Letícia Machado Favery Bertoline

Análise *in silico* de variantes não sinônimas para priorizar o estudo de genes candidatos para frequência cardíaca

Orientador: Prof. Dr. José Eduardo Krieger

São Paulo

2022

Letícia Machado Favery Bertoline

Análise *in silico* de variantes não sinônimas para priorizar o estudo de genes candidatos para frequência cardíaca

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

Programa de Ciências Médicas

Área de concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo

Orientador: Prof. Dr. Jose Eduardo Krieger

São Paulo

2022

Dados Internacionais de Catalogação na Publicação (CIP)

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

Bertoline, Letícia Machado Favery
Análise in silico de variantes não sinônimas para priorizar o estudo de genes candidatos para frequência cardíaca / Letícia Machado Favery Bertoline. -- São Paulo, 2022.
Dissertação(mestrado)--Faculdade de Medicina da Universidade de São Paulo.
Programa de Ciências Médicas. Área de Concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo.
Orientador: Jose Eduardo Krieger.

Descritores: 1.Frequência cardíaca 2.Doenças cardiovasculares 3.Predição de estrutura proteica 4.Simulação de dinâmica molecular 5.Priorização de genes candidatos 6.Dano proteico

USP/FM/DBD-499/22

Responsável: Erinalva da Conceição Batista, CRB-8 6755

AGRADECIMENTOS

Agradeço aos meus pais Cláudia e David, por me escolherem, priorizarem, apoiarem e incentivarem ao longo de toda minha caminhada nos dias ensolarados e nos tempestuosos.

Agradeço aos meus avós Clélia e Antônio Carlos, por acreditarem em mim e estarem ao meu lado neste caminho.

Agradeço a minha avó Inês que nos deixou durante a execução deste trabalho e me ensinou o quão necessário é lutar, resignar-se e ser resiliente, além de sempre rezar por mim e torcer pelas minhas conquistas.

Agradeço aos meus amigos por me ouvirem, apoiarem e incentivarem nesta trajetória.

Agradeço especialmente à Flávia, por sua amizade que durante este período foi fundamental, me ouvindo, auxiliando e animando.

Agradeço ao Professor Dr. Krieger, por me receber no laboratório e me proporcionar realizar esta pesquisa em seu grupo sob sua orientação.

Agradeço à Dra. Samantha Teixeira pelas orientações, discussões e colaborações no desenvolvimento e execução deste trabalho.

Agradeço à Angélica pela colaboração e ensinamento na parte de Dinâmica Molecular bem como conselhos e dicas.

Agradeço à Anna Laura pela colaboração e ensinamento na parte de desenvolvimento de hipóteses bem como conselhos e dicas.

Agradeço à Mariana e Mariliza por me ensinarem as metodologias iniciais deste trabalho.

Agradeço à equipe administrativa do LGCM (Renata, Ana Piesco, Silvana, Ana Reis, Ana Mota, André Ribeiro, André Bueno) por todo auxílio, suporte discussão, e troca.

Agradeço aos colegas do LGCM por todas as conversas e contribuições.

Agradeço a todos que compuseram a rede de apoio ao meu redor que possibilitou a execução deste trabalho.

Agradeço por fim à CAPES pela bolsa ao longo deste período.

EPÍGRAFE

“Seja forte e corajoso”

Josué 1:9

RESUMO

Bertoline LMF. Análise in silico de variantes não sinônimas para priorizar o estudo de genes candidatos para frequência cardíaca [dissertação]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2022.

A frequência cardíaca é um fator de risco independente para doenças cardiovasculares que são responsáveis por 30% das mortes no país e no mundo. A herdabilidade da frequência cardíaca de repouso (RHR) é significativa, estimando-se a sua contribuição em 25% da variação do fenótipo. Estudos de associação ampla do genoma identificaram 64 loci associados a RHR, que juntos explicam apenas 2,5% da variação total do fenótipo. Portanto, há duas importantes lacunas de conhecimento a serem superadas: identificar genes candidatos que influenciam a regulação da frequência cardíaca e desenvolver estratégias eficientes para priorização daqueles que serão validados. Nosso grupo tem contribuído para a primeira identificando sistematicamente genes envolvidos com a frequência cardíaca por meio da estratégia de Forward Genetic Screening, onde os genes de camundongos são sistematicamente modificados para identificar aqueles que influenciam o RHR. Avaliamos nos últimos 5 anos cerca de 23% do genoma do camundongo e identificamos 129 candidatos associados à RHR. Somente 17 destes candidatos são conhecidos por influenciar o fenótipo, 47 mostraram evidências diretas de associação com o fenótipo, enquanto 65 estão em loci contendo mais de um gene candidato. Neste trabalho, propomos uma estratégia de priorização com base em abordagens computacionais *in sílico.* Avaliamos o efeito das alterações em genes/proteínas por meio da 1. conservação da sequência de nucleotídeos, aminoácidos e domínio proteico em três espécies (*Mus musculus*, *Homo sapiens* e *Danio rerio*), usando informações dos bancos de dados públicos Ensembl e Uniprot; e 2. na predição de danos na estrutura proteicas obtidas experimentalmente (proveniente do banco PDBe) ou por predição (provenientes do banco de dados do AlphaFold2 e do algoritmo Phyre[2]) e da ferramenta Missense3d. Finalmente, considerando proteínas selecionadas por ambas metodologias, utilizamos Simulação de Dinâmica Molecular (SDM) para acessar o impacto das trocas de aminoácidos em estruturas terciárias das proteínas julgadas afetadas que trariam danos funcionais. Em conjunto, a abordagem proposta permitiu priorizar 5 genes candidatos. Considerando os genes com associação direta com o fenótipo, selecionamos o Candidato 35, onde a alteração genética afeta aminoácido conservado na região do domínio e é predito de alterar a estrutura da proteína nas três espécies estudadas; e o Candidato 46 selecionado por ambas abordagens na espécie humana e em camundongo, já que este não apresenta ortólogo em no *Danio rerio*. A SDM das mutações em ambos candidatos demonstrou modificações conformacionais nas proteínas, sugerindo alterações em suas funções. Além disso, identificamos os candidatos 91, 99 e 106 dentre os genes em loci contendo mais de um candidato pelas duas abordagens em pelo menos duas espécies estudadas. As análises computacionais aplicadas criaram as bases para priorizar a caracterização de genes candidatos com base na predição de alterações com potencial de comprometer a função proteica conservada em pelo menos duas espécies diferentes. Se validada, esta abordagem será fundamental para aumentar a eficiência da validação *in vivo* dos genes candidatos para RHR.

Palavras-chave: Frequência cardíaca. Doenças cardiovasculares. Predição de estrutura proteica. Simulação de dinâmica molecular. Priorização de genes candidatos. Dano proteico.

ABSTRACT

Bertoline LMF. In silico analysis of non-synonymous variants to prioritize the study of candidate genes for heart rate [dissertation]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2022.

Heart rate is an independent risk factor for cardiovascular diseases that are responsible for 30% of deaths in the country and in the world. The heritability of resting heart rate (RHR) is significant, estimating its contribution to 25% of the phenotype variation. Genome-wide association studies have identified 64 RHR-associated loci, which together explain only 2.5% of the total phenotype variation. Therefore, there are two important knowledge gaps to be overcome: identifying candidate genes that influence heart rate regulation and developing efficient strategies for prioritizing those that will be validated. Our group has contributed to the first by identifying systematically genes involved in heart rate through the Forward Genetic Screening strategy, where mouse genes are systematically modified to identify those that influence the RHR. In the last 5 years, we evaluated about 23% of the mouse genome and identified 129 candidates associated with RHR. Only 17 of these candidates are known to influence the phenotype, 47 showed direct evidence of association with the phenotype, while 65 are at loci containing more than one candidate gene. In this work, we propose a prioritization strategy based on *in silico* computational approaches. We evaluated the effect of gene/protein alterations through 1. conservation of nucleotide, amino acid and protein domain sequence in three species (Mus musculus, Homo sapiens and Danio rerio), using information from public databases Ensembl and Uniprot; and 2. in the prediction of damage to protein structure obtained experimentally (from the PDBe database) or by prediction (from the AlphaFold2 database and the Phyre² algorithm) and the Missense3d tool. Finally, considering proteins selected by both methodologies, we used Molecular Dynamics Simulation (MDS) to assess the impact of amino acid changes on tertiary structures of proteins judged affected that would bring functional damage. Together, the proposed approach allowed prioritizing 5 candidate genes. Considering the genes with direct association with the phenotype, we selected Candidate 35, where the genetic alteration affects conserved amino acid in the domain region and is predicted to alter the structure of the protein in the three studied species; and Candidate 46 selected by both approaches in humans and in mice, since it does not have an ortholog in in Danio rerio. The MDS of the mutations in both candidates showed conformational changes in the proteins, suggesting alterations in their functions. Furthermore, we identified candidates 91, 99 and 106 among genes at loci containing more than one candidate by both approaches in at least two species studied. The applied computational analyzes created the bases to prioritize the characterization of candidate genes based on the prediction of changes with the potential to compromise the conserved protein function in at least two different species. If validated, this approach will be essential to increase the efficiency of *in vivo* validation of candidate genes for RHR.

Keywords: Heart rate. Cardiovascular diseases. Protein structure prediction. Molecular dynamics simulation. Priorização de genes candidatos. Protein damage.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1. Doenças cardiovasculares e seu impacto no mundo

As doenças cardiovasculares (DCVs) são um conjunto de doenças cardíacas e vasculares como: doença coronariana; doença cerebrovascular; doença arterial periférica; doença cardíaca reumática; cardiomiopatia congênita, trombose venosa profunda; embolia pulmonar; entre outras. Segundo a Organização Mundial da Saúde (OMS), as DCVs são a principal causa de óbitos do mundo. Em 2019, as mortes causadas por estas patologias contabilizaram 17,9 milhões, representando 32% do total global.

Em termos nacionais, DCV é responsável por cerca de 33% dos óbitos anuais, afetando principalmente a camada populacional em maior estado de vulnerabilidade. Segundo o site do cardiômetro, até o dia 24/11/2022, houveram 362451 mortes por doenças cardiovasculares no Brasil.

1.2. Relação da frequência cardíaca em doenças cardiovasculares

A frequência cardíaca (FC), quantidade de batidas que o coração realiza por minuto, é um fenótipo complexo e modificável, sendo modulado por fatores ambientais e genéticos. Estudos demonstraram uma forte associação entre alta frequência cardíaca e morbidade e mortalidade por doenças cardiovasculares, incluindo morte súbita e todas as causas de morte[1,2]. Além disso, estudos reportam a FC com um preditor independente para mortalidade após analises multivariadas corrigidas por variáveis demográficas e clínicas[3].

Jouven e colaboradores demonstraram não só que o risco de morte súbita por infarto do miocárdio é aumentada em indivíduos com frequência cardíaca de repouso (FCR) acima de 75 bpm, mas também em indivíduos cuja FC aumenta menos que 89 bpm durante o exercício e que diminui menos do que 25 bpm após a finalização do exercício, concluindo que o perfil da FC durante e após o exercício são preditores de morte súbita[4]. Além disso, alterações no sistema de condução cardíaco observado durante o eletrocardiograma podem levar a fibrilação atrial e arritmias ventriculares.

1.3. A herdabilidade da frequência cardíaca

Devido a influência da FC no desenvolvimento de DCVs é de extrema importância compreender os fatores que influenciam a modulação deste fenótipo. Estudos mostram que a hereditariedade desempenha um papel importante na variação interindividual da frequência cardíaca e que explica aproximadamente a 25% da variação do fenótipo[56]. Estudos em gêmeos adultos indicam uma herdabilidade ainda maior: Russel e colaboradores identificaram uma herdabilidade de 77% para o intervalo R-R e de 36% para intervalo QT, ambas grandezas associadas à FC[7], e Dalageorgou e colaboradores identificaram uma herdabilidade de 56% para frequência cardíaca[8].

Estudos de associação ampla do genoma com polimorfismos de um único nucleotídeo (SNPs) (do inglês Genome-wide Association Studies – GWAS) tem como objetivo estabelecer a relação entre variantes genéticas comuns e alterações singelas em fenótipos complexos, como a FC. Diversos estudos foram realizados em diferentes populações, identificando dezenas de variantes genéticas em genes ou próximas deles associados a esses fenótipos. Dentre estes, o estudo de Eijgelsheim e colaboradores demonstrou associações entre variantes genéticas comuns nos genes MYH6, GJA1 e CD34 ou próximos deles e FC[9].

Um estudo realizado por Eppiga et al avaliou 265.046 indivíduos e 19,9 milhões de variantes genéticas, e identificou 64 loci associados com FCR. Levando em conta o montante do número de alelos presentes nestes loci que conferem alteração na FCR, foi desenvolvido um escore de risco genético para cada indivíduo. Na sequência, fora observado uma associação significativa entre variantes genéticas associadas à FCR e todas as causa de morte, sendo que o aumento de 5 bpm na FCR provoca um aumento relativo de 20% no risco de todas as causas de morte avaliadas. No entanto, cada uma dessas 64 regiões identificadas apresenta efeito no fenótipo muito pequeno, variando de 0,2 a 1,1 bpm, sendo que coletivamente essas variantes explicam apenas 2,5% da variação total da FC de repouso [10].

Sendo assim, uma vez que os estudos executados até então não justificam a totalidade da herdabilidade do fenótipo de FC, faltando explicação da origem de aproximadamente 23,5% da herdabilidade, e considerando a correlação entre variantes genéticas e o aumento do risco de morte cardiovascular, é de extrema importância e interesse acadêmico a execução de estudos para a identificação de novos genes envolvidos no controle da FC. Tal feito conduzirá à identificação de novos alvos terapêuticos para tratamento e prevenção de eventos cardiovasculares.

