the extracellular environment. However, there is very little information on the regulation, expression and localization of transcripts derived from alternative PDI processing and the location of different protein products. Recent advances in NGS (Next Generation Sequencing) and proteomics data provided more accurately information from alternative *splicing* variants. The availability of RNA-seq data from different sources empowers the investigation of splice variants using a reasonable amount of data from cell types and tissues. Using three different RNA-seq databases, we investigated the alternatively processed isoform profile of PDIA1. The analysis of the FANTOM5 bases (Functional Annotation of Mammalian Genome, which collected biological information from 70 RNA-seq samples), ENCODE/Caltech Project and CSHL RNA-seq and GTEx Project (Genotype-Tissue Expression, a huge database with more than 10,000 samples), we identified the expression of *P4HB* *splicing* variants in different cell types and cancer cell lines. The expression of *P4HB* *splicing* variants comprises 10 variants with protein coding sequence. We developed specific approaches to quantify variants of this gene. To validate the specific junction for each *splicing* variant, we designed primers to amplify the splice junction, cloned and sequenced the specific region of each of the three variants. The data indicate that the expression of the isoforms of *P4HB* gene is highly variable between different tissues and cells and even between different types of the same cells, e.g. smooth muscle cells from coronary artery vs aorta. The analysis of isoforms sequence shows that only one, *P4HB*-021, has the 4 domains of thioredoxin, only two retain the retention sequence to endoplasmic reticulum and several truncations of the different domains, suggesting that the derived protein products should not present isomerase activity. We analyzed the gene expression of three *splicing* variants in HEK293 cells and primary smooth muscle cells submitted to serum starvation, treatment with tunicamycin and cobalt chloride to mimic hypoxia. The *P4HB* gene and the *P4HB*-021 isoform were responsive to serum starvation in human smooth muscle cells. The importance of the *P4HB*-021 isoform in this cell type was also confirmed by analysis of other databases and literature experiments. Interestingly, the predicted structure for *P4HB*-021 indicates a protein with greater mobility than the canonical protein. These data indicate a basis for understanding the gene regulation of the PDIA1 gene and may contribute to the understanding of the various functions previously described by our group for PDIA1 in this cell type.

Descriptors: 1. Protein disulfide-isomerases 2. Myocytes, smooth muscle 3. Alternative *splicing* 4. RNA isoforms 5. RNA *splicing* 6. Vascular smooth muscle 7. Endoplasmic Reticulum

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS

Resumo

Summary

1.Introdução	2
1.1 PDI (dissulfeto isomerase proteica): funções canônicas e envolvimento em sinalização redox	2
1.2 <i>Splicing</i> alternativo	6
1.3 Mecanismo de <i>splicing</i> alternativo	7
1.4 <i>Splicing</i> alternativo e doenças	8
2. Hipótese e justificativa	9
3. Objetivos.....	10
3.1 Objetivo geral	10
3.2 Objetivos específicos	10
4. Materiais e Métodos.....	13
4.1 Estratégias para identificação das isoformas derivadas de <i>splicing</i>	13
4.1.1 Identificação de isoformas: bancos de dados de proteínas e mRNAs	13
4.1.2 Bancos de dados de RNA-seq	13
4.1.3 Processamento de dados de RNA-seq.....	14
4.1.4 Bibliotecas de RNA-seq de amostras humanas	15
4.1.5 Análise do sítio de início de transcrição gênica usando <i>CAGE tags</i>	16
4.1.6 Análise de bioinformática	17
4.1.7 Visualização de <i>splicing</i> alternativo em variantes de <i>splicing</i> do gene <i>P4HB</i> ...	17
4.2 Estratégias para identificação e quantificação da expressão gênica das isoformas	18
4.2.1 Cultivo de linhagens celulares e células primárias	18
4.2.2 Síntese de oligonucleotídeos e PCR	19
4.2.3 Extração de RNA.	20
4.2.4 Síntese de cDNA e qPCR	20
4.2.5 Ensaio de careciamento de soro, tratamento com tunicamicina e exposição a	

CoCl ₂	21
4.2.6 Validação experimental das junções de splicing alternativo	21
4.2.7 Clonagem de novas variantes de splicing <i>P4HB</i> em humanos	22
4.2.8 Análise estatística	22
4.3 Modelagem estrutural das isoformas	23
5. Resultados.....	25
5.1 <i>Splicing</i> alternativo do gene <i>P4HB</i>	25
5.2 Alinhamento das sequências de aminoácidos	29
5.3 Análise de CAGE <i>tags</i>	31
5.4 Perfil de expressão das variantes do gene <i>P4HB</i> em amostras de RNA-seq do banco de dados FANTOM5.....	32
5.5 Perfil de expressão das variantes do gene <i>P4HB</i> em amostras de RNA-seq do ENCODE	36
5.6 Perfil de expressão de três variantes de <i>splicing</i> do gene <i>P4HB</i> em amostras de RNA-seq do GTEx.....	38
5.7 Perfil de expressão das variantes do gene <i>P4HB</i> em amostras de RNA-seq de células de músculo liso vascular	42
5.8 Identificação de isoformas do gene <i>P4HB</i> em camundongo e rato	43
5.9 Análise de validação das variantes de <i>splicing</i> de <i>P4HB</i>	44
5.10 Expressão gênica das isoformas do gene da <i>P4HB</i> em linhagens celulares	45
5.11 Modelagem virtual da estrutura protéica de <i>P4HB</i> -021 comparada à <i>P4HB</i>	49
Discussão	52
Conclusão	53
Perspectivas.....	55
ANEXOS	60
ANEXO A – Tabela 1 – Identificação dos arquivos FANTOM5	60
ANEXO B –Tabela 2 – Números de acesso de arquivos ENCODE	62
ANEXO C – Tabela 3 – Amostras GTEx	63
Referências	64

Introdução

1.Introdução

1.1 PDI (dissulfeto isomerase proteica): funções canônicas e envolvimento em sinalização redox

A dissulfeto isomerase proteica (PDIA1 ou PDI) é uma proteína da superfamília da tiorredoxina, com localização primária no lúmen do RE. É responsável por catalisar a introdução e isomerização de pontes dissulfeto em proteínas nascentes ou desenoveladas por meio de sua atividade oxidoreductase de ditiol-dissulfetos. Para realizar suas funções, a PDI tem dois pares de ditióis que podem reduzir, oxidar e isomerizar pontes dissulfeto (Pan *et al.*, 2014). PDI apresenta também atividade chaperona, que é sinérgica, porém independente da atividade oxidoreductase. Essa atividade chaperona pode contribuir ao enovelamento de proteínas que não contêm pontes dissulfeto. A flavoproteína Ero1 (*Endoplasmic reticulum oxidoreductin 1*) é a principal oxidase do RE e responsável por reoxidar os dissulfetos da PDI, para serem então introduzidas em proteínas-alvo. Os sítios ativos das cisteínas da Ero1 são re-oxidadas pela transferência de seus elétrons para o cofator FAD (dinucleotídeo flavina-adenina) e para o oxigênio molecular, produzindo peróxido de hidrogênio (Sevier & Kaiser, 2008; Gross *et al.*, 2006). Dessa forma, PDI participa de uma via do RE que leva à geração de oxidantes como subprodutos, sendo estes parcialmente consumidos por reações catalisadas por enzimas como peroxiredoxina 4 e glutathione peroxidases, capazes ao mesmo tempo de sustentar o enovelamento de proteínas e possivelmente de manter o balanço redox do RE. Além da localização no lúmen do RE, PDI tem sido repetidamente identificada na superfície celular, em associação com o controle tiol-redox de várias funções: internalização de vírus, trombose e hemostasia, ativação plaquetária, regulação de metaloproteinases e adesão celular, entre outras (Wu *et al.*, 2012; Laurindo *et al.*, 2012; Camargo *et al.*, 2013). As vias pelas quais a PDI alcança a superfície celular permanecem obscuras e em células endoteliais envolvem tanto rotas convencionais como não-convencionais (Araújo *et al.*, 2017).

Devido ao seu papel vital no enovelamento protéico, a perda da atividade de PDI tem grande relevância potencial em processos patogênicos, sendo comumente relacionada com a resposta a proteínas mal enoveladas (UPR, unfolded protein response)

que recentemente tem sido um alvo terapêutico em várias doenças como Alzheimer, doença de Parkinson, hepatopatias e diabetes tipo 2 (Galligan & Petersen, 2012). Além disso, a PDI (no caso, o pool específico extracelular) tem importante papel trombogênico e é um dos principais alvos farmacológicos de medicações anti-trombóticas (Jasuja *et al.*, 2012)

A superfamília tiorredoxina (TRX) da qual a PDI faz parte abrange também as isoformas de tiorredoxinas, glutaredoxina e peroxiredoxinas (Jacquot *et al.*, 2002). Uma característica comum entre os membros da família PDI é a presença do domínio TRX, composto por ditióis do formato -CXXC- (onde C é uma cisteína e X outro aminoácido). Além disso, as PDIs compartilham também um característico enovelamento TRX em sua estrutura (que será discutido adiante). A PDI é composta de seis domínios (*a*, *b*, *b'*, *x*, *a'* e *c*) e tem um peso molecular em torno de 57 kDa (Freedman *et al.*, 1994). A Figura 1 mostra a representação esquemática dos domínios *a* e *a'*, que contém os domínios TRX ativos com os resíduos de cisteína conservados e o motivo CGHC (Cis-Gli-His-Cis). Os outros domínios, *b* e *b'*, têm o enovelamento TRX mas não são catalíticos, isto é, não contém os sítos redox-ativos. Por sua vez, o domínio *b'* fornece o principal sítio de ligação para substratos (Klappa *et al.*, 1998) e o domínio *b* assegura o alinhamento de sítos funcionais e deve conferir estabilidade estrutural para a proteína. Há ainda uma região de 19 aminoácidos, o x-linker, localizado entre o domínio *b'* e *a'*, que permite uma flexibilidade inter-domínio (Richards *et al.*, 2016). Um domínio c carboxi-terminal possui a sequência KDEL (lis-asp-glu-leu), que é o sinal de retenção no retículo endoplasmático (Darby *et al.*, 1999).

Figura 1. **Distribuição dos domínios da P4HB humana.** Os números evidenciam a posição dos aminoácidos (Hatahet & Ruddock, 2009)

As proteínas da subfamília da PDI são amplamente expressas no RE, e poucos membros foram encontrados em outros compartimentos subcelulares (Turano *et al.*, 2002). De fato, a PDI propriamente dita é de fato a isoforma PDIA1, que constitui o membro fundador dessa família, que abrange 21 genes conhecidos, que variam em tamanho, expressão, localização e funções enzimáticas. Com a crescente disponibilidade

de bancos de dados de RNA-seq incluindo diferentes tipos celulares e tecidos depositados em banco de dados NCBI, uma constante revisão desses bancos podem mudar o atual número de 21 membros (Galligan & Petersen, 2012).

Figura 2. **Representação esquemática da composição de 21 proteínas na família gênica da PDI.** Os domínios ativos *TRX-like*, em vermelho e os domínios inativo *TRX-like* em azul e roxo. Em branco, sinal de retenção no retículo endoplasmático COOH terminal. Em amarelo, domínio transmembrana (Galligan & Petersen, 2012)

A PDI é codificada pelo gene *P4HB*, o primeiro membro da família a ser descrito e originalmente identificado como a subunidade beta da prolil-4-hidroxilase. Está localizado no cromossomo 17q25 e consiste de 11 éxons (Pihlajaniemi *et al.*, 1987; Tasanen *et al.*, 1988). *P4HB* é altamente expresso em diferentes tecidos e muito abundante nas células (Ferrari *et al.*, 1999). Até o momento, nenhum knockout viável de camundongo para o gene *P4HB* foi reportado, consistente com seu relevante papel funcional no enovelamento proteico (Hatahet & Ruddock, 2009).

Nosso grupo tem investigado ativamente várias vias de sinalização e homeostase redox pela PDI, que pode ocorrer em várias localizações subcelulares (lúmen do RE e um pool específico na superfície celular). O controle redox pela PDI pode envolver vias

associadas à resposta a proteínas mal enoveladas e estresse do RE (Laurindo *et al*, 2012). Particularmente, nosso grupo descreveu uma interação física e funcional entre PDIA1 e NADPH oxidases, em especial a isoforma Nox1. PDI associa-se fisicamente às subunidades catalíticas Nox1, Nox2 e Nox4 e a outras subunidades do complexo NADPH oxidase, como p22phox e p47phox. Além disso, em células de músculo liso vascular (VSMCs – *Vascular Smooth Muscle Cells*), a superexpressão de PDI por si induz aumento de expressão e atividade de Nox1 e depleção de PDI inibe significativamente a atividade de Nox1 induzida por angiotensina II e PDGF (*Platelet-Derived Growth Factor*), interferindo, respectivamente, na ativação de Akt e na migração celular (Pescatore *et al.*, 2012). Os mecanismos desses efeitos da PDI estão sendo ativamente investigados. Análises com ferramentas de biologia de sistemas para inferir interatomoas da PDI indicaram substancial convergência com pequenas GTPases da família Rho (*RhoGTPases*) e o regulador RhoGDI. De fato, depleção de PDI reduziu a atividade de Rac1 e RhoA em VSMCs estimuladas com PDGF e promoveu significativa disfunção e desestruturação do citoesqueleto celular. Vários projetos em andamento no laboratório têm dado suporte à importância dessas interações, em distintos níveis de hierarquia regulatória (*in silico*, *in vitro*, células e modelos animais). Em conjunto, esses dados sugerem que a ação reguladora de PDI sobre NADPH oxidases possa ocorrer no contexto de sua interação com RhoGTPases e, particularmente, de seu tráfego para regiões celulares próximos à membrana plasmática e/ou superfície celular. É plausível que PDI atue no tráfego e montagem de proteínas compondo micro domínios de sinalização de NADPH oxidases, agindo como um adaptador entre sinalização redox e processos celulares dependentes do citoesqueleto (Pescatore *et al*, 2012; Soares Moretti & Martins Laurindo, 2017).

O entendimento dos mecanismos pelos quais a PDI regula a sinalização e homeostase redox, bem como outros de seus múltiplos efeitos, permanece ainda pouco esclarecido. A grande maioria dos estudos abordando a regulação da PDI foca na modulação de sua abundante expressão proteica. Quase nada é conhecido sobre a regulação gênica do gene *P4HB*, apesar de sua grande importância funcional. Há indícios de que a regulação gênica da PDI seja relevante. Por exemplo, a expressão forçada do gene tem consequências funcionais relevantes (Fernandes *et al*, 2009) na ativação da Nox1 em células musculares lisas e gera fenótipos relacionados ao sistema cardiovascular. Por exemplo, camundongos com superexpressão geneticamente

induzida da PDI têm remodelamento vascular exacerbado quando submetidos a ligadura de artéria carótida. Além disso, as células musculares lisas (VSMC) das artérias destes camundongos têm um fenótipo hiper-diferenciado (Fernandes *et al*, dados não publicados). De fato, expressão forçada da PDI em VSMC gera aumento de migração celular, porém a expressão sustentada induz espontaneamente um fenótipo de diferenciação (Fernandes *et al*, dados não publicados). Em paralelo, vários dados de nosso grupo indicam que PDI tem efeitos na regulação do citoesqueleto de VSMC (Pescatore *et al*, 2012). Recentemente, mostramos que o pool extracelular de PDI tem efeito na organização da distribuição polarizada de forças em VSMC (Tanaka *et al*, 2019).

Deste modo, apesar dos altos níveis de expressão basal da PDI, a regulação de sua expressão tem um claro potencial para efeitos funcionais, especificamente na VSMC em condições relevantes a doença cardiovascular. Um aspecto específico e essencial da regulação de expressão gênica é relacionado a *splicings* alternativos, que permanecem essencialmente desconhecidos no caso da PDIA1.

1.2 *Splicing* alternativo

O *splicing* alternativo (SA) pode originar transcritos que podem adquirir novas funções, alterando ou não o quadro de leitura, ou até mesmo causar um atraso no quadro de leitura gerando um códon de terminação prematura que leva à degradação do mRNA (NMD – Nonsense mediated decay), ou uma proteína com um enovelamento diferente, ou ainda a produção de uma proteína truncada. (Poloni & Bonatto, 2018). O processo de *splicing* de RNA origina diferentes transcritos variantes de um mesmo gene, conhecidos como variantes de *splicing*. A sua tradução resulta em diferentes produtos protéicos (Robert & Watson, 2016).