1.4. Identificação de genes associados à frequência cardíaca

Um método que se destacou no campo da identificação e caracterização de genes associados à fenovariância em modelos experimentais foi a metodologia de *Forward Genetic Screening*. Nessa abordagem, animais submetidos a agentes mutagênicos são selecionados pelo seu fenótipo sem conhecimento prévio da base genética que o ocasiona. A técnica é aplicada em camundongos alvo de mutagênese pelo reagente N-etil-N-nitrosouréia (ENU), cujo grupo etil se transfere para o nitrogênio ou oxigênio das bases nitrogenadas, conferindo aleatoriedade à mutação. No entanto, esse grupo etil transferido por si só não constitui uma mutação, mas sua presença ocasiona um erro de identidade na base que o recebeu durante a replicação do DNA, resultando na falta de pareamento entre esta e sua base complementar. Após duas rodadas de replicação, uma substituição de uma única base acontece e isto não é identificado pelo sistema de reparo celular. Dessa forma, nessa estratégia, camundongos machos submetidos ao ENU (G0) são induzidos a mutações pontuais aleatórias nos gametas e sua progênie é submetida a um screening cuidadoso e compreensivo para identificar alterações na variância de fenótipos.[11,12].

O método apontado apresentava restrições no processo de identificar as variantes resultadas pelo ENU e sua correlação com o fenótipo, uma vez que necessitava usar marcadores polimórficos e realizar um mapeamento fino da região candidata, assim como o sequenciamento éxon a éxon. Com o surgimento do sequenciamento de próxima geração (NGS), o sequenciamento total do genoma ou do exoma passou a ser realizado em semanas, bem como

a execução de suas análises, possibilitando identificar as variantes genéticas geradas pelo agente mutagênico de modo rápido e eficiente [12].

Com o objetivo de identificar as mutações resultantes do ENU por NGS bem como sua correspondência com os fenótipos, muitas estratégias foram desenvolvidas e aplicadas. Uma delas, é a desenvolvida por Wang e colaboradores que detecta as variantes geradas pelo ENU associadas à fenótipos de interesse em tempo real [13]. Nela, as mutações são descobertas mediante ao sequenciamento total do exoma do progenitor G1 e a sua zigosidade é estabelecida em camundongos G2 e G3 via genotipagem, aplicando um painel desenhado para o Ion PGM TM System for Next-Generation Sequencing, anteriormente à análise de fenótipo. No final do mapeamento dos animais, as características quantitativas e qualitativas, incluindo efeitos letais, são analisados de maneira integrada com os dados genéticos utilizando um software, Linkage Analyzer, em famílias únicas (pedigrees únicos) ou em superfamílias (super pedigrees). O software, desenvolvido no laboratório do Prof. Bruce Beutler, identifica a ligação significativa entre mutações individuais e escores de fenótipos aberrantes e o dado processado é apresentado em Manhattan plots, como mostrado na Figura 1-A e C. Como o efeito de mutações em um gene pode ser sutil em um background normal, é feita a análise levando em conta superfamílias (análise de superpedigrees) com o objetivo de avaliar o efeito de mutações distintas em um mesmo gene sobre o fenótipo ao decorrer da progressão do mapeamento.

Essa metodologia possibilita realizar imparcialmente o mapeamento genético dos fenótipos, englobando aqueles com penetrância incompleta, a identificação automática de mutações causais concomitantemente com o mapeamento fenotípico e a exclusão de genes não associados ao fenótipo de interesse.

Durante os últimos 5 anos, por meio da colaboração com o pesquisador Bruce Beutler, avaliamos 23% do total de genes autossômicos do genoma do camundongo nos fenótipos de pressão arterial sistólica e frequência cardíaca em 841 pedigrees contendo 43.627 animais G3. A partir da avaliação desses animais, identificamos alterações genéticas em 129 genes associados com

alterações na frequência cardíaca. Destes, 17 apresentaram evidências publicadas de associação com o controle do fenótipo, sendo considerados controles positivos, 47 mostraram evidências diretas de associação com o fenótipo, enquanto 65 estão em loci contendo mais de um gene candidato.

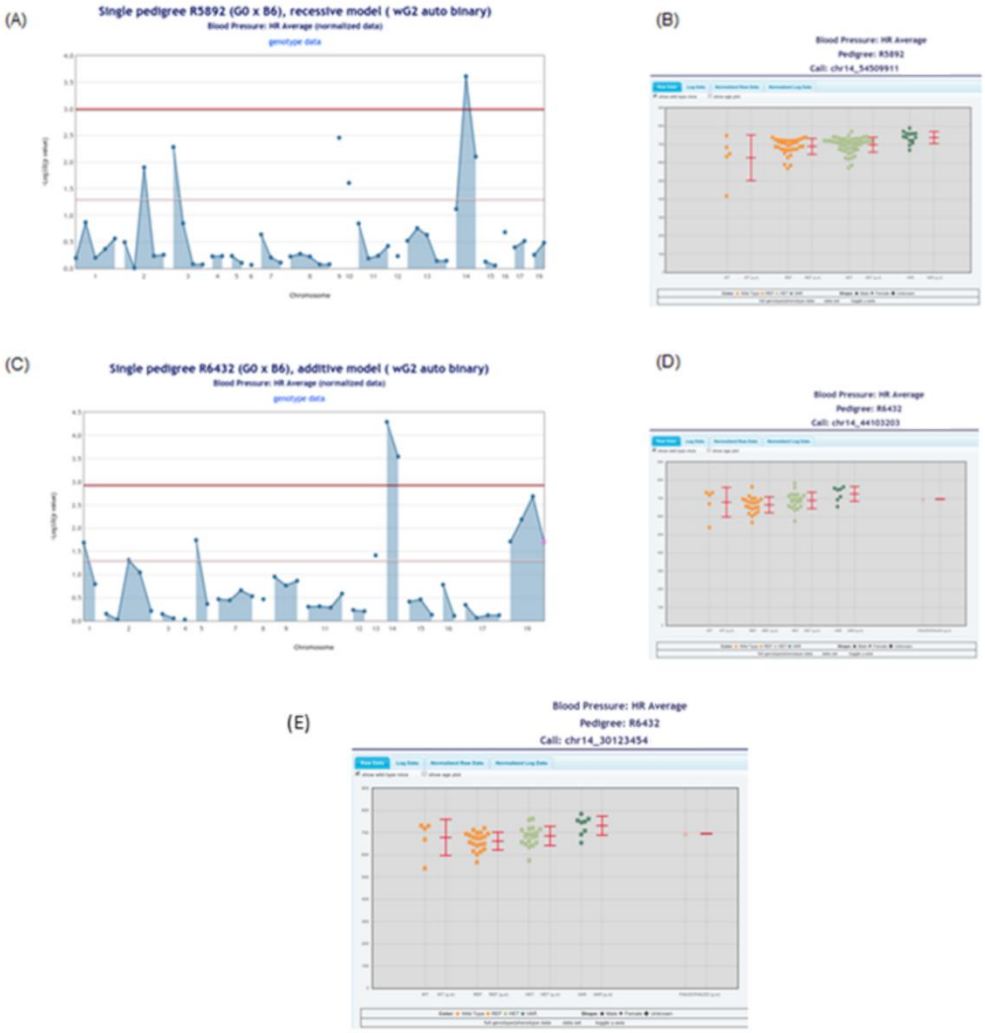

Figura 1 - (A) Mahattan plot do pedigree R5892, mostrando que a mutação no gene candidato foi significativamente associada ao aumento da FC. No eixo x é apresentado os cromossomos e, no eixo y, o $-\log_{10}$(p-valor) do teste de associação de cada uma das mutações identificadas no pedigree e alterações no fenótipo de FC. A linha vermelha determina o p-valor para o teste de associação considerado significativo após correção de múltiplos testes. (B) Gráfico de FC média dos diferentes genótipos para a mutação no gene candidato no pedigree R5892. No gráfico são representados da esquerda para a direita dados dos controles C57 (WT – em laranja), homozigoto selvagem (REF – em laranja), heterozigoto (HET – em verde claro), e homozigoto mutado (VAR – em verde escuro). (C) Mahattan plot do pedigree R6432, no qual esta representado o fato de que duas mutações nos genes candidato e Cacna1d estão associadas com o fenótipo de estudo. (D) Gráfico de FC média dos diferentes genótipos para a mutação no gene candidato no pedigree R6432. (E) Gráfico de FC média dos diferentes genótipos para a mutação no gene Cacna1d no pedigree R6432.

A Figura 1 apresenta tanto o Manhattan plot, gráfico no qual é apresentado o resultado do teste de associação de todas as mutações detectadas na família e a modificações no fenótipo de FC, quanto o gráfico da FC média dos diferentes genótipos para uma alteração genética no gene que apresentou resultado significativo no teste de associação. A Figura 1A e B apresentam o resultado de um gene diretamente associado ao fenótipo, já que apenas uma mutação no pedigree R5892 apresentou significância estatística no teste de associação (Manhattan plot, ponto acima da linha vermelha), enquanto Figura 1 -CC-E apresenta o resultado de um pedigree (R6432) onde mutações em dois genes foram significativamente associadas ao aumento da FC (demonstrado pelos dois pontos acima da linha vermelha no cromossomo 14 do Manhattan plot, assim como pela semelhança dos gráficos da FC média dos diferentes genótipos para mutação nos dois genes, sugerindo que os mesmos animais com mutação em homozigose em um gene apresentam também mutação para o outro gene).

1.5. Necessidade de priorizar genes candidatos para estudo

Em um estudo de levantamento de genes candidatos, costuma-se identificar diversos candidatos e para se avaliar cada um deles, usualmente são desenvolvidos animais knockouts. O valor de custo de um camundongo knockout realizado por um laboratório como “The Jackson Laboratory” é de $4082 (aproximadamente R$ 20588,38). Para um estudo são necessários um casal e a partir dele, uma linhagem será estabelecida e mantida por um biotério para então uma série de experimentos serem realizados de acordo com o fenótipo de interesse. Todo este processo despende tempo e dinheiro, e lidamos com a probabilidade de o gene não apresentar associação com o fenótipo durante o estudo. Visto este quadro e considerando que o número de genes candidatos identificados como associados ao fenótipo de FC no screening fenotípico descrito anteriormente foi bastante amplo (129) surge a necessidade de se desenvolver um método de priorização de genes a serem posteriormente estudados.

1.6. A importância da ortologia e dos padrões de conservação proteico

A homologia é uma palavra de origem grega, que, no contexto de biologia evolutiva, significa o estudo de similaridade entre partes com mesma origem evolutiva. Dentro do conceito de homologia, há três tipos: ortólogos, parálogos e xenólogos. Considera-se um gene parálogo aquele que durante a evolução apresenta em uma das novas espécies uma duplicação do mesmo, mas função distinta. O termo xenólogo refere-se à transferência horizontal de genes entre as espécies, sendo que, dependendo da localidade do gene, a função pode ser preservada ou alterada. A ortologia, em âmbito de biologia molecular, aplica-se quando um gene está presente na espécie ancestral e em todas que provém dele. Este tipo de conservação genética mantém majoritariamente a função do gene nas diversas espécies em que se faz presente. O fato de haver conservação de função biológica, atribui a ortologia uma importância em estudos genômicos, uma vez que a partir de uma espécie é possível predizer a função do gene em outras espécies [14].

Ao comparar dois genomas de espécies distintas, adentramos no campo da genômica comparativa. Nele é possível verificar sua importância em diversos nichos, como, na identificação de novos genes baseados na conservação nucleotídica em diferentes espécies e na identificação de motivos regulatórios[15].

O estudo de sequências de nucleotídeos com baixa taxa de alteração entre espécies ao longo do tempo, consideradas altamente conservadas, correspondem a regiões de elevada importância biológica para o organismo, como sítios associados a funções de replicação de material genético e regiões de interação proteína-proteína. Dito isto, estas regiões atuam como ponto de partida para estudos que buscam localizar a origem de distúrbios genéticos e alvos farmacológicos [16].

Regiões de domínios proteicos são aquelas onde há enovelamento independente, com a finalidade de constituir uma estrutura estável e compacta. Os domínios podem estar associados a uma função biológica ou a estabilidade estrutural da proteína [17]. Um estudo realizado por Ortiz e Sergeev mostram que

se espera que mutações em regiões de domínios levem a efeitos desestabilizantes nessas localidades em aspecto estrutural e funcional. [18]

1.7. A necessidade de predizer estruturas proteicas e suas metodologias

O estudo de estruturas proteicas é de grande importância para um melhor entendimento da sua função biológica. Ele permite uma melhor definição de como proteínas-ligantes e proteínas-proteínas interagem, a importância e função de regiões e domínios proteicos, assim como permite a utilização desse conhecimento para o desenvolvimento de novos fármacos e um melhor entendimento como alterações genéticas afetam o funcionamento de proteínas.

A necessidade de predizer estruturas proteicas emerge da diferença numérica exorbitante, de superior a mil vezes, entre sequências de aminoácidos conhecidos depositadas no UniProtKB e as estruturas tridimensionais obtidas experimentalmente depositadas no Protein Data Bank in Europe (PDBe). Sendo assim, muitos esforços tem sido realizados para possibilitar a obtenção destas estruturas, necessárias para estudos acerca de sua funcionalidade. Esforços estes que remontam da década de sessenta, quando emergiu o “Problema de dobramento proteico”.

Esta incógnita da ciência questiona como a partir de uma dada sequência de aminoácidos podemos determinar a estrutura tridimensional da proteína analisada. Este problema pode ser repartido em três subproblemas: código de dobramento, predição de estrutura proteica, e velocidade de dobramento[19]. O primeiro é um problema de cunho termodinâmico sobre como as forças interatômicas atuantes na sequência geram as estruturas nativas[19]. O segundo se refere a como obter computacionalmente uma estrutura nativa partindo apenas de uma dada sequência de aminoácidos[19]. O terceiro questiona como é possível que uma proteína se enovele tão rapidamente[19].

O subproblema da predição de estrutura proteica tem sido vastamente estudado por grupos de pesquisa ao redor do mundo, que apresentam ferramentas para solucionar em partes esta questão. Os métodos utilizados para obter tal estrutura podem ser segregados em duas classes: aqueles que

se baseiam em um template (*Template based modeling – TBM*), e aqueles que se baseiam em modelagem livre (*Free modeling – FM*)[20]. Para a primeira classe, existem dois métodos: modelagem por homologia e reconhecimento de enovelamento. Já para a segunda classe há apenas o método de aproximação *ab-initio*[20,21].

O método de modelagem por homologia parte do fato de que, se duas sequências de aminoácidos apresentam alta identidade, elas apresentam estruturas semelhantes[20–22]. Dessa forma, utiliza-se um template definido por uma sequência com alta identidade em relação ao alvo, e modela motivos e cadeias laterais, minimiza energia, refina e por fim valida com o diagrama de Ramachandran[20–22].

O método de reconhecimento de enovelamento constrói um template utilizando ferramentas computacionais avançadas de comparação de sequências[20,21,23]. Na sequência, realiza-se um estudo de similaridade de sequências de estruturas secundárias previstas, seguido por uma verificação de afinidade de sequências com dobras tridimensionais e verificação manual[20,21,23].

O método de aproximação *ab-initio* tem suas bases teóricas na hipótese termodinâmica de que estruturas de proteínas nativas apresentam energia livre baixíssimas[20,21,24,25]. Este método utiliza de conceitos físico-químicos tais como ligação de hidrogênio, potencial de contato, propensões de estruturas secundárias derivadas de estruturas PDB no processo de enovelamento para gerar um modelo[20,21,24,25].

Em 2019, DeepMind, uma start-up adquirida pela Google, publicou um novo algoritmo que prediz estruturas proteicas chamado AlphaFold. Este utiliza uma rede neural convolucional que tem como input a sequência proteica e uma série de padrões de alinhamento múltiplo de sequências, que gera o potencial de superfície específico da proteína, seguido por um segundo passo, que consiste em múltiplas corridas de gradiente descendente para identificar a estrutura que melhor minimiza a função do potencial da proteína. A utilização da rede neural convolucional permite que o problema de se encontrar a estrutura 3D de proteínas seja interpretado como um problema de visão

computacional. Para tanto, ele utiliza distogramas, histogramas que mostram as distâncias entre resíduos numa dada proteína, e compõe os perfis extraídos de proteínas com sequencias similares, utilizando alinhamento múltiplo de sequencias para gerar os padrões de perfis. Para o treino da rede, os desenvolvedores utilizaram estruturas experimentalmente adquiridas disponibilizas no Protein Data Bank (PDB) [26].

Este novo método chamou a atenção da comunidade científica após apresentar o melhor desempenho da história entre as metodologias de predição de proteína até então publicadas, na 13ª edição de uma das mais conhecidas avaliações de ferramentas de predição de estrutura proteica, a CASP (*Critical Assessment of Protein Structure Prediction*). Nesta competição, o AlphaFold atingiu a maior pontuação em todas as três categorias, incluindo aquela que avalia ferramentas de modelagem livre (TBM, FM e TBM+FM), ou seja, mesmo sem o treinamento da rede sem o múltiplo alinhamento de sequencias, o AlphaFold teve melhor desempenho que as ferramentas da área [26].

Na CASP14, a nova versão do AlphaFold, AlphaFold2, novamente recebeu reconhecimento, apresentando estruturas proteicas preditas com maior acurácia. Os autores atribuem a alta performance do algoritmo à incorporação de arquiteturas de redes neurais novas e procedimentos de treinamento baseados em restrições evolutivas, físicas e geométricas de estruturas proteicas [27]. Neste novo algoritmo, eles incluíram um transformer iterativo, uma arquitetura auto encoder especializada no mapeamento sequência a sequência, que emprega um mecanismo de atenção que permite que a rede aprenda correlações no dado de input [27].

No mesmo ano de lançamento do AlphaFold2, a DeepMind, em parceria com o Instituto Europeu de Bioinformática (*European Bioinformatics Institute from European Molecular Biology Laboratory – EMBL-EBI*) criaram um banco de dados que disponibiliza estruturas proteicas preditas pelo AlphaFold2 (AlphaFold Protein structure Datatbase – AlphaFold DB)(https://alphafold.ebi.ac.uk)[28]. No seu primeiro lançamento, foram disponibilizadas 360.000 estruturas preditas de mais de 21 organismos.

Atualmente, o banco já contém mais de 200 milhões de estruturas preditas de proteínas humanas e de mais 47 organismos.

Atualmente, novos métodos de predição de estrutura proteica vem sendo desenvolvidos, que utilizam o modelo de linguagem de proteína que se baseia em padrões de correlação entre combinações de amino ácidos e estrutura conformacional[29,30]. Esses novos métodos têm como base os modelos de processamento de linguagem natural (*Natural Language Processing – NLP*), onde as palavras corresponderiam aos aminoácidos e as sentenças às proteínas[30]. Destes, dois exemplos são o ESMfold e o EMBER2 reportados neste ano [29,30]. Esses métodos apresentam acurácia inferior ao AlphaFold mas velocidade muito superior uma vez que não usa múltiplo alinhamento de sequências[29,30]. É importante salientar que apesar dessas metodologias não apresentarem mesma acurácia na predição da estrutura proteica, elas parecem ter melhor desempenho na predição de mutações que acarretam alterações na estrutura proteica, sendo que o AlphaFold2 parece não reconhecer diferenças sutis na estrutura devido a modificação de apenas um aminoácido[31].

1.8. Utilização de simulação de dinâmica molecular para estudo de impacto de variante na interação entre proteína e ligante/proteína

A Simulação de Dinâmica molecular (SDM) é uma ferramenta utilizada para descrever o movimento (posição e velocidade) de cada átomo de uma dada molécula, tendo como base as leis newtonianas[32]. Historicamente fora desacreditada por biólogos e químicos, denominada como perda de tempo. Desde a primeira simulação de dinâmica molecular com proteína realizada em 1997 muitos avanços ocorreram dado a evolução dos algoritmos de simulação e das funções potenciais[32]. A técnica é comumente utilizada em determinação estrutural, refinamento estrutural de macromoléculas e obtenção de estruturas significativas de complexos macromoleculares a partir de dados e baixa resolução [32]. Outra empregabilidade da DM é o estudo de efeito de mutações em proteína de modo a entender seu mecanismo e características.

Um exemplo desta utilidade é o estudo e SDM do sítio de mutação pontual no domínio TPR (do inglês tetratricopeptide repeat) de cyclophilin 40 para identificar os estados conformacionais com diferentes dinâmicas e

propriedades enzimáticas[33]. Outro exemplo é o estudo de Piao e colaboradores, onde foram analisadas as relações cruciais entre a Shank3 e SAPAP E-PBM e explicadas as perdas de afinidade de ligação nos mutantes quando comparados ao tipo selvagem [34].

Tendo em vista a utilização de diferentes métodos e algoritmos computacionais para um melhor entendimento biológico da função proteica e como estas podem ser moduladas ou alteradas por alterações genéticas, no presente projeto, buscamos avaliar se métodos computacionais podem ser utilizados para priorizar o estudo de genes/proteínas candidatas associadas com a frequência cardíaca em modelos experimentais e na espécie humana.

2. OBJETIVO

2.1. Objetivo geral

A dissertação aqui apresentada tem como objetivo desenvolver análises *in silico* para priorizar a validação de genes candidatos associados a frequência cardíaca identificados por meio de *forward genetic screening* em camundongos.

2.2. Objetivos específicos

1. Priorizar genes candidatos identificados no *screening* fenotípicos para frequência cardíaca em camundongos com base na conservação de aminoácido e da presença de domínios proteicos nas regiões afetadas pelas alterações não sinônimas e conservadas em três espécies de vertebrados: *Mus musculus*, *Danio rerio* e *Homo sapiens*.
2. Selecionar genes candidatos com variantes não sinônimas que potencialmente causam danos à estrutura da proteína nas três espécies.
3. Selecionar para validação genes candidatos com base nas duas estratégias propostas acima.
4. Utilizar análise de dinâmica molecular para selecionar as alterações com maior potencial de influenciar comprometimento da função nas proteínas priorizadas.

3. METOLOLOGIA

Devido a novidade dos dados aqui apresentados e como muitos dos genes candidatos identificados por meio de *forward genetic screening* em camundongos nunca foram associados ao fenótipo de frequência cardíaca, estes serão apresentados decodificados. Dessa forma, para os genes em camundongo a nomenclatura será Candidato X, para zebrafish será candidato X e para humano, CANDIDATO X, seguindo-se a norma vigente dos símbolos de genes para cada espécie. Quando tratarmos das proteínas, a nomenclatura seguirá a do gene, mas com fonte em itálico.

Para os genes já conhecidos, com evidência sólida da sua contribuição no fenótipo, seu símbolo será apresentado segundo o banco de dados Ensembl (https://www.ensembl.org/index.html).

Destacamos que as análises foram realizadas considerando apenas variantes genéticas não sinônimas identificadas em genes candidatos, sendo excluídos do estudo variantes que causam *stopgain*, *stoploss* ou que afetam sítios de *splicing*. Em um primeiro momento, as análises forma realizadas considerando as alterações genéticas em genes conhecidos, posteriormente com os genes diretamente associados com o fenótipo estudado, seguindo com aqueles cuja relação com o fenótipo foi dúbia no camundongo, ou seja, quando mutações em dois ou mais genes ocorreram em homozigose nos mesmos animais da mesma família (pedigree), impossibilitando a identificação de qual mutação em qual gene foi responsável pela alteração do fenótipo.

3.1. Análise de padrões conservativos

A fim de avaliar se a variante genética associada à alteração de frequência cardíaca identificada em camundongo por meio de *screening* fenotípico pode ser encontrada em regiões conservadas nas espécies *Mus musculus*, *Danio rerio* e *Homo sapiens*, e possivelmente afetar a função proteica por se localizar em domínios proteicos, seguimos as seguintes etapas.

Primeiramente, verificamos se o gene candidato associado ao controle do fenótipo de interesse possui ortólogo para as espécies *Danio rerio* e *Homo sapiens* com base no banco de dados Ensembl

(https://www.ensembl.org/index.html). Para isto, após buscar o gene em camundongo (GRCm38), acessamos as informações referentes aos ortólogos (“Orthologues”) na aba “Comparitive Genomics, visualizando a existência do mesmo em *Danio Rerio* (GRCz10) e *Homo sapiens* (CRCh38), como mostrado na Figura 2[35–37].

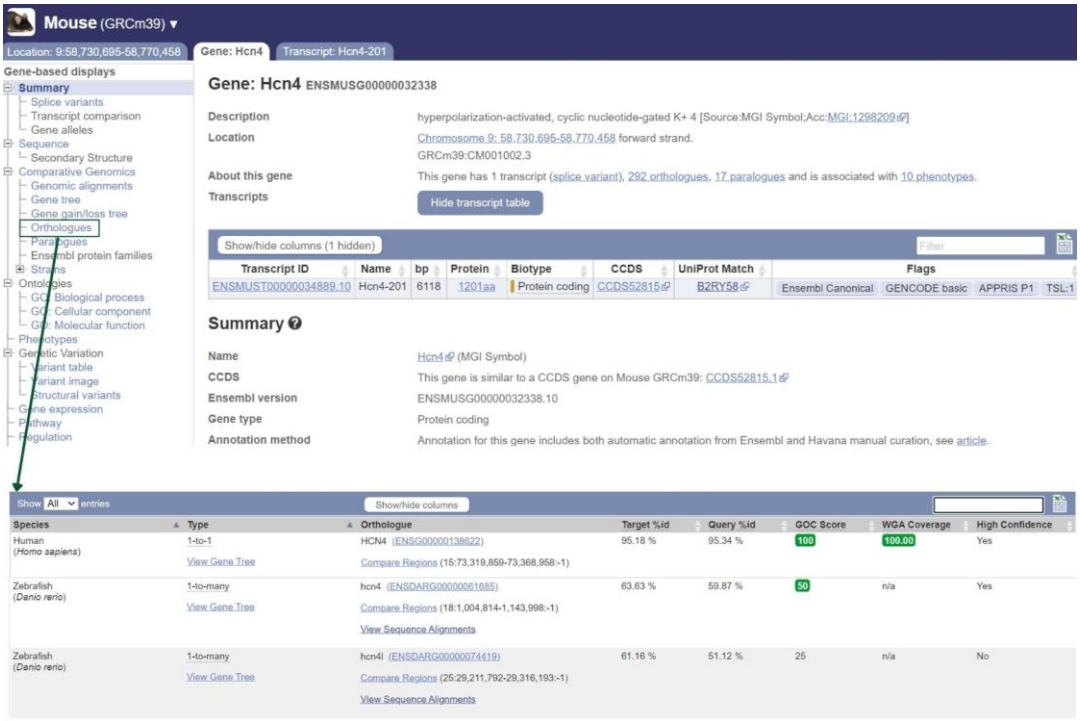

Figura 2 - Interface Ensembl para o gene Hcn4. A página fornece primeiramente informações a respeito do gene seguido pela tabela com o/os transcrito/s. Na parte esquerda da página alguns recursos estão disponíveis para obter mais informações a respeito do gene de interesse, entre eles o item “Orthologues”. Com a ferramenta disponível neste item, é possível verificar se há genes ortólogos em diferentes espécies, inclusive em *Danio rerio* e em *Homo Spiens*, utilizadas neste estudo.

Em caso de apresentar ortólogo, realizamos o alinhamento entre as sequências de cDNAs do camundongo e das outras espécies envolvidas no estudo, identificando o códon, o aminoácido afetado pela variante e sua posição na proteína. Para tanto, utilizamos o recurso “View cDNA” disponibilizado pela própria plataforma Ensembl na aba “View Sequence Alignments”. Assim, foi possível identificar o códon equivalente no cDNA da espécie humana e de peixe paulistinha, assim como o aminoácido e sua posição na proteína da respectiva espécie. Nesta etapa, foram anotados se houve alinhamento do códon onde a mutação inicial se localiza com relação à

espécie comparada e, em caso positivo, se o aminoácido originado pelo códon se conserva entre as espécies analisadas, como ilustrado na Figura 3[35–37].