O SA permite que diversas isoformas sejam transcritas a partir de um único gene, resultando em um mecanismo de geração de diversidade proteica em eucariotos a partir de cerca de 25 mil genes humanos (Venter *et al.*, 2001; Ochoa-Leyva *et al.*, 2013). Esse mecanismo é considerado um fator importante para o aumento da complexidade celular e funcional em eucariotos (Matlin *et al.*, 2005; Blencowe *et al.*, 2006). Além de colaborar com a complexidade proteômica, SA contribui diretamente na geração de novas isoformas e indiretamente na modulação de domínios que são modificados pós-

traducionalmente (Wang, 2004). Por exemplo, a selenoproteína tiorredoxina redutase 1 humana (TrxR1), codificada pelo gene TXNRD1, é um componente chave na regulação redox. A TrxR1 produz diversos transcritos alternativos. Um deles, denominado v3, carrega um domínio glutaredoxina N-terminal atípico e, quando superexpresso, associa-se com a membrana e aciona o filopódio (Dandimopoulou *et al.*, 2009). Outro exemplo é o caso da chaperona GRP78, do retículo endoplasmático, que pode apresentar a variante GRP78va. Em condições de estresse do retículo endoplasmático ocorre o aumento da produção dessa variante, que tem em sua sequência a retenção do primeiro íntron e o sinal interno de tradução. Com a ausência do sinal de retenção, GRP78va localiza-se no citosol e está notavelmente elevada em linhagem celular de leucemia humana submetida ao estresse do RE (Ni *et al.*, 2009). Além disso, por meio de ensaio de ganho (superexpressão) e perda de função (knockdown com siRNA) da GRP78va, foi demonstrada a sua habilidade de regular a sinalização de PERK (uma quinase envolvida no estresse do RE), interagindo e atuando como um antagonista da P58^{IPK}, um inibidor da PERK (Ni *et al.*, 2009).

A detecção de SA permanece desafiadora. Entretanto, com o incremento de novas tecnologias há, atualmente, maior acurácia na detecção das isoformas gênicas como, por exemplo, o uso de tecnologias de sequenciamento de alto rendimento. O RNA-seq é uma tecnologia de segunda geração de sequenciamento que possibilita uma ampla cobertura das sequências de mRNA e a distinção das isoformas. Os resultados obtidos por esta técnica têm mostrado que o SA é bem mais importante do que previamente antecipado. É estimado que 74 a 94% dos genes humanos sofram SA (Wang *et al.*, 2008; Coulombe-Huntington *et al.*, 2009). O RNA-seq proporciona uma medição de níveis transcricionais muito mais precisa, assim como a identificação de suas isoformas (Wang *et al.*, 2009). Atualmente, a disponibilidade de bancos de dados de sequências de cDNA tem aumentado em qualidade e abrangência tecido-específica. De fato, além do controle geral, SA fornece um controle adicional de expressão tecido-específica (Grosso *et al.*, 2008).

1.3 Mecanismo de *splicing* alternativo

Um gene é geralmente caracterizado por éxons intercalados por longos íntrons. A partir dos genes são primeiramente sintetizados os transcritos primários que logo

depois são modificados, com a adição na extremidade 5' do *cap* e na 3' as sequências de poliadenilação. Para a formação do mRNA maduro os éxons são identificados e unidos, enquanto os íntrons são removidos, e depois transportados para fora do núcleo, onde são traduzidos em proteínas (Wessagowit *et al.*, 2005).

O processamento de RNA é realizado pelo *spliceossomo*, um complexo de ribonucleoproteínas responsável pela junção de dois éxons, seguida pela remoção das sequências intrônicas (Wahl *et al.*, 2009). Todo esse processo de junção de éxons é direcionado por sequências presentes nas junções éxon/íntron chamadas sítios de *splicing*. A maquinaria de *splicing* em humanos é complexa, envolvendo cinco pequenas ribonucleoproteínas (U1, U2, U4, U5 e U6 snRNAs), sendo que o spliceossomo humano contém mais de 300 proteínas diferentes (Jurica & Moore, 2003). A informação necessária para decidir quais éxons serão incluídos dependerá do local e da presença das sequências de pré-RNA (Barash *et al.*, 2010), como também do grau de ligação das ribonucleoproteínas com essas sequências (Nilsen & Graveley, 2010, David & Manley, 2010). No caso do SA, diferenças na maquinaria do *splicing*, no reconhecimento dos sítios splice, presença de elementos regulatório em cis e/ou trans, metilação de DNA e outros fatores podem influenciar o evento (Chen & Manley, 2009; Maor *et al.*, 2015). O SA é altamente controlado pela célula com o objetivo de manter a homeostase. A proteína quinase Akt possui três isoformas e tem importante papel na regulação da proliferação, sobrevivência e sinalização celular (Chen *et al.*, 2001). Uma análise de fosfoproteômica identificou que as isoformas Akt possuem substratos únicos e outros que se sobrepõem, porém mostrou-se uma tendência de que vias de sinalização de insulina e IGF sejam reguladas pela Akt2 (Akt isoforma 2) e proteínas envolvidas na modificação pós-transcricional do mRNA sejam reguladas por Akt1 e Akt3. Adicionalmente, a fosforilação de IWS1 por Akt1/Akt3 é requerida para a ligação com SETD2 (histona metil transferase) e a formação do complexo RNA polimerase II seguida pelo recrutamento da proteína MRG15. Este por sua vez irá recrutar PTB (proteína de ligação de polipirimidina) que irá regular a inclusão de éxons (Lee e Blenis, 2014; Sanidas *et al.*, 2014).

1.4 *Splicing* alternativo e doenças

Uma das doenças em que eventualmente é resultado de alterações no mecanismo

de *splicing* é o câncer. O processo de *splicing* é comumente afetado e sua desregulação resulta em produtos finais funcionais e não funcionais (Sveen *et al.*, 2016). Mudanças na concentração, localização e atividade de muitas proteínas que regulam o processo de *splicing* (*splicing factors*) também podem induzir a produção de certas isoformas que podem acelerar a progressão da doença (David & Maley., 2010). Por exemplo, numerosos casos clínicos demonstraram que o perfil de expressão do gene p53 pode estar ligado à progressão do tumor, resposta clínica e prognóstico. Um exemplo marcante é a isoforma p53 β que promove senescência replicativa, enquanto a isoforma Δ 133p53 α promove a proliferação e o escape da senescência (Surget *et al.*, 2013). Muitas desordens Mendelianas, como a fibrose cística, têm sido explicadas por erros no processo de *splicing* que resultam em mutações genéticas (Kralovicova & Vorechovsky, 2007). Outra doença genética é a β -talassemia, que afeta a síntese de hemoglobina subunidade beta (β -globina), cuja redução ou ausência causa eritropoiese ineficaz resultando em anemia. As mutações de *splicing* estão entre as mais comuns entre 200 mutações diferentes que causam β -talassemia, muitas delas envolvendo mutações na junção éxon-intron afetando a produção dos níveis de transcritos canônicos (Vadolas *et al.*, 2006).

A proteína do gene da *Titina* (TTN) é a terceira mais abundante proteína muscular (Bang *et al.*, 2001). É um importante gene causador de doenças do músculo cardíaco, localizado no cromossomo 2 e constituído por 363 éxons. Esses éxons codificam para várias proteínas TTN, devido ao mecanismo de *splicing* regulados por motivos de ligação ao RNA, originando proteínas de tamanhos diferentes. *Splicing* alternativos dessas variantes possuem expressão tecido-específica, e em humanos detectou-se a expressão dessas isoformas em corações com cardiomiopatias (Chen *et al.*, 2018). No músculo cardíaco três isoformas estão presentes: a isoforma N2BA adulto, N2B adulto e a isoforma titina cardíaca fetal (Bang *et al.*, 2001; Lahmers *et al.*, 2004; Greaser *et al.*, 2005)

Um estudo recentemente publicado mostrou que o estresse mecânico em células endoteliais e a presença de polimorfismos de nucleotídeo único (do inglês, single nucleotide polymorphism - SNP) no gene MYCK (miosina quinase de cadeia leve) aumentou a expressão da isoforma chamada nmMLCK2 em células endoteliais vasculares, reduzindo a integridade da barreira de células endoteliais. Essa isoforma não possui o exón 11, que codifica uma região da proteína que influencia na barreira

vascular durante a inflamação (Mascarenhas *et al.*, 2017). Estes exemplos mostram o envolvimento do mecanismo de *splicing* e suas implicações nos diferentes tipos celulares e progressão das doenças.

2. Hipótese e justificativa

Nossa hipótese é que o gene *P4HB* da PDI sofra *splicings* alternativos e que eventuais produtos resultantes deste processo tenham implicações funcionais na regulação das múltiplas funcionalidades dependentes da PDI. Considerando a importância funcional e a essencialidade da PDI, a investigação de uma ou mais isoformas especificamente caracterizadas em uma condição de estresse, tecido-específicas, vinculada a processos fisiopatológicos ou mesmo resultado de uma localização subcelular diferente poderá indicar novos mecanismos de regulação ou outras funções celulares relevantes.

No entanto, não há essencialmente nenhuma informação a respeito de variantes de AS para o gene da *P4HB*. Esta questão é importante não apenas em perspectiva com os múltiplos e relevantes efeitos biológicos da PDI discutidos acima, como também pela peculiar organização estrutura-função da PDI. Tal organização indica que mesmo pequenas alterações na proteína podem produzir mudanças funcionais nas variantes de *splicing*. Esta questão pode ser pertinente a várias doenças, porém considerando os importantes efeitos da PDI descritos em nosso grupo na modulação funcional da célula muscular lisa vascular (VSMC), este é um modelo celular relevante à pergunta do estudo.

3. Objetivos

3.1 Objetivo geral

Investigar a ocorrência e possível implicação funcional de isoformas gênicas derivadas de *splicing* alternativo do gene *P4HB* da dissulfeto isomerase proteica (PDI).

3.2 Objetivos específicos

1. Identificar em bancos de dados de RNA-seq as isoformas do gene *P4HB* derivadas de *splicing*.
2. Construir um perfil de expressão gênica das variantes de *splicing* do gene *P4HB* usando três bancos de dados de RNA-seq (FANTOM5, ENCODE e GTEx).
3. Validar e caracterizar as principais variantes de *splicing* do gene *P4HB*
4. Verificar o perfil de expressão transcricional de três variantes de *splicings* do gene da *P4HB* em tecidos de vasos sanguíneos, em partes do coração e sangue total em dados de RNA-seq usando dados do GTEx.
5. Investigar aspectos funcionais das três variantes de *splicing* do gene *P4HB* usando qPCR em amostras, com foco na célula muscular lisa vascular.

Materias e Métodos

4. Materiais e Métodos

4.1 Estratégias para identificação das isoformas derivadas de *splicing*

4.1.1 Identificação de isoformas: bancos de dados de proteínas e mRNAs

A identificação de isoformas utilizando diferentes bancos de dados com anotações de sequências de aminoácidos e mRNA tem possibilitado investigação e detecção de variantes. O UniProtKB/Swiss-Prot (<https://www.uniprot.org/>) representa um banco curado de sequências proteicas anotadas e com informações sobre as proteínas. As sequências são manualmente anotadas e, por isso, são consideradas sequências proteicas de alta qualidade (Blakeley *et al.*, 2010). O ENSEMBL é um projeto conjunto entre a EMBL-EBI e a *Wellcome Trust Sanger Institute*, que tem como função desenvolver uma plataforma capaz de manter e produzir automaticamente as anotações de genomas eucariotos (<https://www.ensembl.org/index.html>).

4.1.2 Bancos de dados de RNA-seq

Com o uso cada vez mais rotineiro das técnicas de sequenciamento de RNA (RNA-seq) em larga escala, a quantidade de dados publicamente acessíveis tem aumentado significativamente. Com essas técnicas, é possível a obtenção de um perfil transcriptômico de diversos organismos, resultando em dados que podem ser usados para estudos de análise de expressão relativa, *splicing* alternativo, descoberta de novos transcritos e isoformas, edição de RNA, expressão alelo-específica ou ainda a investigação de transcriptomas de organismos não modelo (Anders *et al.*, 2013).

Nesse tipo de análise de RNA-seq são geradas em torno de dez a cem milhões de pequenas sequências chamadas “*reads*”. Tais “*reads*” são organizados e alinhados ao genoma de referência no caso de organismos-modelo, como o genoma de humano e camundongo. É possível contabilizar os “*reads*” para cada gene do genoma e estimar, utilizando ferramentas estatísticas, o grau de expressão de cada gene, como também a detecção de isoformas gênicas.

O projeto Anotação Funcional do Genoma de Mamíferos (FANTOM5) abrange mais de 1000 amostras de humanos e ratos que foram analisados por CAGE, a maioria das amostras derivadas de células primárias (Lizio, *et al.*, 2015). Para esta análise foi utilizado um

subconjunto de setenta amostras humanas do FANTOM5 nas quais foram realizadas RNA-seq (Hon *et al.*, 2017). Mais informações sobre as amostras do estão disponíveis no Anexo A.

As bibliotecas RNA-seq do repositório ENCODE (<https://www.encodeproject.org/>) estão disponíveis publicamente. O Anexo B detalha o número de acesso de cada conjunto de dados. O primeiro grupo de conjunto constitui duas amostras de células primárias de artéria pulmonar de humanos, obtidas de dois doadores do sexo masculino. O segundo grupo é composto por linhagens celulares e linhagem de células cancerígenas representando 13 tipos diferentes de células, contemplando no total 27 amostras. Este conjunto de dados foram submetidos ao repositório ENCODE por laboratórios da *California Institute of Technology* (Caltech) e outro conjunto produzidos por *Cold Spring Harbor Laboratory* (CSHL). Os dados de acesso estão disponíveis no Anexo B. Para o conjunto de RNA-seq publicado por Yu *et al.*, 2018 de células musculares lisas de aorta próxima e células de músculo liso de artéria coronária, o número de acesso do estudo é GSE100081.

O Genotype-Tissue Expression Project (<http://www.gtexportal.org/home/>) é um banco que agrega informações sobre variações genéticas entre indivíduos (humanos) com o objetivo de estudar os efeitos dessa variação na regulação da expressão gênica. Os dados estão disponíveis no link <https://gtexportal.org/home/datasets>. Nesse banco há uma razoável quantidade de dados de RNA-seq compreendendo 30 diferentes tecidos humanos representados por 11690 amostras de RNA-seq, sendo o maior conjunto de dados entre os três bancos. Detalhes sobre as amostras disponíveis no Anexo C.

Os bancos de dados de RNA-seq publicados e disponíveis no site NCBI (<https://www.ncbi.nlm.nih.gov/gds>) foram usados para o estudo. Os dados possuem números de acesso GSE (accession number of series) ou GSM (accession number for samples) disponíveis em Anexos.

4.1.3 Processamento de dados de RNA-seq

O protocolo de análise de dados foi baseado em Trapnell *et al.* (2012), porém algumas modificações foram realizadas para adequar ao objetivo do proposto trabalho. Os dados de sequenciamento estão em formato SRA e precisaram ser convertidos para o formato FASTQ, usando o comando `fastq-dump` da ferramenta SRA Toolkit para a conversão. Para acessar a qualidade das sequências de *reads* foi usado o programa FastQC versão 0.11.2 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Esse programa gera um

relatório para cada biblioteca de RNA-seq sobre o comprimento das sequências e a sua qualidade. Após a verificação da qualidade dos *reads*, essas sequências foram cortadas para a eliminação de problemas, como baixa qualidade de sequências e remoção de adaptadores, utilizando o programa Trimmomatic (Bolger *et al.*, 2014). Após esse programa, os dados foram novamente analisados com o programa FastQC.

Continuando a análise dos dados, usamos o alinhador HISAT, pois esse programa utiliza um sistema altamente eficiente de alinhamento de sequências derivadas de experimentos de RNA-seq (Kim *et al.*, 2015). Utilizamos o programa para indexar o genoma de referência (GRCh37 de *Homo sapiens*) criando índices para o alinhamento das sequências. Após isso, iniciamos o alinhamento das sequências usando o mesmo programa, realizando também a conversão do arquivo fastq para SAM. Os alinhamentos gerados pelo programa HISAT são representados em arquivo formato SAM. Esse precisa ser convertido para o arquivo formato BAM através do programa SAMtools utilizando o comando *view*, seguido pela ordenação, com o comando *sort* do SAMtools. Para a visualização dos arquivos BAM no software Integrative Genome Viewer (IGV) (Robinson *et al.*, 2011), é necessário que o arquivo BAM esteja associado a um arquivo de índice que pode ser feito como comando *index* do SAMtools dentro do ambiente UNIX. Somente os estudos que não disponibilizaram o arquivo BAM seguiram esse protocolo.