Figura 3 - Alinhamento entre o cDNA do camundongo e o de humano para o gene Hcn4. A imagem mostra o alinhamento das sequências de DNA complementar do RNA mensageiro do transcrito consenso das espécies *Mus musculus* (primeira linha) e *Homo sapiens* (segunda linha). Em vermelho está destacado a trinca onde a variante genética associada ao fenótipo foi identificada em camundongo. Em seguida, vemos a sequência de DNA complementar do gene na espécie humana e a sequência de aminoácido (aa) equivalente. Em vermelho temos a trinca e o aminoácido equivalentes a variante identificada em camundongo.

Por fim, com o objetivo de inferir se a variante genética causa efeito na funcionalidade proteica nas três espécies estudadas, verificamos se a posição do aminoácido afetado por ela em cada espécie se encontra em sítios de domínio, região, “coiled coil”, “compositional bias” e repetição, no tópico “Family and Domain” do banco de dados Uniprot (acessado entre 2019 e 2021)(https://www.uniprot.org/)[38,39]. Para essa análise, consideramos domínio como abrangendo os conceitos de domínio, região, “coiled coil”, “compositional bias” e repetição.

Levando em consideração a análise de conservação entre as espécies de interesse e a posição das variantes genéticas nos sítios proteicos com evidência funcional (domínios), identificamos os genes cujos ortólogos nas espécies humana e peixe paulistinha apresentaram alinhamento de cDNA com relação ao camundongo na trinca que sofre alteração causada pela variante genética em camundongo e cujo aminoácido codificado pelo códon nas espécies estudadas fosse idêntico. Complementariamente, selecionamos dentre estes genes aqueles que apresentaram a variante em região de domínio em duas ou três espécies.

3.2. Análise de predição de dano estrutural

Com a finalidade de prever o impacto que a variante genética não sinônima missense causa na estrutura da proteína nas espécies *Homo sapiens*, *Danio rerio* e *Mus musculus* utilizamos a plataforma Missense3d. A ferramenta utiliza a estrutura proteica em formato .pdb, a posição do aminoácido em que ocorre a variante, o aminoácido nativo e o aminoácido gerado pela variante para realizar a análise. A fim de obter as estruturas proteicas codificadas pelos genes candidatos, a priori utilizamos o banco de dados “Protein Data Bank in Europe” (PDBe), onde estão disponibilizadas estruturas proteicas obtidas experimentalmente, seja por Cristalografia de raio X, Crio microscopia eletrônica ou Ressonância Magnética Nuclear. Caso a proteína analisada não apresentasse estrutura disponível neste banco ou a estrutura não cumprisse os pré-requisitos necessários, utilizamos as ferramentas AlphaFold (Release 3) e Phyre², para predizer a estrutura terciária das proteínas codificadas pelos genes candidatos, como mostrado na Figura 4.

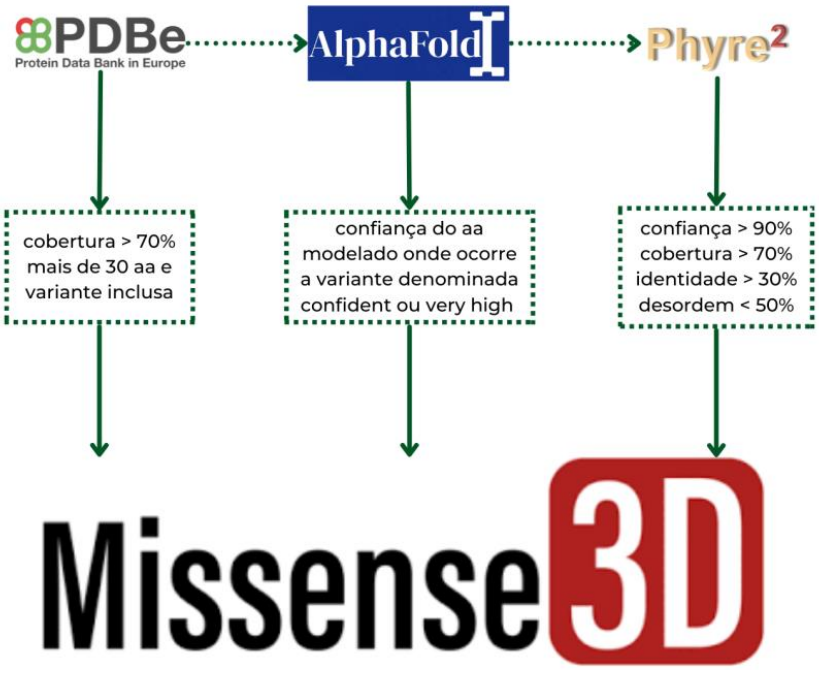

Figura 4 - Fluxograma metodológico de análise de predição de dano estrutural - A análise consistiu de duas etapas sendo que na primeira foram buscadas as estruturas das proteínas codificadas pelos genes candidatos no banco de dados PDBe e, em sua ausência, foram consultadas no banco de dados AlphaFold (https://alphafold.ebi.ac.uk/) e Phyre²(http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). As estruturas foram analisadas seguindo os parâmetros descritos na figura para cada fonte. Por fim, as

estruturas proteicas selecionadas foram submetidas ao Missense3d (http://missense3d.bc.ic.ac.uk/missense3d/).

Para o uso da ferramenta AlphaFold buscamos, no repositório disponibilizado pela software (https://alphafold.ebi.ac.uk/), a estrutura predita para cada sequência de aminoácidos associada ao código Uniprot do transcrito estudado [40]. Cada estrutura foi avaliada tendo como base a confiabilidade da predição do aminoácido determinada pelo fator pLDD (*per-residue confidence score*). Este fator pLDDT vai de 0 à 100, sendo que algumas das regiões com valor abaixo de 50 correspondem a regiões sem predição de estrutura (Figura 5). O modelo só foi utilizado quando o aa afetado pela mutação não-sinônima no camundongo apresentava confiabilidade superior à 70, ou seja, quando era considerado *Confident* ou *Very High*. Em caso de a estrutura não atingir a qualidade desejada, a sequência polipeptídica foi submetida ao Phyre².

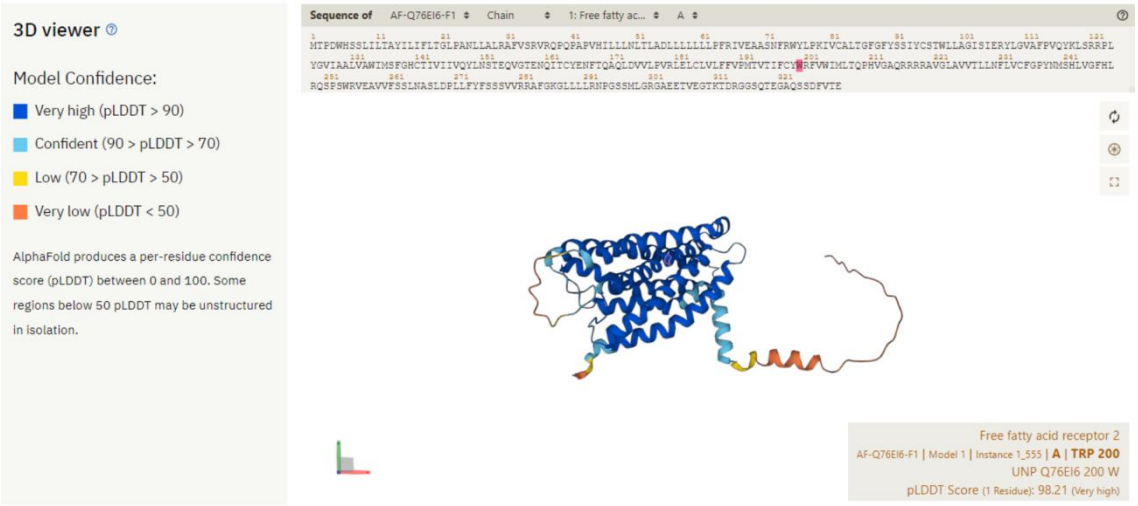

Figura 5 – Representação do resultado da predição de estrutura proteica pelo AlphaFold. Na parte superior é apresentada a sequência de aminoácidos da proteína e abaixo a estrutura predita, onde as cores representam a qualidade da predição. É possível selecionar na sequência de aminoácidos apresentada o aminoácido para visualização de sua posição na estrutura predita. Selecionamos estruturas proteicas que o aminoácido mutado estava em regiões preditas representadas em azul escuro (*Very high*) ou azul claro (*Confident*), como apresentado na legenda a esquerda.

Para o uso da ferramenta Phyre², submetemos a sequência de aminoácidos correspondente a cada transcrito ao modo Normal, conforme recomendado pelos desenvolvedores. Em casos em que essa análise não foi satisfatória para predição da estrutura proteica, as sequências de aminoácidos referentes a cada transcrito foram submetidas à análise em modo intensivo. Este modo utiliza um algoritmo que considera múltiplos templates para

predição da estrutura proteica. Entretanto, o modo intensivo só pode ser utilizado em sequências com comprimento inferior a 1500 resíduos. Nas situações em que a sequência de aminoácidos apresentou comprimento superior a 1500 resíduos, a análise foi interrompida, inviabilizando a predição da estrutura proteica e a avaliação do dano da alteração genética sobre a estrutura via Missense3d.

Para ambos os modos de análise para predição das estruturas proteicas, buscamos estruturas que apresentassem cobertura maior do que 70%, confiança na predição maior do que 90% e identidade maior do que 30%. Esses critérios foram determinados com base nas publicações das ferramentas Phyre² e Missense3d. Caso essas condições não fossem satisfeitas, a estrutura predita foi desconsiderada para as análises subsequentes no Missense3d e submetida ao PhyreAlarm na ferramenta Phyre², que avisa, via e-mail, melhora na predição da proteína investigada com base na publicação de novos templates [41,42].

As estruturas .pdb obtidas por meio do PDBe, AlphaFold e do Phyre² que cumpriram os valores definidos anteriormente foram submetidas ao Missense3d, acompanhadas das demais informações necessárias citadas anteriormente. A plataforma, então, retorna uma figura mostrando a estrutura nativa e a gerada a partir da sequência com aminoácido mutado, salientando com diferentes cores as alterações estruturais decorrentes da alteração do aminoácido, como mostrado na Figura 6. Além disso, como resultado, é apresentado também um parecer explicitando qual o tipo de dano causado e seus detalhes, como é possível ver na Figura 7.

Os possíveis danos detectados pela ferramenta podem ser separados em dois grupos: o primeiro depende de alterações físico-químicas acarretadas pela variante na estrutura selvagem e o segundo depende da qualidade do modelo juntamente com as alterações anteriores[42]. O primeiro grupo é composto por: quebra de ponte dissulfeto, introdução de prolina internalizada, introdução de resíduo hidrofílico internalizado, introdução de resíduo carregado internalizado, troca de resíduo com carga oposta internalizado, ângulo phi/psi não permitido, substituição de resíduo carregado para não carregado

interanlizado, substituição de glicina internalizada, substituição de uma prolina cis e, finalmente, substituição de uma glicina em uma curva (“Disulfide bond breakage”, “Buried Pro introduced”, “Buried hydrophilic introduced”, “Buried charge introduced”, “Buried charge switch”, “Disallowed phi/psi”, “Buried charge replaced”, “Buried Gly replaced”, “Cis Pro replaced” e “Gly in a bend”). O segundo grupo é formado pelas seguintes características: quebra, estrutura secundária alterada, quebra de ponte de hidrogênio internalizada, alteração de cavidade, e troca de exposição/internalização (“Clash”, “Second structure altered”, “Buried H-bond breakage”, “Cavity altered” e “Buried/exposed switch”).

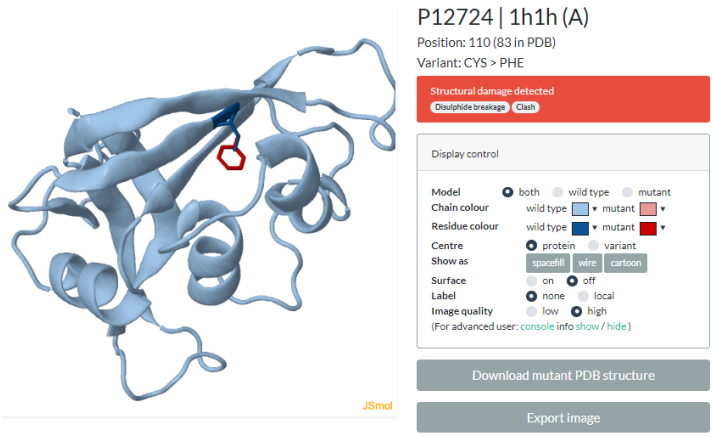

Figura 6 - Análise realizada pelo Missense3d para a estrutura de uma proteína na espécie humana. A janela retornada mostra a detecção de dano estrutural, sendo eles: “disulphidebreakage” e “clash”. Na estrutura, é possível ver em azul claro a cadeia da proteína selvagem, e em azul escuro o resíduo no qual ocorre a variante. Ainda na estrutura, em vermelho é possível ver o resíduo resultante da alteração.

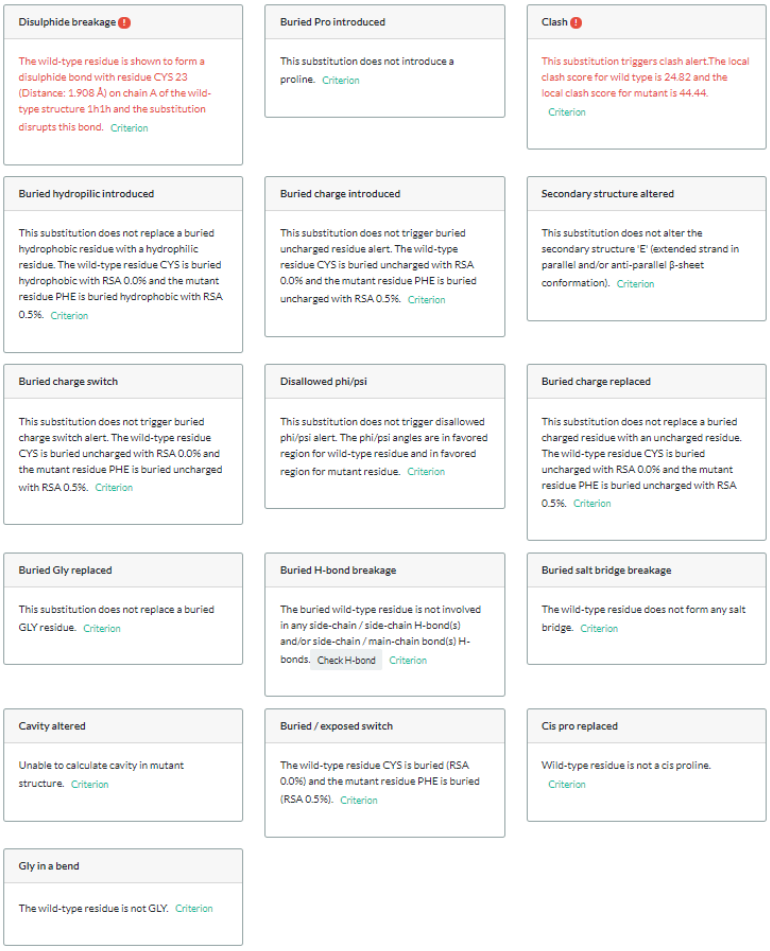

Figura 7 - Descrição das características das estruturas analisadas pelo Missense3d. Em vermelho, estão destacadas as características que apresentam dano à estrutura e sua descrição.

3.3. Integração dos resultados de análise de conservação e de predição de danos estruturais

Ao término das análises nos aspectos conservativos e estruturais, as listas com os genes com maiores quantidades de evidências obtidas por cada metodologia foram comparadas buscando por aqueles presentes em ambas ou presente parcialmente em ambas. Como, por exemplo, cumprir os requisitos conservativos em humanos e camundongos, mas não em peixe paulistinha e possuir dano estrutural em apenas uma das espécies abordadas no trabalho.

3.4. Estudo das estruturas proteicas codificadas pelos genes priorizados

Para selecionar qual estrutura proteica seria utilizada para as etapas posteriores, foram levadas em consideração a qualidade da estrutura e

principalmente a confiança sobre a interação domínio-domínio, proteína-proteína ou proteína-ligante. A fim de avaliar a qualidade da estrutura proteica foi utilizado a plataforma PDBsum [43]. Nela, são obtidas informações como esquemas de estrutura secundária, topologia, lista de ligações (pontes de hidrogênio, ligação dissulfeto, etc), diagrama de Ramachandran, entre outros. No Diagrama de Ramachandran especificamente, foram avaliados se os aminoácidos se encontraram em regiões permitidas. Inicialmente, buscamos estrutura experimentalmente obtida com o ligante. Caso não fosse possível, realizamos o *docking* utilizando a estrutura proteica e o ligante.

3.5. Simulação de dinâmica molecular

Para avaliar o impacto da variante não sinônima *missense* na função da proteína *Candidato 35* e *Candidato 46,* foi realizada uma simulação de dinâmica molecular (SDM). As simulações foram realizadas utilizando o programa GROMACS versão 2018.3, e o campo de forças aplicado foi o CHARMM36[44–46].

3.5.1. *Candidato 46*

Inicialmente, para a proteína *Candidato 46* foram separadas quatro estruturas para a proteína oriunda de camundongo: domínio apo obtido experimentalmente (código PDB: 3BI9), domínio com ligante obtido experimentalmente (código PDB: 3BIB), domínio predito utilizando o código aberto da ferramenta AlphaFold, e a estrutura da proteína inteira predita pelo AlphaFold (AF-Q6U7R4-F1) e disponível em seu banco de dados. Uma primeira simulação foi executada durante 5ns a fim de minimizar energias ao longo do tempo.

Para realização do preparo dos dados para a simulação de dinâmica molecular, foi utilizada a ferramenta CHARMM-GUI[47–49]. No primeiro passo de preparação dos dados, o arquivo em formato .pdb é lido pela ferramenta e então alguns átomos/aminoácidos tem sua nomenclatura alterada. Um exemplo é a troca de HIS, representando a histidina, por HSD (não protonado) ou HSP (protonado nos dois hidrogênios) ou HSE (protonado em um dos hidrogênios) dependendo da protonação da histidina. A caixa na qual a dinâmica molecular ocorreu era retangular e com distância da borda de 10 Å. Na sequência, foram

incluídos íons de sódio e cloro na solução por meio do método de Monte-Carlo com uma concentração de 0,15mM (37 de Na+ e 44 de Cl-).

As etapas seguintes foram realizadas no HPC Saci. Para cada uma das etapas das simulações foram utilizados 1 nó com um total de 4 núcleos, 4 núcleos logicos, 1 placa de GPU NVIDIA Tesla V100-SXM2-16GB. Com relação a memória, o total de NUMA foi de 86 GB. NUMA (Non-uniform memory acess) é a memória do computador usado em multiprocessamento, onde o tempo de acesso a memória é dependente da localização da memória em relação ao processador. Para se ter uma ideia do tempo de processamento para uma proteína com XX resíduos de aminoácidos, foi possível realizar 47,542 ns por dia.

A etapa de minimização da energia do sistema consistiu de uma dinâmica de 5000 passos por meio de integração por *step descent*. A etapa seguinte consistiu da realização de uma SDM curta de 125ps, sendo 125000 passos com duração de 0,001ps cada, com objetivo de equilibrar a energia do sistema. A dinâmica de produção foi realizada com 1000ps e 500000 passos com duração de 0,002ps cada para cada ns de simulação.

A partir dessa SDM, selecionamos o modelo mais indicado para trabalhar levando em consideração a confiança na estrutura proteica e de sua interação com o ligante. Na sequência, com a estrutura selecionada obtida no último nanosegundo da SDM de 5ns, modelamos a estrutura com a variante, utilizando o programa Pymol, para então gerar os inputs da simulação com arquivos do CHARMM-GUI contendo as condições de contorno e solução. Com esses arquivos foi executado uma SDM de 100ns para o sistema nativo e o com a variante.

Com os arquivos da SDM gerados, foram confeccionados dois vídeos, um para o domínio nativa e outro para o domínio que apresenta a variante M30K. Além disso, confeccionamos gráficos de distância entre os resíduos e o ligante para os dois domínios. Para os vídeos, foi utilizado o programa VMD no qual ambos os modelos foram sobrepostos. Para os gráficos, inicialmente foi utilizado o webserver ProteinPlus, especificamente a ferramenta PoseView, para mostrar quais resíduos do domínio, em ambas estruturas WT e M30K,

interagem com o ligante [50–52]. Para esses resíduos, juntamente com a variante de estudo, foram feitos gráficos de distância entre o centro de massa do resíduo e o centro de massa do ligante. Todos os gráficos foram feitos pelo programa Xmgrace (versão 2014).

3.5.2. *Candidato 35*

Utilizamos a mesma metodologia apresentada no candidato 46 para a simulação de dinâmica para o *Candidato 35*. No entanto, consideramos apenas a estrutura proteica predita para *Mus musculus* pelo AlphaFold e realizamos uma SDM com 30ns ao invés de 100ns, já que, pelo tamanho desta proteína (756 aa), uma dinâmica de 100ns demandaria mais de três meses de processamento utilizando o poder computacional apresentado.

Para este candidato, realizamos o processo de *docking* entre a estrutura proteica e o complexo tubulina (código 1TUB no PBDe), para representar a interação com o microtúbulo, utilizando o *webserver* ClusPro[53–56]. Foi realizado o *docking* na configuração default nos instantes inicial (t=0) e final (t=30) para os modelos selvagem (*wildtype*) e mutado (R638H) da SDM e escolhido o cluster resultante com maior número amostral. Por fim, a análise visual foi realizada com o Pymol para verificar o impacto da variante na interação proteína-proteína.

4. RESULTADOS E DISCUSSÃO

4.1. Genes conhecidos

Com base na análise fenotípica realizada em cerca de 45 mil camundongos, identificamos 17 genes candidatos associados à alteração de frequência cardíaca com evidências sólidas da sua participação na regulação do fenótipo, e por isso considerados controle positivo para a estratégia proposta. Este grupo, denominado genes conhecidos, compreende os genes: Ryr2, Hcn1, Kcnq3, Tg, Hcn4, Gad2, Lepr, Tbx5, Cacna1d, Slc8a1, Tln1, Cbs, Kcnj5, Ngfr, Sin3a, Kcnj3 e Duox2.

4.1.1. Análise de padrões conservativos

Dentre os 17 genes conhecidos que apresentam variantes não sinônimas, nove genes apresentam a variante em sítio de domínio proteico em camundongo, sendo eles: Ryr2, Tg, Lepr, Cacna1d, Tln1, Ngfr, Sin3a, Kcnj3 e Duox2 (Figura 8). Enquanto nos oito genes restantes as mutações não se situam em região de domínio proteico.

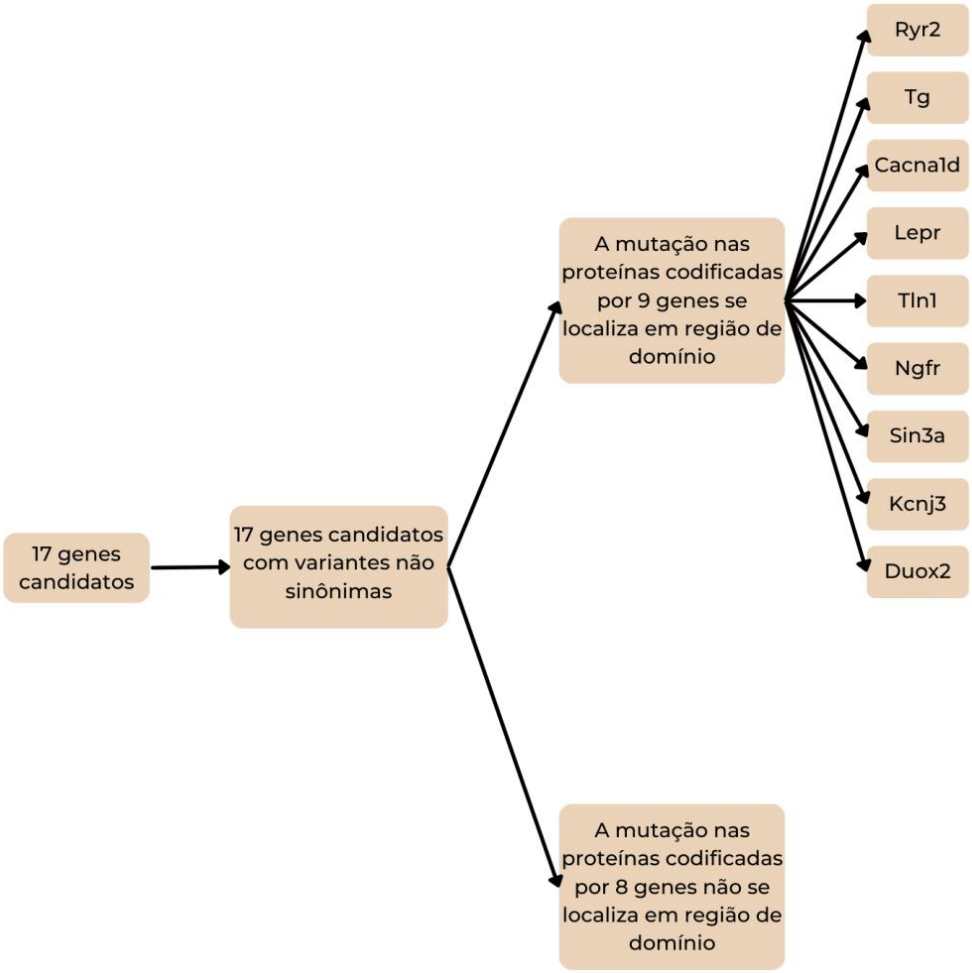

Figura 8 - Árvore de resultados da análise de conservação para *Mus musculus*. Dos 17 genes identificados no screening de camundongo, todos possuem variantes não sinônimas, das quais nove afetam região de domínio proteico (Ryr2, Tg, Cacna1d, Lepr, Tln1, Sin3a, Kcnj3 e Duox2).

Quando a metodologia foi aplicada ao peixe paulistinha, identificamos 16 genes ortólogos aos 17 genes identificados em camundongo, sendo que em todos os casos a variante genética identificada em camundongo se encontra em sequência de cDNA que se alinha entre as duas espécies. Apesar do alinhamento, apenas 10 genes ortólogos codificam o mesmo aminoácido afetado pela mutação em camundongo e seis apresentam aminoácido distinto (Figura 9).

Das sequências nas quais não houve conservação de aminoácido, três genes (Kcnj5, Ngfr e Sin3a) apresentam no mínimo um ortólogo que mostrou ao menos uma variante que afeta aminoácido em sítio de domínio proteico. Já

quando consideramos as sequências em que ocorreu conservação de aminoácido, quatro genes (Ryr2, Hcn1, Cacna1d e Kcnj3) apresentam ao menos um ortólogo que mostrou ao menos uma variante que afeta aminoácido em sítio de domínio proteico, como destacado na Figura 9.