O protocolo requisitou um conhecimento básico de bioinformática, como a familiaridade com a interface de linhas de comando dentro do sistema UNIX e a escrita de scripts, instalação de programas e ferramentas dentro desse sistema, e também a edição, criação de arquivos *txt*, conversão de arquivos, entre outros.

4.1.4 Bibliotecas de RNA-seq de amostras humanas

O desenvolvimento de novos métodos para avaliar o perfil transcricional, como por exemplo, o desenvolvimento de plataformas que produzam *reads* mais longos, como também a escolha de dados de RNA-seq com sequenciamento *paired-end* (sequenciamento a partir das duas extremidades do fragmento), resultando em fragmentos de alta qualidade, facilitando a detecção de novos transcritos e isoformas gênicas, tudo isso é necessário para melhorar a qualidade da análise de dados.

Os dados de RNA-seq do FANTOM5 que correspondem a setenta amostras incluem um conjunto diversificado de amostras biológicas humanas. Estas amostras, com exceção de

amostras de sangue total, sofreram depleção ribossômica usando o kit de remoção rRNA Ribo-Zero (Epicer, *Illumina*). As amostras de sangue total, células B (CD19) e células T (CD8) foram selecionadas por cauda poli-A usando Dynabeads Oligo(dT) (Life Technologies). Esta biblioteca é de 100 bp, sequenciada *single-end* na plataforma *Illumina Hi-Seq 2000* com uma profundidade de ~ 200 milhões de *reads* cada (Hon *et al.*, 2017).

Para dados de RNA-seq da plataforma ENCODE (<https://genome.ucsc.edu>), selecionamos 13 linhagens celulares diferentes: CD20, GM12878, Haoaf, Haoec, Hepg2, HeLa, HSMM, HUVEC, HCT-116, Hlesc, Hmec, Nhek e SK -N-SH no total de 27 amostras. As treze amostras foram selecionadas devido à presença de CAGE *tags* na região de sítio de início de transcrição. Estes conjuntos de dados contêm *reads* com cerca de 75 ~ 200 pb e foram sequenciadas no *Illumina Genome Analyzer* ou *Illumina Hi-Seq 2000*. Os dados do trabalho de Yu *et al.*, (2018) foram sequenciados em plataforma *Illumina Hi-Seq 2500*, 100 bp, *paired-end*.

Os dados do GTEx foram obtidos em <https://gtexportal.org/home/datasets>. O RNA-seq foi realizado usando a preparação da construção da biblioteca *Illumina TruSeq* e o sequenciamento produziu 76 bp, *paired-end reads*. Mais informações podem ser obtidas neste link <https://gtexportal.org/home/documentationPage>. A lista das 11690 amostras de RNA-seq é mostrada no Anexo C.

4.1.5 Análise do sítio de início de transcrição gênica usando CAGE *tags*

O CAGE (*Cap Analysis of Gene Expression*) é uma tecnologia para mapear a maioria dos locais de início da transcrição e seus promotores, decifrando assim a expressão dos RNAs produzidos em cada promotor (Takahashi *et al.*, 2012). Nessa metodologia mensuram-se os níveis de expressão dos sítios de início de transcrição por sequenciamento de grandes quantidades de transcritos na sua extremidade 5', denominado CAGE *tags* (Kawaji *et al.*, 2006). Essa tecnologia permite a aquisição do perfil de expressão gênica, a identificação do uso do promotor e o local específico de início da transcrição (TSS), porém somente é usado o RNA maduro. Com a contagem dos números de CAGE *tags* para cada promotor dentro de um gene, é possível determinar o nível de expressão e o uso de diferentes promotores. Esta *tag* possui 27 nucleotídeos, sendo uma referência para facilitar o mapeamento dessas *tags* ou etiquetas no genoma. A contagem de *tags* CAGE pode ser visualizada e os dados adquiridos através da plataforma ZENBU (Severin *et al.*, 2014).

4.1.6 Análise de bioinformática

Para determinar um perfil da expressão das variantes de *splicing* do gene *P4HB*, investigamos dados de RNA-seq disponíveis no FANTOM5 Project, Consórcio ENCODE e GTEx. Para este estudo, desenvolvemos um *pipeline* para contar a junção de *splicing* usando arquivos BAM (Binary Alignment/map format), que é uma representação binária dos arquivos SAM (Sequence Alignment /map format) desenvolvido para o Projeto 1000 Genomes (Li *et al.*, 2009; Rosenbloom *et al.*, 2009). O pipeline para mapeamento da junção de *splicing* foi escrito especificamente para as junções do gene *P4HB*, quantificando a junção específica para cada variante de *P4HB* descrita na Figura 1. A unidade usada para calcular foi chamada de *tag* de junção de *splicing* por milhão (SJ TPM, *splice junction tag per million*), uma unidade representando grau de expressão gênica, que representa a contagem de *tags* que possuam a junção de *splicing* específica para determinada isoforma dividida pelo número total de *reads* para cada conjunto de dados multiplicado por 1.000.000. Os dados do FANTOM5 e ENCODE Consortium foram normalizados para o número total de *reads* como descrito acima. Os dados do GTEx, devido à grande quantidade de amostras, foram normalizados pela contagem de *tags* que possuam uma junção específica para o gene *P4HB*. Desta forma, tais resultados são expressos como uma fração, pois representam a fração das isoformas em relação ao gene *P4HB*. O pipeline computacional que identifica a junção de *splicing* para um gene específico foi construído com filtros rigorosos para maximizar a especificidade da junção de *splicing*. O RStudio foi usado para visualizar a expressão em SJ TPM. Para o cálculo da área do círculo nos gráficos apresentados, usamos a fórmula $r = \sqrt{\text{área do círculo}/\pi}$, onde r = raio, área do círculo = soma de SJ TPM de todas as amostras. Os pacotes gráficos do R utilizados foram *ggplot* e *symbols*.

4.1.7 Visualização de *splicing* alternativo em variantes de *splicing* do gene *P4HB*

Para visualizar os dados de RNA-seq, usamos o navegador Integrative Genomics Viewer (IGV) (Thorvaldsdóttir *et al.*, 2013), usando os arquivos BAM (*Binary alignment map*) e anotação de modelo gênico no formato GFF (Reese *et al.*, 2010). Os arquivos usados para a visualização no IGV são arquivos do formato BAM de dados provenientes do FANTOM5 (Figura 8D e 9). Essa etapa é crucial para uma análise exploratória inicial dos

dados e possível detecção ou não das variantes de *splicing*.

4.2 Estratégias para identificação e quantificação da expressão gênica das isoformas

4.2.1 Cultivo de linhagens celulares e células primárias

Estudos atualizados da Encyclopedia of DNA Elements (ENCODE) indicam que em média, existem sete transcritos variantes codificados para cada gene codificante de uma proteína (Dunhan *et al.*, 2012; Sveen *et al.*, 2016). Um levantamento realizado para a família da PDI em humanos com base na plataforma *The Human Protein Atlas* e ENSEMBL mostrou que, dos 21 membros da família, a PDIA1 ou PDI (o membro fundador da família) possui dez isoformas preditas (discutidas acima) e a PDIA6 seis isoformas. O restante dos membros apresentou em média de um a três isoformas.

Para os ensaios de expressão gênica das isoformas selecionadas, utilizamos as células de linhagem de rim embrionário, HEK-293 (*human embryonic kidney 293*), assim como células musculares lisas primárias de artéria mamária humana, obtidas de um pool de segmentos coletados durante cirurgia de revascularização.

HEK-293 foram cultivadas em meio DMEM (*high glucose*) e as células primárias de músculo liso de artéria mamárias foram cultivadas em meio DMEM (*low glucose*) com 10% de soro fetal bovino, ambas com antibióticos 1% de penicilina/estreptomicina em incubadora umidificada com 5% de CO₂ a 37°C. A troca do meio de cultura foi realizada a cada três dias para a manutenção das células. Essas foram coletadas na fase logarítmica, confluência 80 a 90% e 0,05% de tripsina/EDTA. Células para manutenção da cultura foram subcultivadas a 1×10^6 em placas de 55 cm². As células primárias de músculo liso de artéria mamária foram doadas pelo Prof. Dr. Alexandre Pereira do Laboratório de Genética e Cardiologia Molecular. Mais informações podem ser obtidas na publicação de Vaquero *et al* (2012).

Para a validação das junções de *splicing* específicas de cada isoformas utilizamos duas linhagens celulares, HCT-116 e SK-H-SN, obtidas do Cell Bank do RIKEN-IMS. A linhagem HCT-116 foi mantida em meio DMEM (*high glucose*) suplementado com 10% de FBS e a SK-H-SK foi mantida em MEM (*Minimum Essential Medium Eagle*) suplementado com 10% de FBS, 1% de penicilina/estreptomicina e 1x L-Glutamina MEM. As células foram cultivadas para extração de RNA (Maxwell RSC, SimplyRNA Tissue Kit, Promega) e síntese

de cDNA (kit PrimeScript 1st cDNA Synthesis Takara).

4.2.2 Síntese de oligonucleotídeos e PCR

Com os dados sobre as isoformas do gene *P4HB* em humanos obtidos de banco de dados como ENSEMBL (<http://www.ensembl.org>) e The Human Protein Atlas, foram desenhados oligonucleotídeos em regiões específicas e conservadas das variantes. Para análise quantitativa (qPCR) de variantes de *splicing* (Vandenbroucke *et al.*, 2001) como também para a validação das junções, projetamos um conjunto de primers para cada uma das três isoformas analisadas na junção específica. Para evitar a co-amplificação de outros transcritos, o requisito era que o primer se sobrepusesse à junção de *splicing* pelo menos 8 bases na extremidade 3' e 8 bases na extremidade 5'. Reações de PCR foram feitas para amplificar os fragmentos e a análise do comprimento do fragmento foi realizada para confirmar a amplificação. A reação de PCR foi feita de acordo com o protocolo da Platinum Taq polymerase (Life Technologies). Os clones foram selecionados e após a extração de DNA plasmidial o material foi purificado e sequenciado.

Para os experimentos de PCR quantitativa em tempo real foram usados os primers descritos na Tabela 1.

Tabela 1. Lista de oligonucleotídeos usados nesse estudo.

Oligo	Gene	Sequência	Amplicon
1	<i>P4HB</i> _exon3_fwd	5' GAGAGGCTGATGACATCGTG 3'	97bp
2	<i>P4HB</i> _exon3_rvs	5' GACTCCACCAAGGACTCTGC 3'	
3	<i>P4HB</i> -02 éxon 5/6 fwd	5' TCACCGAGCAGAGTGTGTCTG 3'	89pb
4	<i>P4HB</i> -02 éxon 6/7 rvs	5' GATGAACAGGATCTTGCCCTTG 3'	
5	<i>P4HB</i> -021 éxon 2 fwd	5' TATCCCACCATCAAGTTCTTCAG 3'	148pb
6	<i>P4HB</i> -021 éxon 2/4 rvs	5' CGACTCCACGTCACCTGTATATT 3'	
7	<i>P4HB</i> -27 éxon 2/3 fwd	5' GAATATACAGCTGCAGAGTCC 3'	211pb
8	<i>P4HB</i> -27 éxon 4/5 rvs	5' CTCATCAAACCTTCTTAAAG 3'	
9	HPRT1 fwd	5' CGTCTTGCTCGAGATGTGATG 3'	61pb
10	HPRT1 rvs	5' GCACACAGAGGGCTACAATGTG 3'	
11	CHOP fwd	5' AAGGCACTGAGCGTATCATGT 3'	105pb
12	CHOP rvs	5' TGAAGATACACTTCCTTCTTGAAC 3'	
13	GRP94 fwd	5' GCTTCGGTCAGGGTATCTTT 3'	111pb
14	GRP94 rvs	5' GGCTCTTCTTCCACCTTTGC 3'	
15	GRP78 fwd	5' CACAGTGGTGCCTACCAAGAAG 3'	151pb

16	GRP78 rvs	5' AGCAGGAGGAATTCCAGTCAGA 3'	
17	ATF6 fwd	5' CCGTATTCTTCAGGGTGCTC 3'	101pb
18	ATF6 rvs	5' CACTCCCTGAGTTCCTGCTG 3'	
19	NOX1fwd	5' CTCTCTCCTGGAATGGCA 3'	156pb
20	NOX1 rvs	5' GACCATCCACTTCAATCC 3'	
21	NOX2 fwd	5' TGCCTTTGAGTGGTTTGCAGAT 3'	121pb
22	NOX2 rvs	5' ATTGGCCTGAGACTCATCCCA 3'	
23	NOX4 fwd	5'-TGTGCCGAACACTCTTGGC 3'	127pb
24	NOX4 rvs	5' ACATGCACGCCTGAGAAAATA 3'	

4.2.3 Extração de RNA.

A extração de RNA das linhagens celulares foi feita com o kit Ilustra RNAspin Mini kit (Cat.No. 25-0500-70). As amostras foram quantificadas com o espectrofotômetro Nanodrop (Thermo Scientific) para avaliar com a qualidade da amostra com os índices A260/280 que devem produzir uma relação ao redor de 1,8 que indica a pureza das amostras obtidas. Qualidade e integridade das amostras foram avaliadas por gel de agarose 2% contendo 0,5 µg/ml de brometo de etídio. O tampão usado foi o TBE 1x e a eletroforese a 100 volts por 15 min. As amostras foram armazenadas no freezer -80°C.

4.2.4 Síntese de cDNA e qPCR

A reação para a síntese de cDNA foi realizada a partir 200 ng de RNA usando o kit SuperScript II reverse transcriptase (Cat.No.18064-014, *Life Technologies*). Para a PCR quantitativa foi usado o sistema SYBR com o kit Platinum SYBR Green qPCR SuperMix-UDG (Cat. No.11733-033, *Life Technologies*) e corante ROX, como referência passiva. Para a amplificação, usamos o seguinte programa: 55°C por 2 min, 95°C por 2 min, 40 ciclos de 95°C por 15 seg e 61°C por 1 min. Para cada par de oligonucleotídeos, foram testadas diferentes concentrações de oligonucleotídeos e a eficiência de amplificação de cada fragmento. A eficiência da reação foi avaliada com a construção de uma curva padrão feita com diluições seriadas de cDNA. O gene constitutivo *HPRT1* (hypoxanthine phosphoribosyltransferase 1) foi usada como normalizador da expressão de mRNA. Todas as amostras foram aplicadas em triplicata. A expressão gênica relativa foi calculada de acordo

com o método C^t ($2^{-\Delta\Delta C^t}$) (Pfaffl *et al.*, 2001). Todos os experimentos foram realizados em triplicata.

4.2.5 Ensaio de carenciamento de soro, tratamento com tunicamicina e exposição a CoCl_2

Para os experimentos com carenciamento de soro, as células foram inicialmente subcultivadas na densidade de 2×10^6 /placa e permaneceram 24h em meio com soro para a adesão celular. Após esse período, foi realizada a troca do meio de cultura e privação de soro e/ou tratamento, seguido de coleta (tempo 16h, 24h ou 40h).

Para os experimentos com tratamento com tunicamicina, as células foram privadas de soro durante 24h, sendo um grupo tratado com tunicamicina por 16h e 40h (0,5 $\mu\text{g/mL}$, 1 $\mu\text{g/mL}$ and 2 $\mu\text{g/mL}$ – um indutor de estresse do retículo endoplasmático) e outro grupo controle somente com privação de soro por 16h e 40h. Para avaliar o estresse do retículo endoplasmático foram usados oligonucleotídeos para os genes *ATF6*, *CHOP*, *GRP78* e *GRP94*.

Foram também realizados ensaios com células HEK293 e células musculares lisas primárias submetidas ao tratamento com CoCl_2 , reagente que mimetiza a hipóxia em cultura celular. Para o tratamento, as células foram privadas de soro durante 24h, sendo um grupo tratado com cloreto de cobalto (250 μM) por 10h e 24h para HEK293, e um grupo de 24h para células musculares lisas. O grupo controle foi submetido somente à privação de soro por 24h e depois submetidas a condições normais de crescimento (meio de cultura específico e 10% de soro fetal bovino) e amostras coletadas em 10h e 24h para HEK293 e 24h para células musculares lisas.

Todas as células após o tratamento ou carenciamento foram lavadas com solução salina tamponada com fosfato (PBS; pH 7,4) e depois recolhidas para extração de RNA e mantidas a -80°C .

4.2.6 Validação experimental das junções de splicing alternativo

Os fragmentos foram amplificados utilizando primers específicos para cada junção de *splicing*. Um gel de agarose a 2,0% foi utilizado para detectar o fragmento, seguido da coloração e fotodocumentação do gel. O produto de PCR foi purificado e clonado no vector

pGEM-T (Promega). Miniprep foi realizado para a extração de DNA plasmidial com o kit de QIAprep Spin Miniprep (Qiagen). As amostras foram sequenciadas através do sequenciamento Sanger. Para o protocolo de sequenciamento, realizamos uma PCR usando o tampão de sequenciamento BigDye (5X), tampão Ready Reaction (2,5X), oligonucleotídeos T7 ou SP6 (3.2 pmol), 30 ng template e 20 µL volume final da reação. Para a amplificação, usamos o seguinte programa: denaturação inicial 95°C por 1 min e 30 seg seguida por 35 ciclos de 95°C por 18 seg, 50°C por 12seg e 60°C por 4 min. Após a reação, o produto foi purificado utilizando a coluna Centri-Sep, Princeton Separations (Applied Biosystems).

4.2.7 Clonagem de novas variantes de splicing *P4HB* em humanos

Duas isoformas diferentes e o gene *P4HB* foram clonados (*GeneScript*) no vetor pFC14a *HaloTag*® CMV Flexi Vector (N° G282A, Promega), com o *HaloTag* no local do C-terminal da proteína. Esse vetor é usado para a expressão de proteínas e também para a potencial localização subcelular da proteína de interesse. A marcação pode ser realizada com um ligante fluorescente (*HaloTag*® *TMR Direct*, excitação: 552 nm e emissão: 578 nm) covalentemente marcada a proteína de fusão *HaloTag*®. As isoformas selecionadas foram: *P4HB*-02 (825 pb) e *P4HB*-021 (1419 pb). A sequência de *P4HB*-021 clonada nesse vetor foi usada como template para validação da junção de *splicing* específica. Para a isoforma *P4HB*-021 a junção específica é entre o éxon 2 e 4 do gene da *P4HB* com a ausência do éxon 3.

4.2.8 Análise estatística

As análises estatísticas foram realizadas usando o programa GraphPad Prism versão 6.0 (GraphPad Softwares Inc., La Jolla, USA). A normalidade dos dados foi verificada através do teste de Kolmogorov-Smirnov. As variáveis que apresentaram distribuição normal usou-se ANOVA de duas vias seguido pelo teste *post hoc* de Bonferrini. Para os dados não paramétricos utilizamos o teste de Kruskal Wallis com nível de significância 0,05. Os dados são descritos como média $\pm$ desvio padrão ou somente média.

4.3 Modelagem estrutural das isoformas

Para a construção de modelos proteicos das isoformas de *P4HB* foi utilizada a plataforma SWISS-MODEL (Benkert *et al.*, 2011; Guex *et al.*, 2009; Bertoni *et al.*, 2017; Bienert *et al.*, 2017; Waterhouse *et al.*, 2018). Disponível no site <https://swissmodel.expasy.org/>. A construção de modelos por homologia é uma técnica útil para prever a estrutura de proteínas-alvo. Foram utilizadas as sequências obtidas no banco de dados da UniProtKB (<http://www.uniprot.org>), com o número de identificação da proteína mostrado à Tabela 2. As sequências usadas para a predição da isoforma e da versão canônica estão anotadas no UniProtKB/Swiss-Prot.

Para os modelos já descritos na literatura o banco de dados de proteína RCSB PDB (Protein Data Bank, <http://www.rcsb.org/pdb/home/home.do>) foi usado. A estrutura da tiorredoxina oxidada (Pan & Bardwell, 2006) foi usada na Figura 6E como base de comparação com a PDI, pois contém a motivo C-X-X-C e mostra o enovelamento peculiar à família.

Resultados

5. Resultados

5.1 *Splicing* alternativo do gene *P4HB*

O gene *P4HB* possui 24 transcritos no genoma humano (ENSEMBL, versão GRCh38.p10), compreendendo no total a isoforma canônica, dez seqüências codificantes de proteínas, um transcrito com códons de parada prematuro (estes geralmente são eliminados via processo de decaimento por mutação sem sentido - nonsense mediated decay), três transcritos processados e nove transcritos com retenção de introns. As principais variantes descritas como codificantes de proteínas (ENSEMBL) são mostradas na Figura 1 e estão disponíveis em <http://www.ensembl.org> (Zerbino *et al.*, 2018). Todas estas 10 isoformas são corroboradas pelo *The Human Protein Atlas* (<http://www.proteinatlas.org>) e anotados no UniProt. A Tabela 1 resume as informações sobre as variantes de *splicing* da *P4HB*, incluindo número de identificação do transcrito (ENSEMBL), número de identificação da UniProt, comprimento de nucleotídeo e proteína, massa molecular e possível presença do peptídeo de sinal.

A PDI humana apresenta os domínios mostrados à Figura 1, em sequência: *a*, *b*, *b'*, *x*, *a'*, *c*. As isoformas presentes na Tabela 2 estão organizadas na Figura 4 apresentando um arranjo diferencial de domínios, com certos domínios presentes apenas parcialmente ou ausentes. Considerando que a atividade isomerase da PDI, a mais característica e típica da família, requer a presença simultânea e cooperativa dos dois domínios tiorredoxina redox-ativos (*a* e *a'*), a ausência de quaisquer destes domínios indica potencial mudança de atividade da isoforma específica. Como os domínios *b* e *b'* são responsáveis pela ligação ao substrato, deleções destas seqüências podem mudar a afinidade da isoforma da PDI por substratos específicos. Além disso, como a mobilidade do segmento *x* é importante para a atividade da PDI, alterações em sua estrutura proporcionarão mudanças importantes na atividade enzimática e conformação geral da molécula. A organização dos domínios protéicos prevista para cada isoforma é representada na Figura 4. É importante ressaltar que, embora o *P4HB-02* não exiba, de forma previsível, o códon de inicialização ATG clássico, foi possível detectar CAGE tags naquela região. Além disso, exceto para *P4HB-02* e *P4HB-021*, não se prevê que as isoformas tenham um códon de parada clássico, enquanto *P4HB-023* (detectado) e *P4HB-027* (possivelmente) tenham códons de parada nas regiões 3'-UTR. O tamanho da UTR 5' em algumas isoformas pode regular a transcrição e tradução dos genes e produtos protéicos.

Alguns estudos indicam que variações da UTR 5' em genes como *TGF-beta*, *BRCA1* e *MDM2*, que apresentam UTR curtas, são traduzidas mais eficientemente (Pickering *et al.*, 2005). As UTR influenciam a expressão gênica, modulando a estabilidade do mRNA e/ou a eficiência de tradução (Palaniswamy *et al.*, 2010).

Com exceção da isoforma *P4HB*-021, que apresenta todos os domínios da tiorredoxina (embora com truncamento parcial dos domínios *a* e *b*), os produtos previstos para as isoformas *P4HB* carecem de um ou mais domínios ou apresentam formas incompletas de alguns domínios, gerando combinações variáveis com potencial para exibir funções alternativas, uma vez que a atividade tiol isomerase exclusiva da PDIA1 requer todos os 4 domínios tiorredoxina (*a*, *a'*, *b*, *b'*) (Gilbert *et al.*, 2011). Por exemplo, *P4HB*-19 tem um domínio fragmentado sem 36 aminoácidos que incluem o motivo redox-ativo e *P4HB*-27 tem uma arquitetura truncada, com um motivo ativo-redox e um domínio *b* truncado. A isoforma *P4HB*-02 mostra o segundo sítio ativo (*a'*) contendo duas cisteínas (CGHC). Essa isoforma não tem a sequência de peptídeo de sinal e mostra a sequência C-terminal KDEL de direcionamento para o retículo endoplasmático (ER). Apresenta o domínio *b* parcialmente, e no domínio *b'* mostra a ausência de 10 aminoácidos. É a única da lista de isoformas que não tem códon de início de tradução, que está presente em todas as outras. A isoforma *P4HB*-023 possui somente o domínio *a*, com o sítio ativo CGHC e é o menor em número de aminoácidos. A *P4HB*-024, por outro lado, tem o peptídeo de sinal e o primeiro sítio ativo CGHC. Não tem, entretanto, o domínio *a'* e toda a sequência C-terminal incluindo o sítio KDEL, além da ausência de 90 aminoácidos entre o domínio *a* e *b*. Já a isoforma *P4HB*-025 possui somente o domínio *a* e *a'* contendo parcialmente um sítio ativo. Um caso mais comum de variantes derivadas de *splicing* é do *P4HB*-026, que usa um sítio alternativo de início de tradução, próximo ao domínio *b*, que está presente parcialmente e possui o domínio completo *b'*. *P4HB*-026 e *P4HB*-021 são as únicas que possuem o domínio *b'* completo. Recentemente, um estudo publicado por Wang *et al.* (2019) mostrou a importância do domínio *b'* na ligação com substratos, nesse caso com plaquetas ativadas. A isoforma *P4HB*-027 tem o domínio *a* e *b* parcialmente presentes com um sítio ativo. A última da lista, *P4HB*-029 possui os três domínios *a*, *b* e *b'* parcialmente sem o sítio ativo.

Prevê-se que *P4HB*-21 tenha o peptídeo de sinal, os 2 domínios CGHC redox-ativos e apresente apenas uma ausência de 44 aminoácidos entre os domínios *a* e *b*. *P4HB*-02 e *P4HB*-021 são os únicos a exibir o motivo C-terminal intacto com a sequência KDEL, indicando em geral que eventuais produtos proteicos gerados a partir de outras isoformas não exercem atividade isomerase no lúmen do retículo endoplasmático.

Além da PDIA1, um levantamento realizado para a família da PDI em humanos com base na plataforma *The Human Protein Atlas* (HPA) e ENSEMBL mostrou que, dentre os 21 membros da família, a PDIA1 (que é o protótipo ou o fundador da família) possui as dez isoformas preditas discutidas acima, enquanto a PDIA6 tem seis isoformas. O restante dos membros apresenta em média de uma a três isoformas. Estudos atualizados da Encyclopedia of DNA Elements (ENCODE) indicam que em média, existem sete transcritos variantes codificados para cada gene codificante de uma proteína (Dunhan *et al.*, 2012; Sveen *et al.*, 2016).

Figura 3. Estrutura gênica do gene *P4HB* humano e suas variantes de *splicing* preditas. Os éxons são representados por caixas e o UTR (região não traduzida) da extremidade 5' ou 3'. ATG indica o códon de partida e STOP, o códon de parada. O trecho sublinhado na *P4HB-23* indica a região traduzida com um códon de parada dentro da UTR. Na sequência de codificação de *P4HB-027* está incluída a região UTR 3' (região não traduzida) que não contém um códon de terminação identificado

Tabela 2. Variantes de *splicing* do gene *P4HB* em Humanos

Gene/ Variante de <i>splicing</i>	Transcrito ID	Comprimento (bp)	UniProt	Comprimento (aa)	Massa molecular (kDa)	Peptídeo sinalização (predição)
<i>P4HB-001</i>	ENST00000331483.4	2603bp	A0A024R8S5 / P07237	508	57.1	Sim
<i>P4HB-002</i>	ENST00000415593.1	1669bp	H0Y3Z3	274	31.4	Não
<i>P4HB-019</i>	ENST00000574914.1	571bp	I3L3P5	156	16.9	Sim
<i>P4HB-021</i>	ENST00000439918.2	1502bp	H7BZ94	464	52.5	Sim
<i>P4HB-022</i>	ENST00000573778.2	688bp	I3L312	208	23	Sim
<i>P4HB-023</i>	ENST00000576390.1	439bp	I3L514	106	11.6	Sim
<i>P4HB-024</i>	ENST00000576052.1	562bp	I3NI03	166	18.5	Sim
<i>P4HB-025</i>	ENST00000571617.1	505bp	I3L0S0	148	16.2	Sim
<i>P4HB-026</i>	ENST00000576380.1	943bp	I3L398	200	23	Não
<i>P4HB-027</i>	ENST00000576541.1	571bp	I3L3U6	188	20.7	Sim
<i>P4HB-029</i>	ENST00000570907.1	353bp	I3L4M2	117	13.5	Não

Para os experimentos subsequentes, focamos na investigação de 3 isoformas específicas de acordo com as seguintes características (discutidas em seguida): (1) dados de expressão gênica em bancos de dados de RNA-seq, (2) tamanho da proteína compatível com bandas detectadas em ensaios western prévios do grupo, (3) evidência de transcrito e proteína e (4) presença de códon de início de tradução e de parada. A primeira isoforma escolhida foi a *P4HB-02*, devido aos dados de expressão gênica como descrito no item 3.4. O tamanho da isoforma *P4HB-021* se mostrou semelhante a achados recorrentes em vários estudos prévios no nosso laboratório (análise de *western blot*). A isoforma *P4HB-027* tem um produto de 188 aminoácidos e possui o códon de início de tradução, mas não o códon de parada detectado na região UTR 3'.

Figura 4. Organização dos domínios da proteína PDI e as isoformas de proteínas preditas. A proteína PDI é composta por cinco domínios: *a*, *b*, *b'*, *x*, *a'* e *c*. O domínio catalítico da tiorredoxina (CGHC) é representado por domínios *a* e *a'* e os domínios não catalíticos, *b* e *b'*, em caixas cinzentas. O x-linker está localizado entre *b'* e *a'*, o que confere flexibilidade à proteína PDI. O domínio *c* contém a sequência KDEL, um sinal de retenção ER.

5.2 Alinhamento das sequências de aminoácidos

O alinhamento das sequências de aminoácidos da PDI e suas três isoformas foi realizado com a ferramenta *Align* usando o programa Clustal Omega no site da UniProt (<http://www.uniprot.org/align/>). As sequências de identificação das proteínas (Tabela 2). Os números de identificação da PDIA1 completa e das isoformas *P4HB-02*, *P4HB-021* e *P4HB-*

027 são mostrados, respectivamente, como: P07237, H0Y3Z3, H7BZ94, I3L3U6.