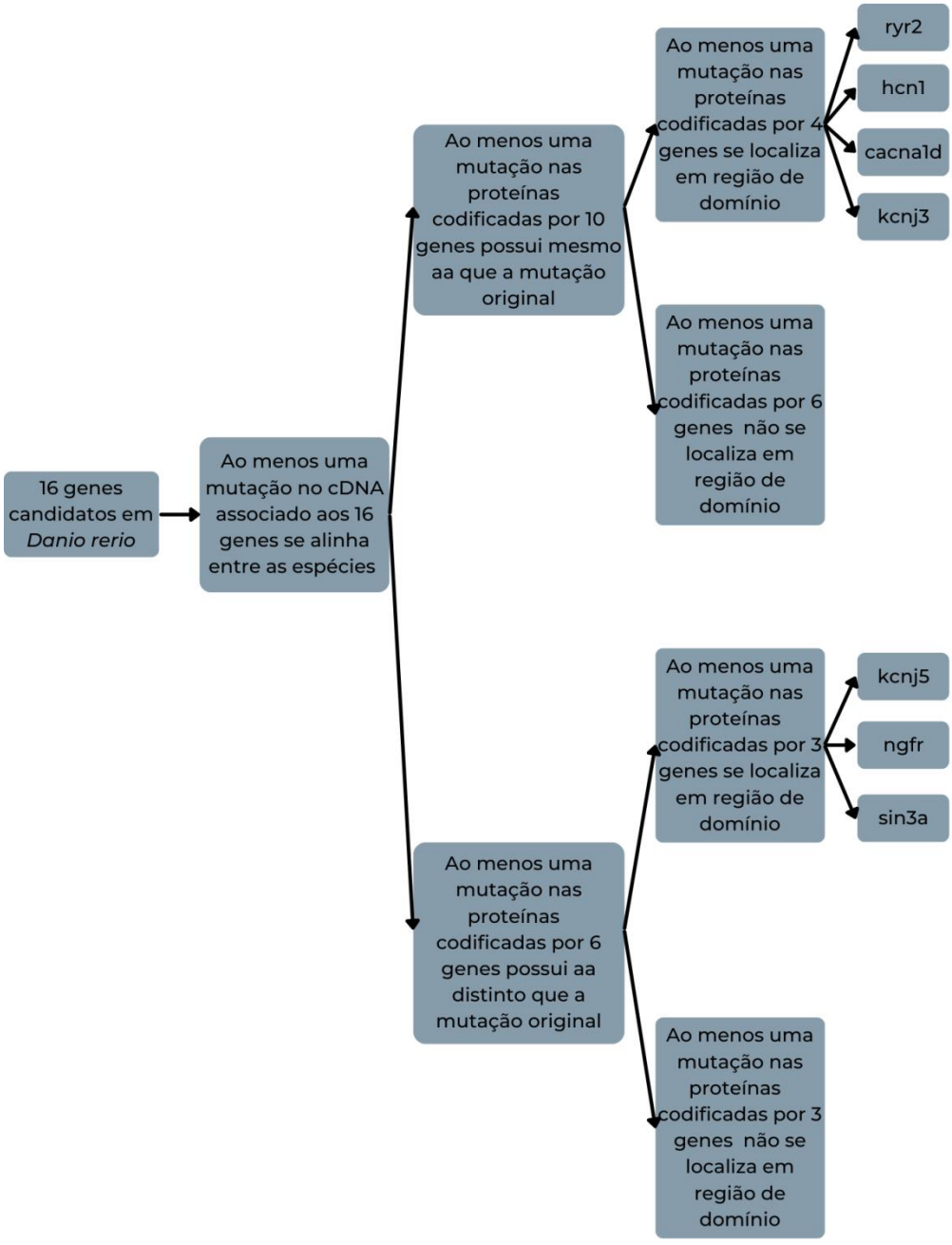

Figura 9 - Árvore de resultados da análise de conservação para *Danio rerio*. O esquema mostra a quantidade de genes obtidos em cada etapa da análise, e por fim a lista de genes que cumprem os requisitos estabelecidos.

Ao reproduzir a análise para a espécie humana, temos que todos os genes inicialmente identificados em camundongo possuem ortólogo em humano. Sendo assim, 15 dos 17 genes possuem aa afetado pela mutação conservado entre camundongos e humanos. Destes, em nove genes ao menos uma mutação afeta aa que se localiza em região de domínio proteico (Figura 10).

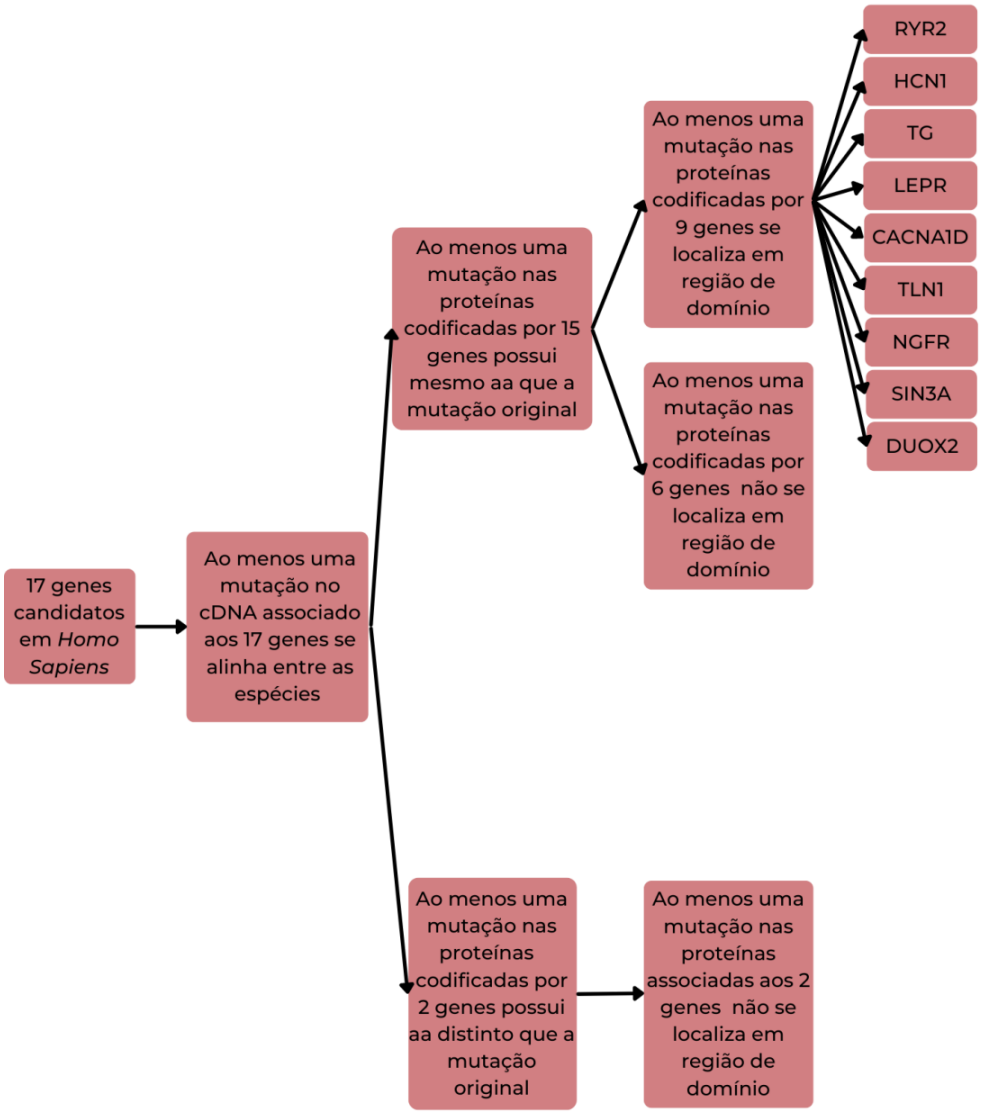

Figura 10 - Árvore de resultados da análise de conservação para *Homo sapiens*. O esquema mostra a quantidade de genes obtidos em cada etapa da análise, e por fim a lista de genes que cumprem os requisitos estabelecidos.

Ao considerar as análises individuais em cada espécie estudada e suas comparações com o genoma da espécie *Mus musculus*, avaliamos quais dos genes apresentaram em sua sequência proteica conservação de aminoácido

afetado pela variante genética identificada em camundongo em região de domínio proteico concomitante nas três espécies (Figura 11).

A análise apontou que dois genes possuem as características nas três espécies, sendo um associado à redução de frequência cardíaca (Ryr2) e um relacionado ao aumento do fenótipo (Cacnad1). Além disso, verificamos quais genes apresentaram concordância das análises em duas espécies: camundongo e peixe paulistinha - um gene associado à redução de frequência cardíaca (Kcnj3); peixe paulistinha e espécie humana - um gene associado à redução de frequência cardíaca (Hcn1); e espécie humana e camundongo - seis genes, sendo cinco associados à redução de frequência cardíaca (Tg, Lepr, Ngfr, Sin3a e Duox2) e um relacionado ao aumento do fenótipo (Tln1) (Figura 11).

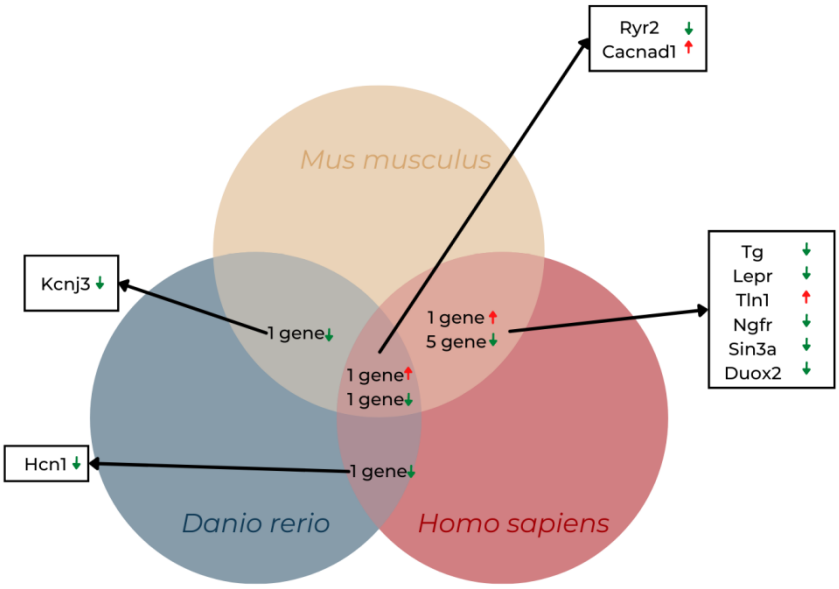

Figura 11 - Diagrama de Venn para a análise de conservação em genes conhecidos. O círculo bege representa a espécie *Mus musculus*, o azul *Danio rerio* e o vermelho *Homo sapiens*. Os genes representados são aqueles cujo cDNA se alinha com o cDNA de camundongo na região da variante associada à alteração do fenótipo de frequência cardíaca em camundongo, possui conservação do aminoácido afetado pela mutação entre as três espécies e que estão em região de domínio proteico. Em cada área de intersecção entre os círculos estão os genes que cumprem esses requisitos nas espécies indicadas.

4.1.2. Análise de predição de dano à estrutura proteica

Considerando as estruturas das proteínas codificadas pelos genes no camundongo, uma não foi analisada, pois as variantes genéticas nela identificadas como associadas ao fenótipo ocasionam um códon de parada, fato que não permite a análise pelo programa utilizado. Das 16 proteínas restantes, as estruturas obtidas experimentalmente não estão disponíveis. Sendo assim, buscamos estruturas preditas pelo AlphaFold. Com essa estratégia, identificamos 10 estruturas no AlphaFold, dos quais cinco estruturas apresentam a região que contém o aminoácido afetado predita com confiança *very high* ou *cofident*. Destas, ao serem submetidas ao Missense3d, apenas uma indicou possível dano estrutural (Cacnad1). As 11 sequências proteicas restantes, que não possuíam registro no banco de dados do AlphaFold ou que não cumpriam os requisitos estipulados foram submetidas ao software Phyre². Destas, apenas três continham os aspectos necessários para seguir para a predição de dano pela variante pelo Missense3d, sendo que o resultado obtido mostrou que as três estruturas apresentam possível dano devido a mutação (Figura 12). De todas as sequências proteicas codificadas pelos genes com variantes missense, oito não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na estrutura tridimensional da proteína.

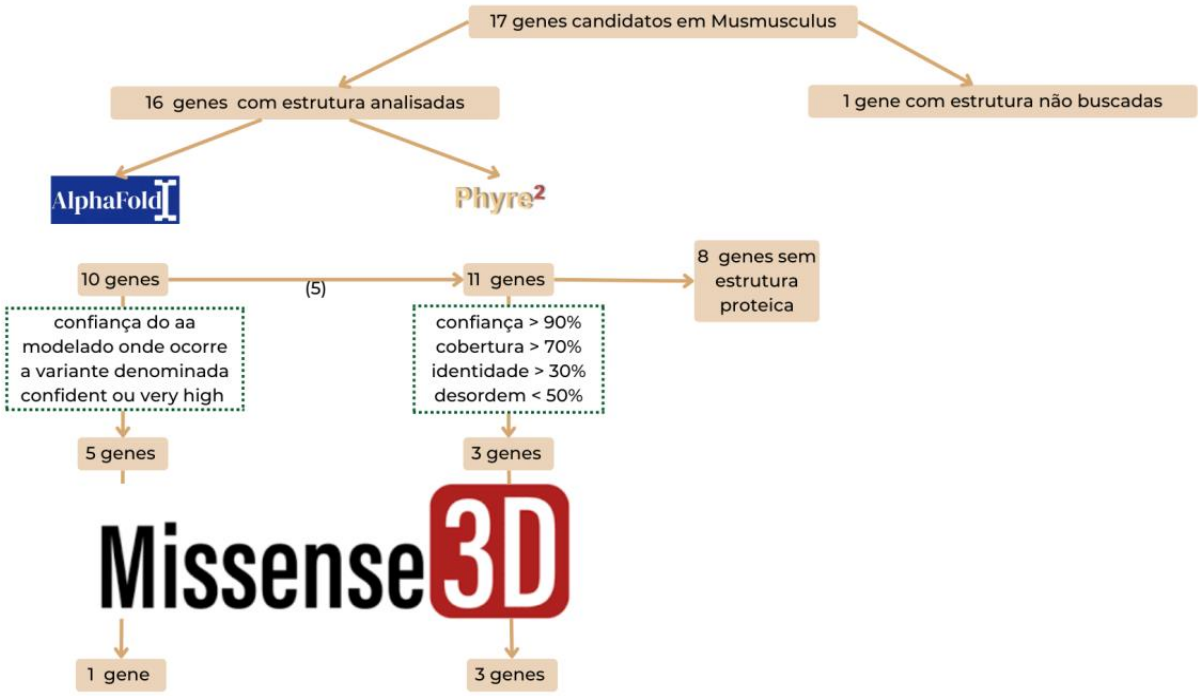

Figura 12 - Árvore de resultados da análise de predição de dano estrutural para os genes em camundongo. O esquema mostra o número de estruturas relacionadas aos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína associada.

As análises estruturais realizadas em peixe paulistinha utilizaram apenas as estruturas preditas pelo AlphaFold, visto que o banco de dados de estruturas obtidas experimentalmente dispõe de número limitado de estruturas proteicas (aproximadamente 480) e o Phyre² utiliza homologia para criar seus modelos considerando como molde estruturas proteicas provenientes do banco de dados essencialmente humanos e de camundongo. Sendo assim, das 16 com variantes *missense* foram identificadas nove proteínas com estrutura predita no AlphaFold, dos quais quatro estruturas contém a região com o aminoácido afetado predito com confiança *very high* ou *confident*. Ao submetê-las ao missense3d apenas uma apresentou dano estrutural devido a mutação (cacna1d) (Figura 13). Das sequências proteicas codificadas pelos genes com variantes missense 11 não foram analisadas por não possuir estrutura proteica com parâmetros adequados para prever o impacto da variante nela.

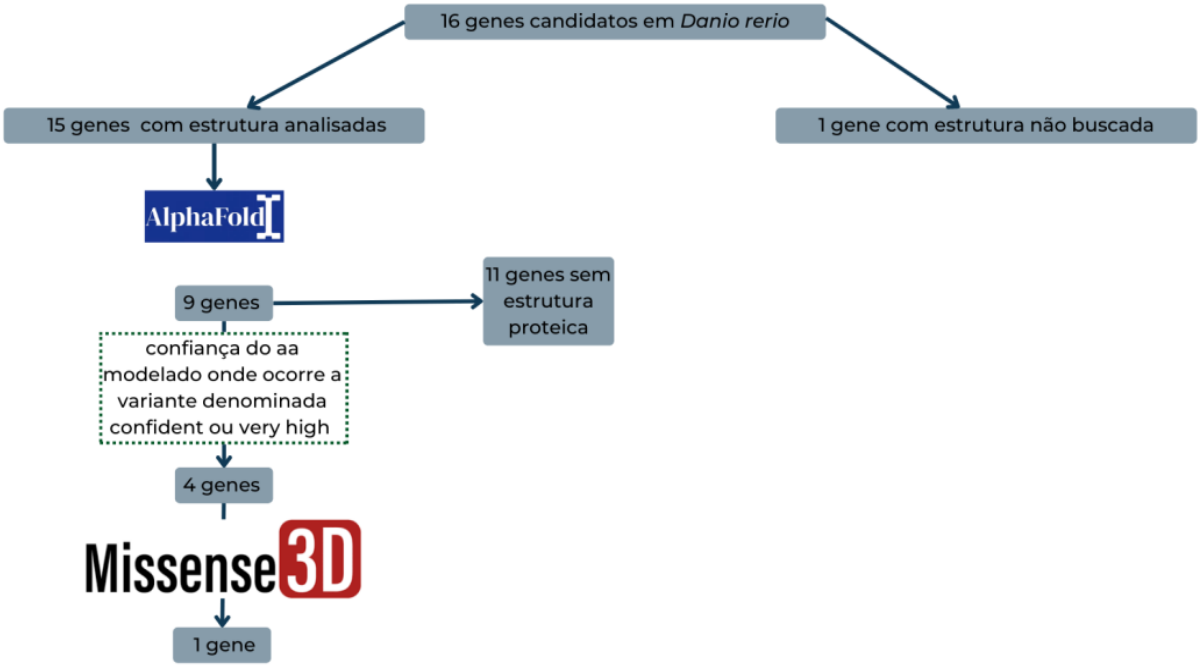

Figura 13 - Árvore de resultados da análise de predição de dano estrutural para os genes em peixe paulistinha. O esquema mostra o número de estruturas relacionadas aos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína associada.

A análise foi reproduzida para as proteínas em humano. Das 17 proteínas ortólogas, três tinham suas estruturas depositadas no banco de dados de proteína com estrutura solucionada experimentalmente, das quais duas apresentaram dano estrutural ao serem submetidas com as alterações missense correspondente ao camundongo, sendo elas codificadas pelos genes TG e TLN1. As demais 14 proteínas foram buscadas na plataforma AlphaFold, na qual se encontraram cinco estruturas proteicas cumprindo os requisitos para serem submetidas ao Missense3d. Destas, apenas duas delas apresentaram dano estrutural: CACNA1D e SLC8A1(Figura 14). Finalmente, das sete estruturas modeladas pelo Phyre2, apenas três atingiram os critérios necessários e apenas uma apresentou dano quando modelada com a alteração missense correspondente ao camundongo (RYR2) (Figura 14). Das sequências polipeptídicas codificadas pelos genes com variantes missense cinco não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na mesma.

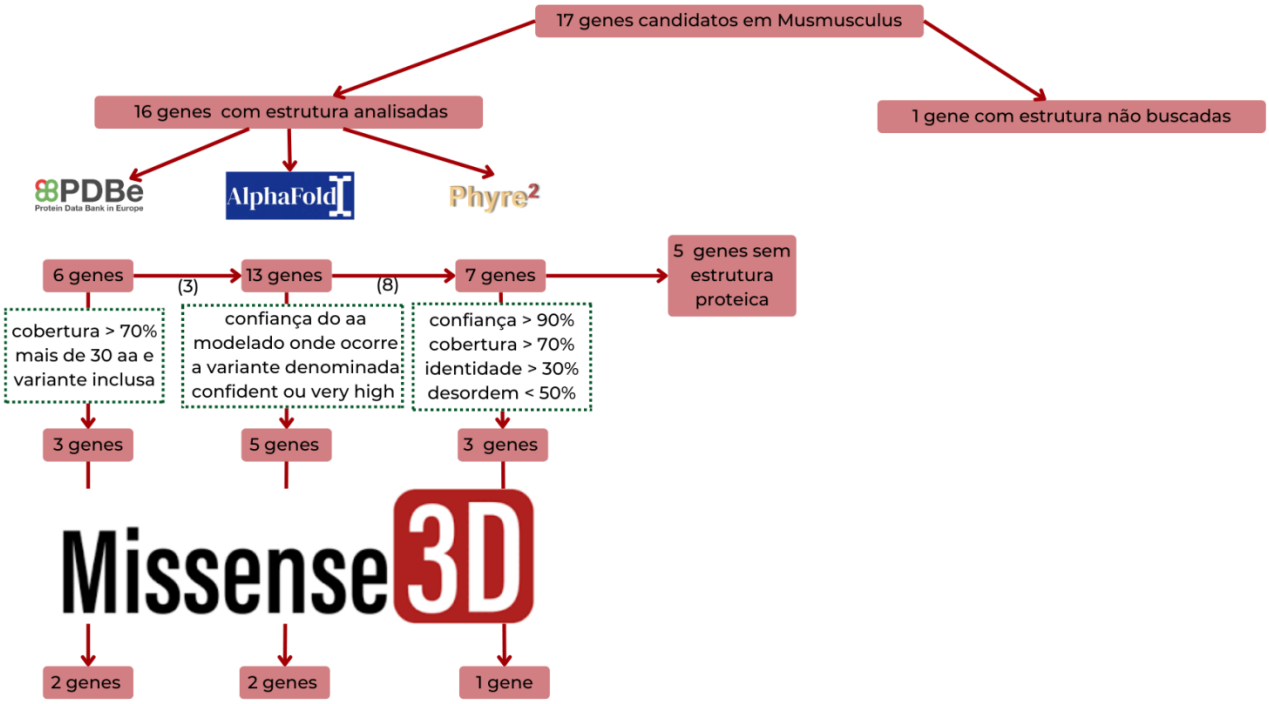

Figura 14 - Árvore de resultados da análise de predição de dano estrutural para os genes em humanos. O esquema mostra o número de estruturas relacionadas aos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína associada.

Ao término da análise de predição de dano à estrutura da proteína codificada pelos genes avaliados, realizamos uma comparação entre as três espécies (Figura 15). O único gene cuja proteína apresentou dano estrutural quando modelada com a mutação nas três espécies foi o Cacna1d. Vale salientar que para as três espécies a estrutura proteica utilizada foi a predita pelo AlphaFold. Além do gene Cacna1d, os genes Ryr2 e Tg também apresentaram dano da estrutura proteica quando modelados com a mutação na proteína do camundongo e na humana, sendo que as estruturas da proteína Ryr2 para ambas as espécies foram preditas pelo Phyre2, enquanto que a estrutura da proteína Tg humana foi experimentalmente obtida e a de camundongo foi predita pelo Phyre2 (Figura 15).

Finalmente, os genes SLC8A1 e TLN1 tiveram predição de dano a estrutura proteica quando modelados com a mutação apenas com proteína humana, enquanto que a proteína *Duox2* codificada pelo gene Duox2 apresentou dano à sua estrutura quando modelado com a mutação apenas em camundongo (Figura 15).

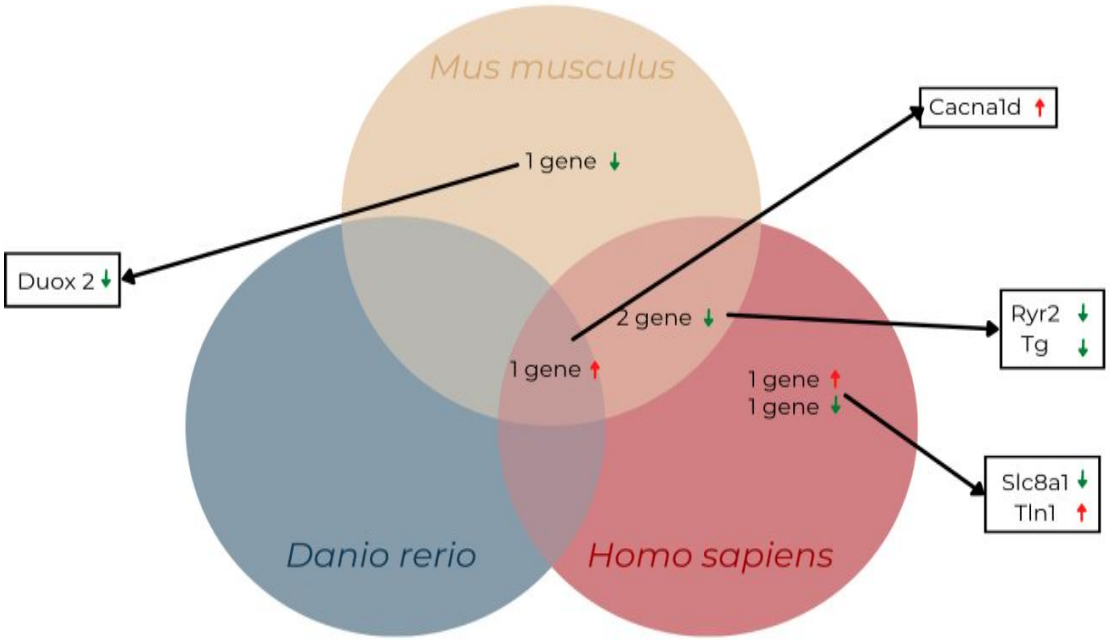

Figura 15 -- Diagrama de Venn de apresentação de resultados de dano estrutural. Temos no diagrama os genes cuja proteína apresenta dano estrutural ocasionado pela variante por espécie. As setas em verde mostram a diminuição de frequência cardíaca enquanto as vermelhas referem-se ao aumento da frequência.

4.2.3. Comparação entre análises

Comparamos, então, os resultados da análise de conservação e de dano estrutural a proteína (Figura 16). Durante a análise comparativa, é possível observar que a mutação do gene Cacnad1 ocorre em aa que é conservado e está em domínio proteico, assim como é predito de afetar a estrutura da proteica nas três espécies analisadas. Da mesma forma, o gene Ryr2 aparece em ambos os diagramas, mas, na análise de conservação, a mutação afeta aa conservado em região de domínio nas três espécies, enquanto a análise de predição de dano estrutural evidencia o efeito da mutação apenas na estrutura proteica de camundongo e de humano, uma vez que não há estrutura cristalográfica resolvida nem predita para zebrafish. Os genes Tg, Tln1 e Duox2 também foram selecionados em ambas as análises. No entanto, a mutação no gene Tg ocorre em aa conservado em domínio proteico e é predito de afetar estrutura proteica em *Mus musculus* e *Homo Sapiens*, enquanto que para a mutação no gene TLN1 há predição de afetar a estrutura proteica apenas em

humano. Finalmente, a mutação no gene Duox2 segue mesmo padrão que o Tln1, mas afeta apenas a estrutura proteica em camundongos.