P07237 PDIA1 HUMAN	1	MLRRALLCLAVAALVRADAPEEEDHVLVLRKSNFAEALAAHKYLLVEFYAPWCGHCKALA	60
H0Y3Z3 H0Y3Z3 HUMAN	1	-----	0
H7BZ94 H7BZ94 HUMAN	1	MLRRALLCLAVAALVRADAPEEEDHVLVLRKSNFAEALAAHKYLLVEFYAPWCGHCKALA	60
I3L3U6 I3L3U6 HUMAN	1	MLRRALLCLAVAALVRADAPEEEDHVLVLRKSNFAEALAAHKYLLVEFYAPWCGHCKALA	60
P07237 PDIA1 HUMAN	61	PEYAKAAGKCLKAEGSEIRLAKVDATEESDLAQYGVGRGYPTIKFFRNGDTASPKKEYTAGR	120
H0Y3Z3 H0Y3Z3 HUMAN	1	-----	0
H7BZ94 H7BZ94 HUMAN	61	PEYAKAAGKCLKAEGSEIRLAKVDATEESDLAQYGVGRGYPTIKFFRNGDTASPKKEYTG--	118
I3L3U6 I3L3U6 HUMAN	61	PEYAKAAGKCLKAEGSEIRLAKVDATEESDLAQYGVGRGYPTIKFFRNGDTASPKKEYTA--	118
P07237 PDIA1 HUMAN	121	EADDIVNWLKRTGPAATTLPGDAAAESLVESSEVAVIGFFKDVESDSAKQFLQAAEAID	180
H0Y3Z3 H0Y3Z3 HUMAN	1	-----	0
H7BZ94 H7BZ94 HUMAN	119	-----DVESDSAKQFLQAAEAID	136
I3L3U6 I3L3U6 HUMAN	119	-----AESLVESSEVAVIGFFKDVESDSAKQFLQAAEAID	153
P07237 PDIA1 HUMAN	181	DIPFGITSNSDVFSKYQLDKDGVVLFKKFDEGRNFEGEVTKENLLDFIKHNQLPLVIEF	240
H0Y3Z3 H0Y3Z3 HUMAN	1	-----XFEGEVTKENLLDFIKHNQLPLVIEF	26
H7BZ94 H7BZ94 HUMAN	137	DIPFGITSNSDVFSKYQLDKDGVVLFKKFDEGRNFEGEVTKENLLDFIKHNQLPLVIEF	196
I3L3U6 I3L3U6 HUMAN	154	DIPFGITSNSDVFSKYQLDKDGVVLFKKFDEGRN-----	188
P07237 PDIA1 HUMAN	241	TEQTAPKIFGGEIKTHILLFLPKSVSDYDGKLSNFKTAESFSGKILFIFIDSDHTDNQR	300
H0Y3Z3 H0Y3Z3 HUMAN	27	TEQ-----SVSDYDGKLSNFKTAESFSGKILFIFIDSDHTDNQR	66
H7BZ94 H7BZ94 HUMAN	197	TEQTAPKIFGGEIKTHILLFLPKSVSDYDGKLSNFKTAESFSGKILFIFIDSDHTDNQR	256
I3L3U6 I3L3U6 HUMAN	189	-----	188
P07237 PDIA1 HUMAN	301	ILEFFGLKKEECPAVRLITLLEEMTKYKPESEELTAERITEFCHRFLEGKIKPHLMSQEL	360
H0Y3Z3 H0Y3Z3 HUMAN	67	ILEFFGLKKEECPAVRLITLLEEMTKYKPESEELTAERITEFCHRFLEGKIKPHLMSQEL	126
H7BZ94 H7BZ94 HUMAN	257	ILEFFGLKKEECPAVRLITLLEEMTKYKPESEELTAERITEFCHRFLEGKIKPHLMSQEL	316
I3L3U6 I3L3U6 HUMAN	189	-----	188
P07237 PDIA1 HUMAN	361	PEDWDKQPVKVLVGKNFEDVAFDEKKNVFEFYAPWCGHCKQLAPIWDKLGETYKDHENI	420
H0Y3Z3 H0Y3Z3 HUMAN	127	PEDWDKQPVKVLVGKNFEDVAFDEKKNVFEFYAPWCGHCKQLAPIWDKLGETYKDHENI	186
H7BZ94 H7BZ94 HUMAN	317	PEDWDKQPVKVLVGKNFEDVAFDEKKNVFEFYAPWCGHCKQLAPIWDKLGETYKDHENI	376
I3L3U6 I3L3U6 HUMAN	189	-----	188
P07237 PDIA1 HUMAN	421	VIAKMDSTANEVEAVKVHSFPTLKFFPASADRTVIDYNGERTLDGFKKFLESQGGQDGAGD	480
H0Y3Z3 H0Y3Z3 HUMAN	187	VIAKMDSTANEVEAVKVHSFPTLKFFPASADRTVIDYNGERTLDGFKKFLESQGGQDGAGD	246
H7BZ94 H7BZ94 HUMAN	377	VIAKMDSTANEVEAVKVHSFPTLKFFPASADRTVIDYNGERTLDGFKKFLESQGGQDGAGD	436
I3L3U6 I3L3U6 HUMAN	189	-----	188
P07237 PDIA1 HUMAN	481	DDDLLEDLEAAEFPDMEEDDDQKAVKDEL	508
H0Y3Z3 H0Y3Z3 HUMAN	247	DDDLLEDLEAAEFPDMEEDDDQKAVKDEL	274
H7BZ94 H7BZ94 HUMAN	437	DDDLLEDLEAAEFPDMEEDDDQKAVKDEL	464
I3L3U6 I3L3U6 HUMAN	189	-----	188

Figura 5. Alinhamento das sequências de aminoácidos de PDI e suas isoformas. Comparação entre a PDI (P07237) de humano e as três isoformas: *P4HB*-02 (H0Y3Z3), *P4HB*-021(H7BZ94) e *P4HB*-027 (I3L3U6). A linha preta indica o sítio ativo CGHC e a linha pontilhada a sequência KDEL

O perfil de organização e o resultado do alinhamento das sequências de aminoácidos mostrados às Figuras 4 e 5 corroboram a noção de que as isoformas observadas não correspondem provavelmente a modificações pós-traducionais. Conforme discutido acima, a estrutura das isoforma *P4HB*-021 é consistente com um produto de *splicing* alternativo (éxon skipping). Por outro lado, a mesma consistência não está evidente para as isoformas *P4HB*-02 e *P4HB*-027, considerando que *P4HB*-02 não mostra claramente o códon de iniciação (metionina) e *P4HB*-027 não tem um stop-códon, como discutido anteriormente.

5.3 Análise de CAGE tags

Utilizando-se das informações disponíveis na plataforma ZENBU sobre CAGE tags, analisamos uma região a montante da variante de *splicing* *P4HB*-02. Para o *P4HB*-02, usamos 250pb a montante da região de codificação putativa para verificar quais amostras apresentou um maior número de CAGE tags (unidade TPM, tags per million). Para isto, um subconjunto de dados de FANTOM5 chamado FANTOM5 CAGE Phase1 CTSS (CAGE transcription start site) foi analisado. As amostras que apresentaram maior TPM foram amostras de pâncreas, células de Sertoli, células musculares lisas (aórtica), linhagem de leiomioma e fibroblastos (adventícia aórtica) (Figura 2). Além disso, as informações do ENCODE CAGE também foram analisadas mostrando as linhagens celulares Hepg2, K562, HUVEC e Nhek para *P4HB*-02 (Figura S2). Checamos a presença de CAGE tags somente a montante da isoforma *P4HB*-02 com o objetivo de verificar uma especificidade de expressão para essa isoforma, no entanto, as demais isoformas que se iniciam na mesma região que o gene *P4HB* também se apresentaram tags similares. O benefício dos CAGE tags foi revelar uma gama de eventos alternativos de iniciação da transcrição (TSS), mesmo em regiões codificantes, em éxons, como ocorre no caso da *P4HB*-02 (Carninci *et al.*, 2005; Mercer *et al.*, 2010). Os dados de CAGE tags são importantes, pois evidencia outro nível de regulação, o uso de sítios de início de transcrição alternativos. Estes dados foram importantes para selecionar amostras do ENCODE Consortium (dados de RNA-seq) que apresentaram um maior nível de expressão.

As amostras da Figura 6 apresentam linhagens celulares e, em alguns casos, frações celulares diferentes como: citoplasmática, nuclear ou total.

Figure 6. Amostras representativas para a isoforma P4HB-02 de um subgrupo do (A) FANTOM5 (FANTOM CAGE) e amostras do (B) ENCODE CAGE. CAGE *tags* estão representadas como TPM (*tags* por milhão)

5.4 Perfil de expressão das variantes do gene *P4HB* em amostras de RNA-seq do banco de dados FANTOM5

Em seguida, focando em um aspecto importante da caracterização de distintas isoformas, abordamos seu perfil de expressão em diferentes linhagens celulares e tecidos, usando bancos de dados distintos: FANTOM5, ENCODE e GTEx.

Primeiro, o projeto FANTOM5 fornece um atlas de RNAs e microRNAs não codificadores e seus promotores, com dados de transcriptoma RNA-seq e micro-RNAs (Lizio *et al.*, 2019). Foi utilizada a informação de todas as bibliotecas de RNA-seq FANTOM5 (70 amostras) publicadas por Hon *et al.*, 2017, com o objetivo de analisar prospectivamente as junções de *splicing* de variantes do gene *P4HB*. Estas amostras são compostas por linhagens celulares ($n = 32$), células primárias (27) e tecidos (11). Em alguns casos ($n = 6$), utilizou-se a média das triplicatas, como no caso das amostras de sangue total, células B CD19 e células T CD8, provenientes do mesmo doador. Estas 70 amostras do projeto FANTOM5 foram usadas para construir as Figuras 7A e B, que mostram o perfil de expressão para as 10 isoformas do

gene *P4HB* descritos na Figura 3. Na Fig.8C, dois exemplos representativos das isoformas mais expressas (*P4HB-02* e *P4HB-021*) são mostrados para diferentes células e tecidos.

B

Média de SJ TPM das 70 amostras do FANTOM5

Figura 7. Distribuição da expressão das variantes de *splicing* do gene *P4HB* nas amostras de RNA-seq do FANTOM5. Junção de *splicing* em tag por milhão (SJ-TPM) é uma unidade para contar o número de uma junção de isoforma específica, normalizada pelo número total de *reads* para cada conjunto de dados de RNA-seq. (A) O gráfico mostra a distribuição das variantes do gene *P4HB*. O raio do círculo indica a expressão média de todas as amostras em SJ-TPM para cada variante de junção detectada. (B) Fração de abundância de isoformas em FANTOM5 compreendendo as 10 variantes de *splicing* de *P4HB*.

Figura 8. Perfil de expressão das variantes de *splicing* *P4HB-02* e *P4HB-021*. A expressão da variante *P4HB-02* (A) e *P4HB-021* (B) em três de amostras: células primárias, tecido e linhagem celular. (C) Diagrama representativo da expressão de *P4HB-021* no total de amostras analisadas. (D) Visualização do evento de *splicing* de *P4HB-021*; o diagrama azul indica parte do gene *P4HB*, na direção reversa do éxon 1 ao éxon 4. A seta preta indica o éxon 4 com uma junção de *splicing* no éxon 2. Na variante *P4HB-021* ocorre a ausência do éxon 3. O número total de dados de RNA-seq foram 70 amostras.

Estes resultados mostraram, resumidamente, a seguinte visão geral: dez isoformas do gene *P4HB* são expressas no conjunto de amostras do FANTOM5. A isoforma *P4HB-021* está significativamente representada, particularmente na célula do músculo liso da aorta, seguida por *P4HB-019*, *P4HB-023*, *P4HB-027* e *P4HB-02*. As amostras com número global mais elevado de isoformas de *P4HB* expressas são células do músculo liso da aorta, seguidas pelas células B CD19 e linhagem de adenocarcinoma mucinoso. A isoforma expressa com maior frequência em todas as 70 amostras foi *P4HB-02*, presente em 28 amostras, enquanto

P4HB-021 e *P4HB-027* representaram os SJ TPMs mais altos. A isoforma *P4HB-021* tem seu maior nível de expressão nas células musculares lisas da aorta, mas devido ao número de amostras, analisamos mais dados de células de músculo liso (SMC) no ENCODE e também em dados de RNA-seq em células de músculo liso vasculares publicado recentemente por Yu *et al*, 2018.

Para visualizar o arquivo BAM na plataforma IGV, selecionamos a variante *P4HB-027* para verificar a junção de *splicing* em 4 diferentes tipos de células. Além disso, na Figura 8 D, há um gráfico para o *P4HB-021* exibir o evento de *splicing*. *Sashimi* plot foram gerados no IGV-Sashimi, o que nos permite selecionar uma região genômica específica e detectar eventos de *splicing* (Katz *et al.*, 2015). A Figura 9 ilustra o evento de *splicing* no meio do éxon 3 do *P4HB-027*. Usando essa ferramenta, foi possível visualizar a junção de *splicing* de diferentes variantes de *splicing* entre múltiplas amostras. Além disso, a junção de *splicing* para *P4HB-027* no éxon 3 não está presente em todas as 4 amostras analisadas, como indicado pela seta preta (Figura 9).

Figura 9. Gráficos representativos de quatro amostras de RNA-seq de FANTOM5 mostrando quatro linhagens celulares diferentes para a região do cromossomo 17: 79796651-79822949 obtidas utilizando a ferramenta sashimi-plot no programa IGV. No painel superior apresenta-se o gene *P4HB* inteiro com 11 éxons (azul). No painel inferior, as linhas indicam o éxon 3, onde é a junção específica para *P4HB-027*. A seta preta indica a junção de *splicing* dessa isoforma.

5.5 Perfil de expressão das variantes do gene *P4HB* em amostras de RNA-seq do ENCODE

A Enciclopédia de Elementos de DNA (ENCODE, Djebali *et al.*, 2012) tem um conjunto amplo de dados com diferentes metodologias, como Exon Arrays, Chip-Seq e análise de RNA-seq, disponível em <http://www.encodeproject.org>. Aqui utilizamos os dados ENCODE Caltech RNA-seq e os dados CSHL/ENCODE RNA-seq para analisar 27 conjuntos de dados de RNA-seq incluindo 12 linhagens celulares diferentes, 5 das quais são linhagens celulares de câncer. A seleção das amostras foi justificada pela presença nos CAGE tags (Forrest *et al.*, 2014), que são tags para expressão gênica, conforme detalhado em Métodos. A distribuição das variantes de *splicing* foi calculada pela junção de *splicing* (SJ TPM) nos conjuntos de dados ENCODE é mostrada na Figura 10. Neste gráfico, o mais representativo (expresso na maioria das amostras) é o *P4HB*-029, mas as isoformas mais expressas (em SJ TPM) são *P4HB*-02 e 021. Em geral, HCT-116 (câncer de cólon humano), Gm12878 (linhagem celular linfoblástico humana), Hmsc (linhagem de células estaminais mesenquimais humanas) e Hsmm (linhagem celular de mioblastos do músculo esquelético humano) foram as células mais representadas neste conjunto (Figura 11).

Figura 10. Distribuição da expressão das variantes de *splicing* do gene *P4HB* em amostras de RNA-seq do ENCODE Consortium. Junção de *splicing* em tag por milhão (SJ-TPM) é uma unidade para contar o número junções de uma isoforma específica, normalizada pelo número

Em seguida, aplicamos o mesmo pipeline acima para identificar e contar a junção de *splicing* das isoformas do gene *P4HB* em conjuntos de dados de RNA-seq humanos de células primárias de músculo liso de artéria pulmonar de dois doadores do sexo masculino com idades de 26 e 28 anos disponíveis no ENCODE (<https://www.encodeproject.org/>). *P4HB-019*, 23 e 26 foram mais expressos nestas células (Figura 12C) representando cerca de 0,7% (*P4HB-019*) e 0,5% (*P4HB-023*) do total de *P4HB*. Em todos esses casos, no entanto, a expressão de isoformas é relativamente pequena em relação à isoforma canônica.

Figura 12. (A) SJ-TPM das variantes de *splicing* do gene *P4HB* em dois tipos celulares: células endoteliais (HUVEC e HAoEC) e fibroblastos adventiciais da aorta (HaoAF). (B) Percentagem de distintos tipos celulares para cada uma das variantes do gene *P4HB*. (C) SJ-TPM das variantes de *splicing* do gene *P4HB* em células primárias de músculo liso de artéria pulmonar.

5.6 Perfil de expressão de três variantes de *splicing* do gene *P4HB* em amostras de RNA-seq do GTEx

Com o objetivo de investigar um maior número de amostras, usamos os dados de RNA-seq do The Genotype Tissue Expression Project (GTEx) (<http://www.gtexportal.org/home/>), que é um dos amplos estudos de coortes que continuam gerando um número significativo de dados transcriptômicos, incluindo RNA-seq de vários tecidos. Neste estudo foram utilizados dados de 11690 RNA-seq de diferentes tecidos e condições listadas no Anexo C - Tabela 3. A Figura 13 mostra a quantificação de três variantes de *splicing* *P4HB* (*P4HB*-02, 021 e 027) em 30 tecidos diferentes. A

representatividade dessas isoformas em cada um dos tecidos é diferente. Entre as três variantes analisadas, *P4HB-02* e *027* apresentam uma fração maior que a *P4HB-021*. Entre os tecidos, a fração de isoforma de *P4HB-021* no coração e testículo é maior em relação a outros tecidos. Na Figura 13, analisamos os 30 tecidos, mesclando todas as sub-regiões de cada tecido (Anexo C - Tabela 3).

Na próxima análise, focando nas sub-regiões de cada tecido (coração e vasos sanguíneos), analisamos primeiro a expressão de isoformas em diferentes sub-regiões do coração (apêndice atrial e ventrículo esquerdo): não ocorreram diferenças entre as sub-regiões (Figura 14A). Estes conjuntos de dados são enriquecidos com amostras de sangue total (407 amostras), vasos sanguíneos (913 amostras) e coração (600 amostras). Já para os subtipos de vasos sanguíneos - artéria aorta ($n = 299$), artéria coronária ($n = 172$) e artéria tibial ($n = 400$), a fração *P4HB-021* é menos representada comparada às variantes *P4HB-02* e *027*. De acordo com as frações representadas, a variante de *splicing* *P4HB-027* apresenta uma fração mais abundante comparada à *P4HB-02* e *021* tanto nos vasos como no coração.

Na Figura 15, mostra a quantificação das variantes de *splicing* (*P4HB-02*, *P4HB-021* e *P4HB-027*) do gene *P4HB* em amostras de sangue total do conjunto de dados de RNA-seq do banco de dados GTEx. A expressão das três isoformas derivadas de *splicing* do gene *P4HB* apresentam valores de fração de isoformas significantes.

Figura 13. Quantificação das variantes de *splicing* *P4HB-02*, *P4HB-021* e *P4HB-027* no GTEx. Fração da isoforma é representada por SJ TPM da isoforma normalizada por *P4HB* (expressão). O número total de amostras foi de 11960 com 30 tecidos diferentes. A linha preta indica a média e o desvio padrão é representado em branco

Figura 14. Quantificação de variantes de *splicing* do gene *P4HB* em amostras de vasos sanguíneos e coração. Para detectar a fração de abundância de isoformas (*P4HB*-02, 021 e 027) normalizada pela fração do gene *P4HB*. (A) Fração de *P4HB*-02, 021 e 027 em vasos sanguíneos divididos em três subtipos: artéria aorta ($n = 299$), artéria coronária ($n = 172$) e artéria tibial ($n = 400$) (B) Fração de variantes de *splicing* *P4HB* em coração com duas sub-regiões: apêndice atrial ($n = 300$) e ventrículo esquerdo ($n = 300$)

Figura 15. Quantificação de variantes de *splicing* do gene *P4HB* em amostras de sangue total. Para detectar a fração de abundância de isoformas (*P4HB*-02, 021 e 027) normalizada pela fração do gene *P4HB* em amostras de sangue total (n=407). A linha preta indica a média e o desvio padrão é representado em preto. As diferenças estatisticamente significativas estão indicadas por letras. O teste de Kruskal Wallis foi usado.

5.7 Perfil de expressão das variantes do gene *P4HB* em amostras de RNA-seq de células de músculo liso vascular

Considerando nosso foco em células vasculares e os resultados acima apresentados (FANTOM5 e ENCODE), seguimos a análise da isoforma *P4HB* em VSMC. Um estudo publicado recentemente por Yu *et al.* (2018) produziram dados de RNA-seq de células musculares lisas vasculares da aorta e coronárias (VSMC) com o objetivo de investigar padrões de expressão gênica durante alterações na rigidez da matriz extracelular (MEC), uma vez que as interações mecanobiológicas da matriz extracelular estão envolvidas na fisiopatologia de doenças vasculares. A Figura 16A indica que a expressão de *P4HB* tende a ser menor sob condições patológicas (rígidas, *stiff*), em comparação com condições fisiológicas (complacentes, *soft*). Com respeito às variantes de *splicing* da *P4HB* nas VSMC da artéria coronária, o SJ TPM tende a ser mais alto nas condições fisiológicas. Similarmente, em VSMC de aorta proximal, as isoformas *P4HB*-02 e *P4HB*-021 são mais representativas nas condições fisiológicas.

Figura 16. Análise de dados de RNA-seq de um estudo (Yu *et al.*, 2018) usando células VSMC (células de músculo liso vasculares) mimetizando as condições patológicas (stiff) e fisiológicas (soft). (A) Expressão gênica do gene *P4HB* no experimento. É também mostrada a expressão de *P4HB* e suas variantes em VSMC de coronária (B) e VSMC de aorta proximal (C).

Em conjunto, uma conclusão parcial até o momento é que a análise por estas três distintas bases de dados reforça não apenas a expressão altamente tecido-específica de diferentes isoformas do gene *P4HB*, como também indica que as variantes de *splicing P4HB-02* e *P4HB-021* são particularmente importantes e comuns às três distintas bases. Embora nem todas as VSMC de localizações distintas tenham mostrado o mesmo perfil de expressão das isoformas, a isoforma *P4HB-021* foi particularmente expressa em um número considerável de amostras.

5.8 Identificação de isoformas do gene *P4HB* em camundongo e rato

Em camundongos, o número de isoformas do gene *P4HB* também é reduzido. Apresenta três isoformas, uma das quais é a versão canônica com 2861 pares de bases *P4HB-201* (ENSMUST000000026122.10) com 509 aminoácidos (UniProt: P09103); a segunda

codifica uma proteína de 165 aminoácidos (UniProt: E9Q8G8) com identificação do transcrito *P4HB*-203 (ENSMUST00000168360.1) com 770 pares de bases (pb); a terceira isoforma identificada como *P4HB*-202 (ENSMUST00000166620.1) não produz uma proteína e é classificada como transcrito processado com 414 pb. Esses dados estão disponíveis no site http://www.ensembl.org/Mus_musculus com a versão de montagem do genoma de camundongo, Mouse GRCm38.p6 (Genome Reference Consortium Mouse Reference 38).

Em ratos, usando o mesmo banco de dados (ENSEMBL), não identificamos outras isoformas, somente a canônica com 509 aminoácidos (UniProt: P04785) e um transcrito (ENSRNOT00000054958.2) com 2489 pb. Esses dados estão disponíveis no site http://www.ensembl.org/Rattus_norvegicus para a versão de montagem do genoma Rnor_6.0.

Estas análises indicam o gene *P4HB* evidencia formas de *splicing* alternativo em distintas espécies. No entanto, o número de isoformas codificantes de proteínas para o gene *P4HB* em humanos é significativamente maior, quando comparado a camundongos e ratos. A diferença no número variantes de *splicing* sugere que esse evento possa conferir uma vantagem evolutiva, como isoformas alternativas originando novas combinações de éxons, fornecendo uma oportunidade para a aquisição de novas funções (Krawczak *et al.*, 1992; Ward *et al.*, 2010).

5.9 Análise de validação das variantes de *splicing* de *P4HB*

Um aspecto técnico importante para a análise das isoformas da PDI por RT-PCR foi o desenho dos oligonucleotídeos, considerando as peculiaridades de cada isoforma de interesse. Os oligonucleotídeos foram desenhados especificamente para cada isoforma em regiões de intersecção entre os éxons. A isoforma *P4HB*-02 possui parcialmente o final do éxon 5 e final do éxon 6. A isoforma *P4HB*-021 não tem o éxon 3 e os oligonucleotídeos foram desenhados para a junção entre o éxon 2 e 4. (Figura 6). Os oligonucleotídeos amplificam fragmentos com os seguintes tamanhos: 89 pb (*P4HB*-02), 148 pb (*P4HB*-021) e 211 pb (*P4HB*-027).

Para certificar que esses fragmentos correspondiam com a junção específica para cada isoforma, os mesmos foram clonados em vetor de clonagem e sequenciados. Para isso, usamos cDNA de células de linhagem HCT-116 e células de neuroblastoma (SK-N-SH). Anteriormente, foram analisadas e confirmadas a presença daquelas isoformas nesses tipos celulares utilizando os dados de RNA-seq. Na Figura 17 é possível observar o fragmento de cada isoforma amplificada e seu respectivo tamanho.

Como o tamanho dos amplicons são pequenos, esses mesmos oligonucleotídeos foram

usados para analisar a expressão dessas três isoformas, descrita no tópico a seguir.

Esquema de três isoformas do gene *P4HB* em humanos

Figura 17. Amplificação por PCR da junção específica das variantes de *splicing*: *P4HB*-02, *P4HB*-021 e *P4HB*-027. Os fragmentos das isoformas *P4HB*-02 (89 bp) e 027 (211 bp) foram amplificadas em células HCT-116 e células SK-N-SH. Para a amplificação de *P4HB*-021, o fragmento esperado foi de 148 pb. A amplificação deste fragmento foi realizada utilizando um vetor pFN14a contendo a sequência da isoforma *P4HB*-021.

5.10 Expressão gênica das isoformas do gene da *P4HB* em linhagens celulares

Em seguida, buscamos a validação da detecção das variantes de *splicing* usando RT-qPCR, com foco VSMC primárias humanas da artéria mamária e da linhagem HEK-293 (rim embrionário humano). Em ambos os casos, as células foram investigadas em seu estado basal, após carenciamento de soro (16h ou 24h) e após exposição ao CoCl_2 (24h) como um mimético de hipóxia, uma vez que algumas PDIs são regulados por hipóxia (Tanaka *et al.*, 2000).

Importante, em linha com nossas análises anteriores nos bancos de dados, *P4HB*-021 foi claramente detectada em VSMC na condição basal e sua expressão, juntamente com a da isoforma canônica *P4HB*, foi positivamente regulada após 24 horas de carenciamento, quando comparada aos dados de 16 horas de carenciamento (Figura 18A). Com relação aos demais genes analisados (Figura 18B), as análises mostram aumentos na expressão do gene *GRP94*.

Mostra tendências a aumento de *P4HB* e *P4HB-021* e diminuição para as isoformas *P4HB-02* e *P4HB-027*, apresentam níveis de expressão baixos. Assim, como os genes da família NOX apresentaram baixa expressão. A exposição ao CoCl_2 , entretanto, não afetou de modo estatisticamente significativo a expressão de *P4HB* ou de suas variantes nas condições de nossos experimentos (Figura 19). O tratamento com CoCl_2 em VSMC não afetou os genes relacionados ao estresse do retículo endoplasmático. O gene ATF6 apresentou um aumento na expressão em relação aos outros genes analisados. Já os genes da família NOX, NOX2 e NOX4, apresentaram um aumento na expressão, mas não estatisticamente significativo.

Figura 18. qPCR de células de músculo liso de artéria mamária para o gene *P4HB*. Expressão gênica de (A) três variantes de *splicing* do gene *P4HB* e (B) 4 genes relacionados ao estresse do retículo endoplasmático submetidas a carençamento de soro por 16h e 24h. O experimento foi repetido três vezes. O asterisco indica diferença estatística entre 16h e 24h após carençamento.

A seguir, analisamos os efeitos de tunicamicina, um potente inibidor de *N*-glicosilação, em células HEK-293 para investigar a influência do estresse do retículo endoplasmático resultante na expressão de *P4HB* e suas variantes. As células foram incubadas por 24 horas com 3 concentrações de tunicamicina e as expressões dos genes foram então avaliadas (Figuras 20 e 21). Outros genes envolvidos na resposta de estresse do retículo endoplasmático (*unfolded protein response*), como *ATF6*, *CHOP*, *GRP78* e *GRP94*, assim como as NADPH oxidases *NOX1*, *NOX2* e *NOX4* foram também analisados. A expressão de *GRP78*, *CHOP* e *GRP94*, que são marcadores clássicos de estresse do retículo endoplasmático, aumentaram após 16h e 40h de exposição à tunicamicina, indicando que esta de fato induziu estresse.

Figura 19. qPCR de células de músculo liso de artéria mamária tratado com CoCl₂ durante 24h. (A) Expressão do gene P4HB e suas três variantes de *splicing* e (B) expressão de seis genes relacionadas com estresse do retículo endoplasmático em células VSMC tratada com CoCl₂ durante 24h (B). Ambos os experimentos foram repetidos três vezes. O teste não paramétrico de Kruskal Wallis ($p < 0,05$) foi usado.

Aumentos menos robustos foram também observados para o fator de transcrição *ATF6* e o gene da família Nox, NOX4. Enquanto as expressões do gene *P4HB* e de suas variantes mostrou tendência a aumento quando analisadas em relação ao controle com o aumento das concentrações de tunicamicina ao longo do tratamento (Figura 20). Já para os genes Nox1 e *CHOP*, a expressão mostrou-se estatisticamente significativa quando comparado a expressão em 16h de tratamento e 40h (Figura 21). Demonstrando que o tratamento com tunicamicina foi efetivo no aumento de expressão de genes relacionados ao estresse do RE. Nenhuma dessas diferenças, entretanto, foi estatisticamente significativa para o gene *P4HB*, assim como suas variantes de *splicing*, não são diretamente responsivos à *unfolded protein response* (como por exemplo, *GRP78*, *GRP94* e a outras PDIs). De modo análogo às VSMCs, a exposição ao CoCl₂ por 24h mostrou uma diferença pequena, mas não estatisticamente significativa, nas expressões de *P4HB* e suas variantes em células HEK-293 (Figura 22).

Figura 20. qPCR de células HEK-293 para o gene *P4HB* e as três variantes de *splicing* escolhidas tratadas com tunicamicina por 16h e 40h. A expressão relativa representada como amostras tratadas/não tratadas em três concentrações diferentes de tunicamicina (0,5 µg/mL, 1 µg/mL e 2 µg/mL). Nenhuma diferença estatística foi observada.

Figura 21. qPCR de células HEK-293 tratados com tunicamicina por 16 e 24h para a análise de expressão de sete genes. (A) HEK-293 tratados com tunicamicina por 16 e 24h (B) HEK-293 tratada com CoCl_2 durante 10h e 24h. As diferenças estatisticamente significativas estão indicadas por letras. O teste de Kruskal Wallis foi usado.

Figura 22. qPCR de células HEK-293 para a análise de expressão do gene *P4HB* e três variantes de *splicing*, *P4HB-02*, *P4HB-021* e *P4HB-027*. HEK-293 tratada com CoCl₂ durante 10h e 24h. O teste de Kruskal Wallis foi usado.

Na Figura 22 a expressão do gene *P4HB* e suas três variantes em células HEK-293 submetidas ao tratamento com CoCl₂ por 10h e 24h não apresentaram diferenças estatísticas em relação ao tempo e tratamento. Os demais genes analisados como *GRP78*, *GRP94*, *ATF6*, *CHOP*, *NOX1*, *NOX2* e *NOX4* não apresentaram diferenças de expressão estatisticamente significativas.

5.11 Modelagem virtual da estrutura protéica de *P4HB-021* comparada à *P4HB*

Dada a relevância da isoforma *P4HB-021* e sua peculiar estrutura primária, analisamos a sua modelagem tridimensional. Para analisar o modelo proteico da isoforma *P4HB-021* comparada à *P4HB*, utilizamos a plataforma SWISS-MODEL (<https://swissmodel.expasy.org>) usando as sequências de aminoácidos anotadas no UniProtKB. Na Figura 22 são apresentadas as estruturas previstas para *P4HB* e *P4HB-021*. A estrutura prevista para *P4HB* está de acordo com as outras estruturas já publicadas (Kozlov *et al.*, 2006; Inaba *et al.*, 2010), indicando acuidade de nosso modelo. Como discutido acima, a única diferença da *P4HB-021* com a *P4HB* é a ausência do exon 3, que causa uma mudança conformacional no domínio *a* e *b*. A ausência de 44 aminoácidos na transição desses domínios promove a ausência de uma α -hélice no domínio *a*, a ausência de duas folhas β e uma α -hélice no domínio *b* e a ausência de uma α -hélice no domínio *b'*. Tais alterações resultam primeiramente na perda da arquitetura do enovelamento tiorredoxina nos domínios *a* e *b* (Figura 23). Além disso, resultam em um segmento parcialmente não-estruturada na transição dos domínios *a-b*, que poderia se associar a um significativo incremento da mobilidade da proteína, de modo análogo ao já conhecido para o *x*-linker na transição dos domínios *b'* e *a'*. Esta maior mobilidade poderia causar a

acomodação de outros tipos de substratos maiores ou mais complexos do que aqueles normalmente acessíveis à *P4HB* canônica.

Figura 23. Modelagem proteica predita da isoforma *P4HB-021*. (A) Organização do domínio proteico do PDIA1 com a indicação de quatro domínios: *a*, *b*, *b'*, *a'*. (B) A variante de *splicing* *P4HB-021* mostra uma conformação diferente na transição entre os domínios *a* e *b* em comparação com PDIA1. O modelo foi construído usando a sequência UniProt e programa SWISS MODEL (<https://swissmodel.expasy.org>). A ausência de 44 aminoácidos entre os domínios *a* e *b* se associa a uma mudança estrutural do domínio *b*, o que poderia conferir mobilidade adicional de maneira análoga àquela fornecida pelo x-linker entre os

domínios b' e a'

(A) Estrutura da tioredoxina oxidada (Pan & Bardwell, 2006)

(B) Enovelamento tioredoxina (Martin, J. 1995)

Figura 24. (A) Estrutura da tioredoxina (12 kDa) e (B) enovelamento da tioredoxina.

A estrutura da tioredoxina oxidada (Figura 23a) é composta por quatro folhas β e rodeada por três α hélices. O enovelamento tioredoxina é o princípio estrutural básico da superfamília da tioredoxina e da família das PDIs (Pan & Bardwell, 2006). A Figura 23b apresenta o enovelamento tioredoxina que contém as α -hélices em cinza ($\alpha 1$ e $\alpha 3$), estão atrás das folhas β , enquanto que a α -hélice ($\alpha 2$) está à frente das folhas β (Martin, J. 1995).

Os domínios (unidades compactas dentro do padrão de enovelamento de uma cadeia única que parecem possuir estabilidade independente) são mais similares entre *P4HB-021* e a PDI humana e a isoforma *P4HB-021* preserva somente o enovelamento tioredoxina no domínio a' .

Discussão

Discussão

O *splicing* alternativo expande significativamente o perfil de proteínas codificadas a partir de um subconjunto de genes, proporcionando maior diversidade de isoformas e subtipos de proteínas potencialmente envolvidas em funções celulares específicas. Estas incluem a sinalização subcelular envolvida em funções restritas de órgãos e particularmente em translocação para locais subcelulares distintos, como núcleo (Mussil *et al.*, 2013), mitocôndrias (Casas *et al.*, 1999) e aparelho de Golgi (Li *et al.*, 2019). Esses processos aumentam muito o potencial de adaptabilidade a condições externas distintas (Chisa & Burke, 2007), mas também podem contribuir para a fisiopatologia de doenças por termo conhecido como “isoform switch” (Vitting-Seerup & Sandelin, 2017). Transcritos alternativos também podem regular a expressão do gene canônico, como por exemplo, no caso do gene CTCF (Li *et al.*, 2019). Assim, o conhecimento das isoformas alternativas de um dado gene é crucial para entender as implicações de sua regulação gênica. Neste estudo, realizamos uma análise abrangente do perfil de *splicings* alternativos do gene *P4HB*. Dadas as importantes e múltiplas funções do gene *P4HB*, a investigação de seu panorama de *splicing* alternativo é particularmente relevante. Além disso, a arquitetura modular peculiar de todas as PDIs é suscetível a produzir profundas diferenças na função da proteína mesmo com pequenas modificações de seus domínios específicos.

Em paralelo, a análise combinada de tais isoformas fornece um cenário relevante para entender as relações estrutura-função do PDIA1. Neste sentido, o conjunto de possibilidades funcionais evocadas pelas distintas arquiteturas de domínio de isoformas de *P4HB* incluem: 1) alteração nas funções redox tais como oxidase, redutase ou isomerase, relacionadas principalmente aos domínios *a*. A função isomerase de *P4HB* requer todos os 4 domínios, portanto é improvável que a maioria das isoformas de *splicing* alternativo de PDIA1 exiba a atividade de tiol isomerase, com possível exceção de *P4HB*-021; 2) mudança na função chaperona, que depende principalmente da presença e integridade dos domínios *b*; 3) alterações na especificidade do substrato e afinidade de ligação, também dependentes principalmente dos domínios *b*, mas também da conformação global da proteína; d) alteração na localização, dada pelo peptídeo sinal N-terminal e sequência C-terminal KDEL, além de outros motivos potenciais. Quando

todos esses aspectos são considerados em conjunto, é provável que a maioria das isoformas de P4HB não exerça atividade tiol isomerase no retículo endoplasmático, com a possível exceção de *P4HB-021*. Provavelmente, eles podem exercer outros tipos de atividade em localizações subcelulares distintas, expandindo significativamente o alcance funcional do gene da *P4HB*.

Nossos resultados ainda corroboram que o padrão de expressão das isoformas de *P4HB* é consistente com múltiplas funções específicas, uma vez que a expressão é distinta entre os diferentes tipos de células e tecidos. Embora tais níveis de expressão sejam geralmente baixos, deve-se considerar que as consequências funcionais podem ser relevantes se a alteração funcional ocorrer dentro de um compartimento específico. De fato, enquanto os níveis de expressão de PDIA1 são bastante altos no ER, os níveis de PDIA1 na superfície celular ou no meio extracelular são < 2% dos níveis totais de PDIA1 (Araújo *et al.*, 2017). Ainda assim, o pool extracelular tem importantes funções na regulação tiol-redox da adesão celular, trombose, remodelamento vascular e organização do citoesqueleto (Araújo *et al.*, 2017; Tanaka *et al.*, 2018).

Em particular, a variante *P4HB-021* foi significativamente expressa na VSMC, no entanto com um padrão variável conforme suas distintas origens e localizações. Tal variabilidade provavelmente reflete fatores como diversas origens embrionárias, distintas histórias mecanobiológicas e exposição variável a mediadores parácrinos de células endoteliais ou interações com matriz extracelular. A modelagem virtual de *P4HB-21* é consistente com uma proteína estruturada e eventualmente funcional, embora a mudança de arquitetura dos domínios *a* e *b* devam causar alterações funcionais importantes. Um estímulo particular que regulou positivamente a expressão de *P4HB-021* foi à privação de soro, sugerindo os possíveis efeitos dessa variante na sinalização metabólica e na diferenciação celular subjacente.

Nossos resultados indicam novos aspectos a serem considerados com relação aos efeitos mediados por PDIA1 em funções incluindo ativação de NOX NADPH oxidase, migração celular, diferenciação celular, mecanorregulação e modulação de RhoGTPase. Em todos esses casos, os processos mediados por PDIA1 foram avaliados principalmente no nível de expressão proteica, dada a alta expressão de proteínas PDI em geral. Nossos dados fornecem uma base para adicionar à compreensão de como as funções dependentes de PDIA1 podem ser potencialmente moduladas pela regulação gênica. Perspectivas futuras incluem a identificação de funções específicas da variante

de *splicing* *P4HB-021* e também a análise da expressão gênica em células musculares lisas de diferentes origens embrionárias.

Conclusão

Conclusão

- A disponibilidade de bancos de dados de RNA-seq contendo amostras de diferentes linhagens celulares e órgãos possibilitou a construção de perfil de expressão gênica de todas as variantes de *splicing* do gene *P4HB* em três bancos de dados distintos de RNA-seq
- Entre as 10 isoformas estudadas, duas delas, a *P4HB-02* e *P4HB-021*, apresentam características de potenciais transcritos reguladores ou produtos proteicos funcionais.
- Além dessas duas isoformas, focamos nossa análise na *P4HB-027* cujos dados de expressão se mostraram representativos. Nessas três variantes, realizamos uma análise quanto à presença de domínios da proteína PDI, número de éxons, número de aminoácidos, tamanho de nucleotídeos, aminoácidos, peso molecular (kDa).
- Conseguimos validar (amplificação, clonagem e sequenciamento) as junções de *splicing* específicas para a cada uma das três isoformas em tipos celulares identificados previamente por ferramentas de bioinformática e acesso aos bancos de dados.
- Foi possível detectar a expressão de três isoformas em diferentes linhagens celulares humanas revelando a isoforma *P4HB-021* como uma variante potencialmente relevante na VSMC.
- Em ensaios de qPCR de VSMC submetidas a carenciamento, tanto o gene *P4HB* quanto a *P4HB-021* mostraram-se reguladas positivamente.
- O panorama de transcritos alternativos da *P4HB* pode fornecer novos elementos para o entendimento da regulação gênica da expressão da PDIA1 e variantes associadas, bem como de suas implicações funcionais.

Perspectivas

Perspectivas

Uma das perspectivas é a realização de ensaios funcionais como o experimento de transfeção de células de músculo liso com vetor, pFN21a *HaloTag*® que contenham as variantes de *splicing* *P4HB*-02 e *P4HB*-021, clonadas e expressem as isoformas das proteínas. Primeiramente, com o intuito de verificar a sua sub-localização celular e posteriormente purificar as isoformas proteicas e realizar ensaios de sua atividade (atividade redutase, isomerase e chaperona). Esse vetor (pFN21a) possui o gene da *HaloTag*® na porção N-terminal. O vetor pFC14a contendo o gene da *P4HB* e também as variantes *P4HB*-02 e *P4HB*-021, já foram estudados e não foi possível detectar a presença da proteína HaloTag, porém nesse estudo a proteína está fusionada no C-terminal da proteína de interesse.

A segunda perspectiva é investigar em dados de Terceira Geração de Sequenciamento principalmente os dados da PacBio (Pacific Biosciences) ou MinION (Oxford Nanopore Technologies) que possuem *reads* longos. O enfoque nesses dados seria interessante, pois comprovaria por outras tecnologias a presença das variantes de *splicing* do gene *P4HB*.

Anexos

ANEXOS

ANEXO A – Tabela 1 – Identificação dos arquivos FANTOM5

Anexo A - Tabela 1		
Amostra	Tipo da amostra	Arquivo ID
Whole blood (ribopure), donor090309, donation1	primary cell	DRR013801
Whole blood (ribopure), donor090309, donation2	primary cell	DRR013802
Whole blood (ribopure), donor090309, donation3	primary cell	DRR013803
Whole blood (ribopure), donor090612, donation1	primary cell	DRR013804
Whole blood (ribopure), donor090612, donation2	primary cell	DRR013805
Whole blood (ribopure), donor090612, donation3	primary cell	DRR013806
Smooth Muscle Cells - Aortic, donor0 (nuclear fraction)	primary cell	DRR062865
CD19+ B Cells (pluriselect), donor090309, donation1	primary cell	DRR013789
CD19+ B Cells (pluriselect), donor090309, donation2	primary cell	DRR013790
CD19+ B Cells (pluriselect), donor090309, donation3	primary cell	DRR013791
CD19+ B Cells (pluriselect), donor090612, donation1	primary cell	DRR013792
CD19+ B Cells (pluriselect), donor090612, donation2	primary cell	DRR013793
CD19+ B Cells (pluriselect), donor090612, donation3	primary cell	DRR013794
Small Airway Epithelial Cells, donor3 (nuclear fraction)	primary cell	DRR062866
CD8+ T Cells (pluriselect), donor090309, donation1	primary cell	DRR013795
CD8+ T Cells (pluriselect), donor090309, donation2	primary cell	DRR013796
CD8+ T Cells (pluriselect), donor090309, donation3	primary cell	DRR013797
CD8+ T Cells (pluriselect), donor090612, donation1	primary cell	DRR013798
CD8+ T Cells (pluriselect), donor090612, donation2	primary cell	DRR013799
CD8+ T Cells (pluriselect), donor090612, donation3	primary cell	DRR013800
Preadipocyte - breast, donor2 (nuclear fraction)	primary cell	DRR062867
Mesothelial Cells, donor3	primary cell	DRR062868
CD14+ Monocytes, donor3	primary cell	DRR062869
Smooth Muscle Cells - Pulmonary Artery, donor3	primary cell	DRR062870
Hepatic Sinusoidal Endothelial Cells, donor3	primary cell	DRR062871
Universal RNA - Human Normal Tissues Biochain, pool1	tissue	DRR062872
testis, adult, pool1	tissue	DRR062873
thymus, fetal, pool1	tissue	DRR062874
heart, fetal, pool1	tissue	DRR013807
epididymis, adult	tissue	DRR062875
parotid gland, adult	tissue	DRR062876
extraskelatal myxoid chondrosarcoma cell line:H-EMC-SS	cell line	DRR062877
renal cell carcinoma cell line:OS-RC-2	cell line	DRR062878
colon carcinoma cell line:COLO-320	cell line	DRR062879
adult T-cell leukemia cell line:ATN-1	cell line	DRR062880
Burkitt's lymphoma cell line:DAUDI	cell line	DRR062881
choriocarcinoma cell line:BeWo	cell line	DRR062882
acute lymphoblastic leukemia (T-ALL) cell line:HPB-ALL	cell line	DRR062883
myeloma cell line:PCM6	cell line	DRR062884
retinoblastoma cell line:Y79	cell line	DRR062885
ductal cell carcinoma cell line:MIA Paca2	cell line	DRR062886
neuroblastoma cell line:CHP-134	cell line	DRR062887
small cell lung carcinoma cell line:LK-2	cell line	DRR062888

argyrophil small cell carcinoma cell line:TC-YIK	cell line	DRR062889
testicular germ cell embryonal carcinoma cell line:NEC15	cell line	DRR062890
neuroblastoma cell line:NBSusSR	cell line	DRR062891
small-cell gastrointestinal carcinoma cell line:ECC4	cell line	DRR062892
gastrointestinal carcinoma cell line:ECC12	cell line	DRR062893
mucinous adenocarcinoma cell line:JHOM-1	cell line	DRR062894
hairy cell leukemia cell line:Mo	cell line	DRR062895
leiomyoblastoma cell line:G-402	cell line	DRR062896
leukemia, chronic megakaryoblastic cell line:MEG-01	cell line	DRR062897
NK T cell leukemia cell line:KHYG-1	cell line	DRR062898
rhabdomyosarcoma cell line:KYM-1	cell line	DRR062899
anaplastic large cell lymphoma cell line:Ki-JK	cell line	DRR062900
adenocarcinoma cell line:IM95m	cell line	DRR062901
anaplastic squamous cell carcinoma cell line:RPMI 2650	cell line	DRR062902
Wilms' tumor cell line:G-401	cell line	DRR062903
adrenal cortex adenocarcinoma cell line:SW-13	cell line	DRR062904
medulloblastoma cell line:D283 Med	cell line	DRR062905
small cell lung carcinoma cell line:DMS 144	cell line	DRR062906
small cell lung carcinoma cell line:NCI-H82	cell line	DRR062907
merkel cell carcinoma cell line:MKL-1	cell line	DRR062908
Mast cell - stimulated, donor1	primary cell	DRR062909
Olfactory epithelial cells, donor1	primary cell	DRR062910
thyroid, fetal, donor1	tissue	DRR062911
substantia nigra, adult, donor10252	tissue	DRR062912
pineal gland, adult, donor10252	tissue	DRR062913
pituitary gland, adult, donor10252	tissue	DRR062914
cerebellum, adult, donor10252	tissue	DRR062915

ANEXO B –Tabela 2 – Números de acesso de arquivos ENCODE

Tabela 1 - GEO número de acesso	
Identificação da amostra	GEO ID
Homo sapiens smooth muscle of the pulmonary artery primary cell male adult (26 years) (2 biological replicates)	GSM2072330
Homo sapiens smooth muscle of the pulmonary artery primary cell male adult (28 years) (2 biological replicates)	GSM2072331
wgEncodeCaltechRnaSeqGm12878R2x75Il200AlignsRep2V2	GSM958728
wgEncodeCaltechRnaSeqHct116R2x75Il200AlignsRep1V2	GSM958749
wgEncodeCaltechRnaSeqHct116R2x75Il200AlignsRep2V2	GSM958749
wgEncodeCaltechRnaSeqHsmmR2x75Il200AlignsRep2V2	GSM958744
wgEncodeCshlLongRnaSeqHepg2CytosolPapAlnRep1	GSM758576
wgEncodeCshlLongRnaSeqHuvecCellPapAlnRep1	GSM758563
wgEncodeCshlLongRnaSeqHuvecCytosolPapAlnRep3	GSM758569
wgEncodeCshlLongRnaSeqHuvecCytosolPapAlnRep4	GSM758569
wgEncodeCshlLongRnaSeqHuvecNucleusLongnonpolyaAlnRep3	GSM767857
wgEncodeCshlLongRnaSeqHuvecNucleusPapAlnRep4	GSM758565
wgEncodeCshlLongRnaSeqCd20CellPapAlnRep1	GSM981256
wgEncodeCshlLongRnaSeqGm12878CellLongnonpolyaAlnRep2	GSM758572
wgEncodeCshlLongRnaSeqGm12878CytosolPapAlnRep1	GSM758560
wgEncodeCshlLongRnaSeqGm12878CytosolPapAlnRep2	GSM758560
wgEncodeCshlLongRnaSeqH1hescNucleusPapAlnRep2	GSM758574
wgEncodeCshlLongRnaSeqHaoaf609010111CellTotalAlnRep2	GSM984614
wgEncodeCshlLongRnaSeqHaoec70717061CellTotalAlnRep1	GSM984618
wgEncodeCshlLongRnaSeqHaoecCellTotalAlnRep2	GSM984618
wgEncodeCshlLongRnaSeqHepg2CellLongnonpolyaAlnRep1	GSM758567
wgEncodeCshlLongRnaSeqHepg2CellLongnonpolyaAlnRep2	GSM758567
wgEncodeCshlLongRnaSeqHepg2CytosolLongnonpolyaAlnRep2	GSM767840
wgEncodeCshlLongRnaSeqHepg2CytosolPapAlnRep1	GSM758576
wgEncodeCshlLongRnaSeqHmscuc00525017CellTotalAlnRep1	GSM984607
wgEncodeCshlLongRnaSeqHmscuc00811017CellTotalAlnRep2	GSM984607
wgEncodeCshlLongRnaSeqNhekCellLongnonpolyaAlnRep1	GSM765398
wgEncodeCshlLongRnaSeqSknshCellPapAlnRep3	GSM981253
wgEncodeCshlLongRnaSeqSknshCellPapAlnRep4	GSM981253

ANEXO C – Tabela 3 – Amostras GTEx

Tabela 3 - Amostras GTEx

Amostra	Número de amostras
Baço	162
Bexiga	11
Cérebro	1671
Colo uterino	11
Cólon	507
Coração	600
Esofago	1021
Estômago	262
Fígado	175
Glândula adrenal	190
Glândula pituitária	183
Glândula salivar	97
Intestino delgado	137
Mama	290
Músculo	564
Nervo	414
Ovário	133
Pâncreas	248
Pele	1203
Próstata	152
Pulmão	427
Rim	45
Sangue - linfócitos	130
Sangue total	407
Tecido adiposo	797
Testículo	259
Tireóide	446
Trompas de falópio	7
Útero	111
Vagina	115
Vasos sanguíneos	913

Referências

1. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010;11(10):R106.
2. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc.* 2013;8(9):1765-86.
3. Andreadis A. Tau splicing and the intricacies of dementia. *J Cell Physiol.* 2012;227(3):1220-5.
4. Araujo TLS, Zeidler JD, Oliveira PVS, Dias MH, Armelin HA, Laurindo FRM. Protein disulfide isomerase externalization in endothelial cells follows classical and unconventional routes. *Free Radic Biol Med.* 2017;103:199-208.
5. Asano K. Why is start codon selection so precise in eukaryotes? *Translation (Austin).* 2014;2(1):e28387.
6. Bang ML, Centner T, Fornoff F, Geach AJ, Gotthardt M, McNabb M, et al. The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ Res.* 2001;89(11):1065-72.
7. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. *Nature.* 2010;465(7294):53-9.
8. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics.* 2011;27(3):343-50.
9. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep.* 2017;7(1):10480.

10. Bienert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* 2017;45(D1):D313-d9.
11. Blakeley P, Siepen JA, Lawless C, Hubbard SJ. Investigating protein isoforms via proteomics: a feasibility study. *Proteomics.* 2010;10(6):1127-40.
12. Blencowe BJ. Alternative splicing: new insights from global analyses. *Cell.* 2006;126(1):37-47.
13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-20.
14. Bouayad D, Pederzoli-Ribeil M, Mocek J, Candalh C, Arlet JB, Hermine O, et al. Nuclear-to-cytoplasmic relocation of the proliferating cell nuclear antigen (PCNA) during differentiation involves a chromosome region maintenance 1 (CRM1)-dependent export and is a prerequisite for PCNA antiapoptotic activity in mature neutrophils. *J Biol Chem.* 2012;287(40):33812-25.
15. Camargo LL, Babelova A, Mieth A, Weigert A, Mooz J, Rajalingam K, et al. Endo-PDI is required for TNFalpha-induced angiogenesis. *Free Radic Biol Med.* 2013;65:1398-407.
16. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol.* 2009;10(11):741-54.
17. Chen WS, Xu PZ, Gottlob K, Chen ML, Sokol K, Shiyanova T, et al. Growth retardation and increased apoptosis in mice with homozygous disruption of the Akt1 gene. *Genes Dev.* 2001;15(17):2203-8.
18. Chen Z, Song J, Chen L, Zhu C, Cai H, Sun M, et al. Characterization of TTN Novex Splicing Variants across Species and the Role of RBM20 in Novex-Specific Exon Splicing. *Genes (Basel).* 2018;9(2).
19. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther.* 2012;22(4):271-4.

20. Coulombe-Huntington J, Lam KC, Dias C, Majewski J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.* 2009;5(12):e1000766.
21. Darby NJ, van Straaten M, Penka E, Vincentelli R, Kemmink J. Identifying and characterizing a second structural domain of protein disulfide isomerase. *FEBS Lett.* 1999;448(1):167-72.
22. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.* 2010;24(21):2343-64.
23. De Faria Poloni J, Bonatto D. Influence of transcriptional variants on metastasis. *RNA Biol.* 2018;15(8):1006-24.
24. Fernandes DC, Manoel AH, Wosniak J, Jr., Laurindo FR. Protein disulfide isomerase overexpression in vascular smooth muscle cells induces spontaneous preemptive NADPH oxidase activation and Nox1 mRNA expression: effects of nitrosothiol exposure. *Arch Biochem Biophys.* 2009;484(2):197-204.
25. Ferrari DM, Soling HD. The protein disulphide-isomerase family: unravelling a string of folds. *Biochem J.* 1999;339 (Pt 1):1-10.
26. Forrest AR. Deep sequencing blood transcriptomes. *Blood.* 2014;123(17):2595-6.
27. Freedman RB, Hirst TR, Tuite MF. Protein disulphide isomerase: building bridges in protein folding. *Trends Biochem Sci.* 1994;19(8):331-6.
28. Galligan JJ, Petersen DR. The human protein disulfide isomerase gene family. *Hum Genomics.* 2012;6:6.
29. Greaser ML, Krzesinski PR, Warren CM, Kirkpatrick B, Campbell KS, Moss RL. Developmental changes in rat cardiac titin/connectin: transitions in normal animals and in mutants with a delayed pattern of isoform transition. *J Muscle Res Cell Motil.* 2005;26(6-8):325-32.
30. Gross E, Sevier CS, Heldman N, Vitu E, Bentzur M, Kaiser CA, et al. Generating disulfides enzymatically: reaction products and electron acceptors of the endoplasmic reticulum thiol oxidase Ero1p. *Proc Natl Acad Sci U S A.* 2006;103(2):299-304.

31. Grosso AR, Gomes AQ, Barbosa-Morais NL, Caldeira S, Thorne NP, Grech G, et al. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res.* 2008;36(15):4823-32.
32. Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis.* 2009;30 Suppl 1:S162-73.
33. Hartley SW, Mullikin JC. Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. *Nucleic Acids Res.* 2016;44(15):e127.
34. Hatahet F, Ruddock LW. Protein disulfide isomerase: a critical evaluation of its function in disulfide bond formation. *Antioxid Redox Signal.* 2009;11(11):2807-50.
35. Inaba K, Masui S, Iida H, Vavassori S, Sitia R, Suzuki M. Crystal structures of human Ero1alpha reveal the mechanisms of regulated and targeted oxidation of PDI. *Embo j.* 2010;29(19):3330-43.
36. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, et al. The transcriptional program in the response of human fibroblasts to serum. *Science.* 1999;283(5398):83-7.
37. Jacquot JP, Gelhaye E, Rouhier N, Corbier C, Didierjean C, Aubry A. Thioredoxins and related proteins in photosynthetic organisms: molecular basis for thiol dependent regulation. *Biochem Pharmacol.* 2002;64(5-6):1065-9.
38. Jasuja R, Passam FH, Kennedy DR, Kim SH, van Hessem L, Lin L, et al. Protein disulfide isomerase inhibitors constitute a new class of antithrombotic agents. *J Clin Invest.* 2012;122(6):2104-13.
39. Jurica MS, Moore MJ. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell.* 2003;12(1):5-14.
40. Kawaji H, Kasukawa T, Fukuda S, Katayama S, Kai C, Kawai J, et al. CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res.* 2006;34(Database issue):D632-6.

41. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357-60.
42. Klappa P, Ruddock LW, Darby NJ, Freedman RB. The b' domain provides the principal peptide-binding site of protein disulfide isomerase but all domains contribute to binding of misfolded proteins. *Embo j*. 1998;17(4):927-35.
43. Kozlov G, Maattanen P, Schrag JD, Pollock S, Cygler M, Nagar B, et al. Crystal structure of the bb' domains of the protein disulfide isomerase ERp57. *Structure*. 2006;14(8):1331-9.
44. Kralovicova J, Vorechovsky I. Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Res*. 2007;35(19):6399-413.
45. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet*. 1992;90(1-2):41-54.
46. Lahmers S, Wu Y, Call DR, Labeit S, Granzier H. Developmental control of titin isoform expression and passive stiffness in fetal and neonatal myocardium. *Circ Res*. 2004;94(4):505-13.
47. Laurindo FR, Pescatore LA, Fernandes Dde C. Protein disulfide isomerase in redox cell signaling and homeostasis. *Free Radic Biol Med*. 2012;52(9):1954-69.
48. Lee G, Blenis J. Akt-ivation of RNA splicing. *Mol Cell*. 2014;53(4):519-20.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
50. Li J, Huang K, Hu G, Babarinde IA, Li Y, Dong X, et al. An alternative CTCF isoform antagonizes canonical CTCF occupancy and changes chromatin architecture to promote apoptosis. *Nat Commun*. 2019;10(1):1535.

51. Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, et al. Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.* 2019;47(D1):D752-d8.
52. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 2015;16:22.
53. Mascarenhas JB, Tchourbanov AY, Fan H, Danilov SM, Wang T, Garcia JG. Mechanical Stress and Single Nucleotide Variants Regulate Alternative Splicing of the MYLK Gene. *Am J Respir Cell Mol Biol.* 2017;56(1):29-37.
54. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol.* 2005;6(5):386-98.
55. Mattick JS, Makunin IV. Small regulatory RNAs in mammals. *Hum Mol Genet.* 2005;14 Spec No 1:R121-32.
56. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008;5(7):621-8.
57. Ni M, Zhou H, Wey S, Baumeister P, Lee AS. Regulation of PERK signaling and leukemic cell survival by a novel cytosolic isoform of the UPR regulator GRP78/BiP. *PLoS One.* 2009;4(8):e6868.
58. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature.* 2010;463(7280):457-63.
59. Ochoa-Leyva A, Montero-Moran G, Saab-Rincon G, Briebe LG, Soberon X. Alternative splice variants in TIM barrel proteins from human genome correlate with the structural and evolutionary modularity of this versatile protein fold. *PLoS One.* 2013;8(8):e70582.
60. Palaniswamy R, Teglund S, Lauth M, Zaphiropoulos PG, Shimokawa T. Genetic variations regulate alternative splicing in the 5' untranslated regions of the mouse glioma-associated oncogene 1, Gli1. *BMC Mol Biol.* 2010;11:32.
61. Pan JL, Bardwell JC. The origami of thioredoxin-like folds. *Protein Sci.* 2006;15(10):2217-27.

62. Pan S, Chen HH, Correia C, Dai H, Witt TA, Kleppe LS, et al. Cell surface protein disulfide isomerase regulates natriuretic peptide generation of cyclic guanosine monophosphate. *PLoS One*. 2014;9(11):e112986.
63. Pescatore LA, Bonatto D, Forti FL, Sadok A, Kovacic H, Laurindo FR. Protein disulfide isomerase is required for platelet-derived growth factor-induced vascular smooth muscle cell migration, Nox1 NADPH oxidase expression, and RhoGTPase activation. *J Biol Chem*. 2012;287(35):29290-300.
64. Pickering BM, Willis AE. The implications of structured 5' untranslated regions on translation and disease. *Semin Cell Dev Biol*. 2005;16(1):39-47.
65. Pihlajaniemi T, Helaakoski T, Tasanen K, Myllyla R, Huhtala ML, Koivu J, et al. Molecular cloning of the beta-subunit of human prolyl 4-hydroxylase. This subunit and protein disulphide isomerase are products of the same gene. *Embo j*. 1987;6(3):643-9.
66. Richards KL, Rowe ML, Hudson PB, Williamson RA, Howard MJ. Combined ligand-observe (19)F and protein-observe (15)N,(1)H-HSQC NMR suggests phenylalanine as the key Delta-somatostatin residue recognized by human protein disulfide isomerase. *Sci Rep*. 2016;6:19518.
67. Robert C, Watson M. The incredible complexity of RNA splicing. *Genome Biol*. 2016;17(1):265.
68. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 29. United States 2011. p. 24-6.
69. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, et al. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res*. 2010;38(Database issue):D620-5.
70. Sanidas I, Polytarchou C, Hatzia Apostolou M, Ezell SA, Kottakis F, Hu L, et al. Phosphoproteomics screen reveals akt isoform-specific signals linking RNA processing to lung cancer. *Mol Cell*. 2014;53(4):577-90.

71. Severin J, Lizio M, Harshbarger J, Kawaji H, Daub CO, Hayashizaki Y, et al. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat Biotechnol.* 2014;32(12):217-9.
72. Sevier CS, Kaiser CA. Ero1 and redox homeostasis in the endoplasmic reticulum. *Biochim Biophys Acta.* 2008;1783(4):549-56.
73. Soares Moretti AI, Martins Laurindo FR. Protein disulfide isomerases: Redox connections in and out of the endoplasmic reticulum. *Arch Biochem Biophys.* 2017;617:106-19.
74. Surget S, Khoury MP, Bourdon JC. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *Onco Targets Ther.* 2013;7:57-68.
75. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene.* 2016;35(19):2413-27.
76. Takahashi H, Lassmann T, Murata M, Carninci P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc.* 2012;7(3):542-61.
77. Tanaka LY, Araujo TLS, Rodriguez AI, Ferraz MS, Pelegati VB, Morais MCC, et al. Peri/epicellular protein disulfide isomerase-A1 acts as an upstream organizer of cytoskeletal mechanoadaptation in vascular smooth muscle cells. *Am J Physiol Heart Circ Physiol.* 2019;316(3):H566-h79.
78. Tanaka S, Uehara T, Nomura Y. Up-regulation of protein-disulfide isomerase in response to hypoxia/brain ischemia and its protective effect against apoptotic cell death. *J Biol Chem.* 2000;275(14):10388-93.
79. Tasanen K, Parkkonen T, Chow LT, Kivirikko KI, Pihlajaniemi T. Characterization of the human gene for a polypeptide that acts both as the beta subunit of prolyl 4-hydroxylase and as protein disulfide isomerase. *J Biol Chem.* 1988;263(31):16218-24.

80. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 2013;14(2):178-92.
81. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* 2013;31(1):46-53.
82. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012;7(3):562-78.
83. Turano C, Coppari S, Altieri F, Ferraro A. Proteins of the PDI family: unpredicted non-ER locations and functions. *J Cell Physiol.* 2002;193(2):154-63.
84. Vadolas J, Nefedov M, Wardan H, Mansooriderakshan S, Voullaire L, Jamsai D, et al. Humanized beta-thalassemia mouse model containing the common IVSI-110 splicing mutation. *J Biol Chem.* 2006;281(11):7399-405.
85. Vandenbroucke, II, Vandesompele J, Paepe AD, Messiaen L. Quantification of splice variants using real-time PCR. *Nucleic Acids Res.* 2001;29(13):E68-8.
86. Vaquero AR, Ferreira NE, Omae SV, Rodrigues MV, Teixeira SK, Krieger JE, et al. Using gene-network landscape to dissect genotype effects of TCF7L2 genetic variant on diabetes and cardiovascular risk. *Physiol Genomics.* 2012;44(19):903-14.
87. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science.* 2001;291(5507):1304-51.
88. Vitting-Seerup K, Sandelin A. The Landscape of Isoform Switches in Human Cancers. *Mol Cancer Res.* 2017;15(9):1206-20.
89. Wahl MC, Will CL, Luhrmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell.* 2009;136(4):701-18.
90. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* 2008;456(7221):470-6.

91. Wang J, Gao QS, Wang Y, Lafyatis R, Stamm S, Andreadis A. Tau exon 10, whose missplicing causes frontotemporal dementia, is regulated by an intricate interplay of cis elements and trans factors. *J Neurochem.* 2004;88(5):1078-90.
92. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10(1):57-63.
93. Ward AJ, Cooper TA. The pathobiology of splicing. *J Pathol.* 2010;220(2):152-63.
94. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 2018;46(W1):W296-w303.
95. Wessagowit V, Kim SC, Woong Oh S, McGrath JA. Genotype-phenotype correlation in recessive dystrophic epidermolysis bullosa: when missense doesn't make sense. *J Invest Dermatol.* 124. United States 2005. p. 863-6.
96. Wu Y, Ahmad SS, Zhou J, Wang L, Cully MP, Essex DW. The disulfide isomerase ERp57 mediates platelet aggregation, hemostasis, and thrombosis. *Blood.* 2012;119(7):1737-46.
97. Yu CK, Xu T, Assoian RK, Rader DJ. Mining the Stiffness-Sensitive Transcriptome in Human Vascular Smooth Muscle Cells Identifies Long Noncoding RNA Stiffness Regulators. *Arterioscler Thromb Vasc Biol.* 2018;38(1):164-73.
98. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. *Nucleic Acids Res.* 2018;46(D1):D754-d61.