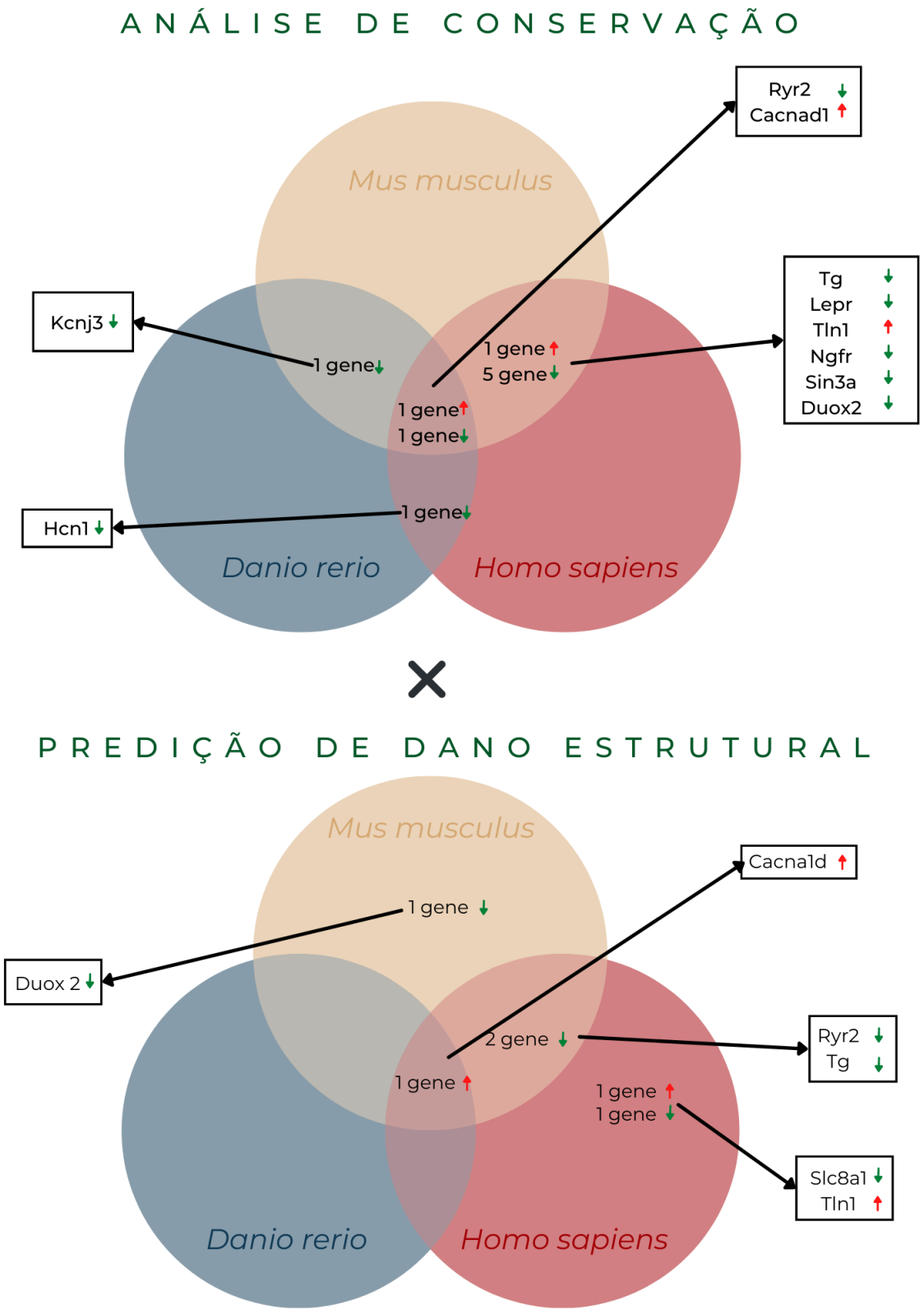

Figura 16 - Comparação entre as análises de conservação e análise de predição de dano estrutural.

As análises realizadas nos genes conhecidos apontam limitações na abordagem de priorização proposta, já que fica evidente que esta é dependente de informações disponíveis de bancos de dados públicos tanto para a avaliação de regiões conservadas e de domínio proteico, quanto para obtenção de estruturas proteicas. No caso das estruturas proteicas, há o acréscimo de uma camada na complexidade, pois além da dependência da disponibilidade das estruturas, estas precisam conter a região afetada pela mutação e terem alta qualidade de predição no caso de estruturas preditas. Nesse sentido, notamos que obtivemos um baixo número de estruturas proteicas com qualidade suficiente para a análise de dano, sendo apenas oito em camundongo, quatro em peixe paulistinha e 11 em humano, dentre as 16 proteínas estudadas.

4.2. Genes diretamente associados ao fenótipo

4.2.1. Análise de padrões conservativos

Dentre os 47 genes cujas variantes estão associadas à alteração de frequência cardíaca diretamente, 46 genes apresentam variantes não sinônimas, sendo que 13 possuem a variante em aminoácido em região de domínio proteico no camundongo (Figura 17). O gene que não apresentou variante não sinônima leva a alterações no *splicing*, ocasionando a deleção ou inserção de regiões exônicas ou intrônicas, alterando o quadro de leitura da proteína. Essas variações são de difícil predição e interpretação fazendo com que fossem excluídas das próximas análises.

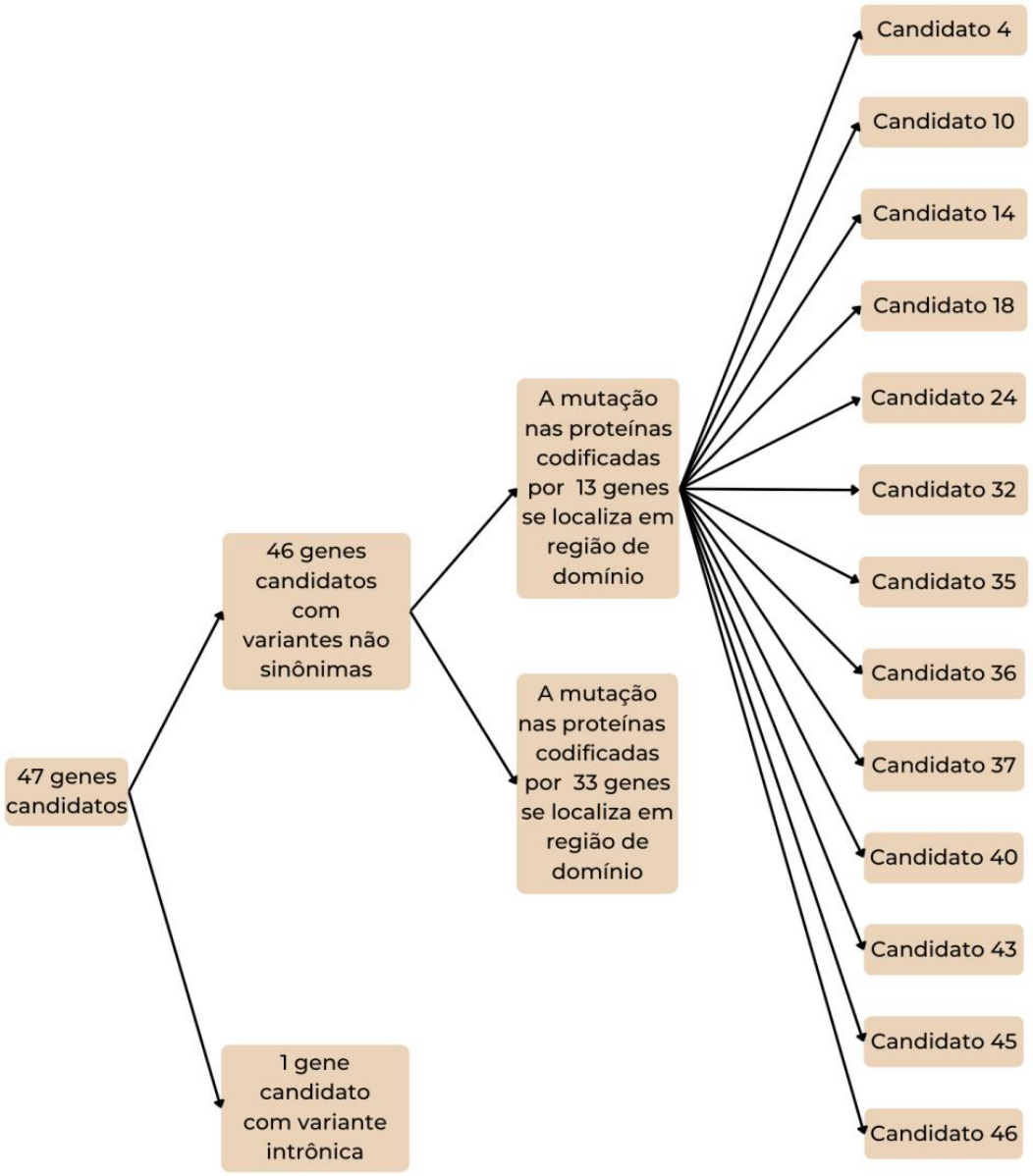

Figura 17 - Árvore de resultados da análise de conservação para Mus musculus. Dos 47 genes identificado no screening de camundongo, um apresenta variante intrônica enquanto 46 possuem variantes não sinônimas, das quais 13 afetam região de domínio proteico.

Em peixe paulistinha, 39 dentre os 46 genes identificados possuem pelo menos um ortólogo cuja variante genética identificada é não sinônima em camundongo e se encontra em sequência de cDNA que se alinha entre as duas espécies. Destes, 26 demostram conservação de aminoácido entre as duas espécies na posição onde se localiza a variante relacionada ao fenótipo em camundongo e 13 apresentam aminoácidos distintos entre as duas

espécies (Figura 18). Dos genes em que não houve conservação de aminoácido afetado pela variante genética, o aa afetado de cinco está em região de domínio proteico, enquanto que, dos genes em que houve conservação de aminoácido afetado pela variante genética, nove apresentam variante que afeta aa em região de domínio (Figura 18).

Figura 18 - Árvore de resultados da análise de conservação para Danio rerio. O esquema mostra a quantidade de genes obtidos em cada etapa da análise, e por fim a lista de genes que cumprem os requisitos estabelecidos.

Quando a análise é reproduzida para a espécie humana, temos que, dos 44 genes com variantes não sinônimas ortólogos aos identificados em camundongos, todos apresentam alinhamento entre os cDNAs das duas espécies no sítio afetado pela variante genética (Figura 19). Destes, dois não possuem conservação do aminoácido entre as espécies no sítio afetado pela variante genética e os aas não se localizam em região de domínio proteico. Dos 42 que têm pelo menos um ortólogo e possuem no mínimo uma variante genética que afeta o aminoácido conservado entre as espécies, 14 apresentaram domínio na região de ocorrência de troca de aa em decorrência de variante genética.

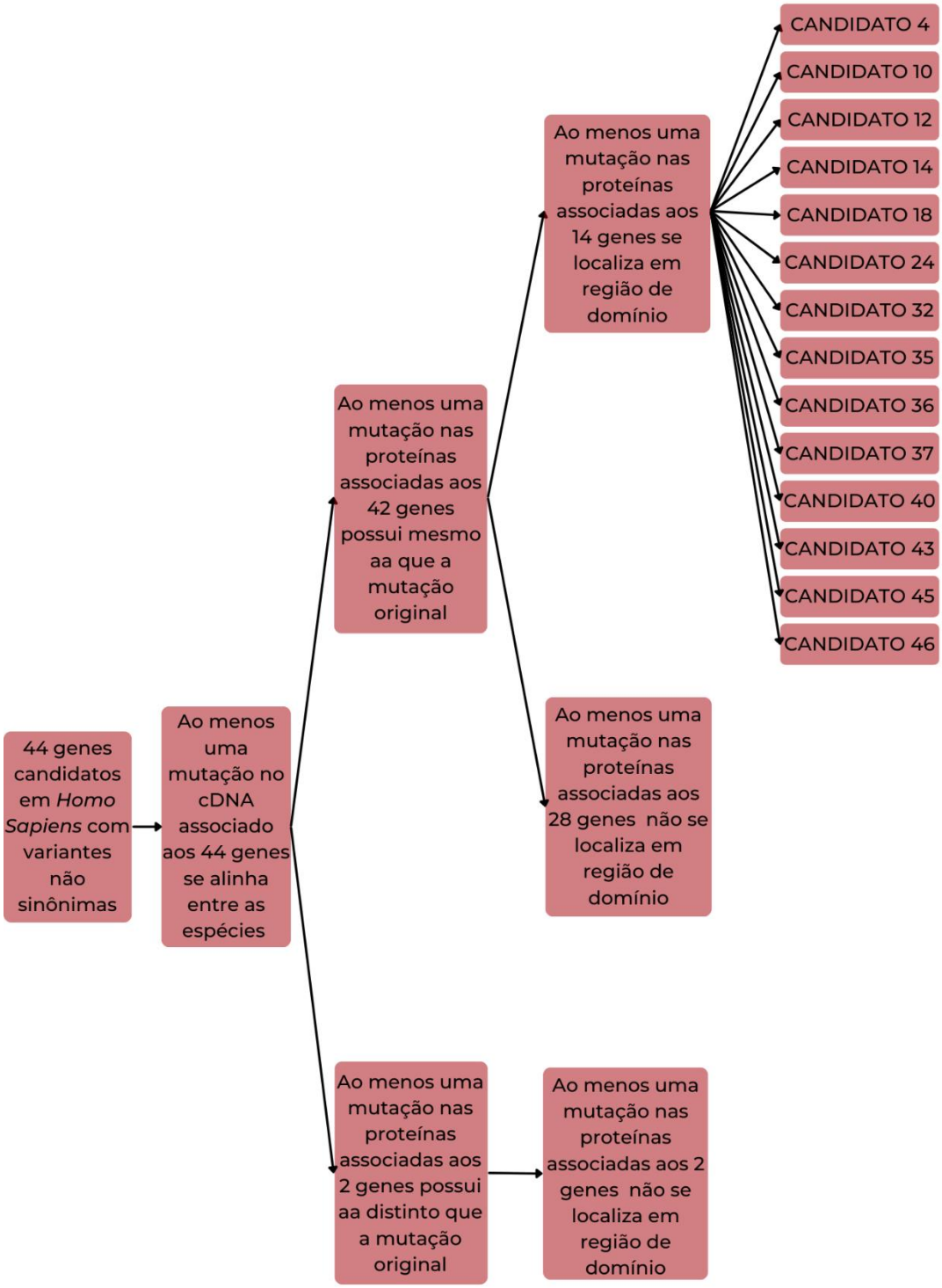

Figura 19 - Árvore de resultados da análise de conservação para *Homo sapiens*. O esquema mostra a quantidade de genes obtidos em cada etapa da análise, e por fim a lista de genes que cumprem os requisitos estabelecidos.

Considerando as análises em cada espécie estudada e suas comparações com o genoma do camundongo, obteve-se que seis genes apresentam variantes que afetam aa conservados em região de domínio nas

três espécies, sendo dois associados ao aumento de frequência cardíaca (Candidato 10 e Candidato 45) e quatro relacionados à diminuição da frequência cardíaca (Candidato 14, Candidato 32, Candidato 35 e Candidato 43). Além disso, outros seis genes apresentam variantes que afetam aa conservados em região de domínio em camundongo e na espécie humana, sendo também dois associados ao aumento de frequência cardíaca (Candidato 36 e Candidato 40) e quatro relacionados à diminuição da frequência cardíaca (Candidato 4, Candidato 18, Candidato 24 e Candidato 46). Finalmente, existem quatro genes que apresentam variantes que afetam aa conservados em comparação com camundongo em domínio proteico em apenas uma espécie, sendo três em zebrafish e um em humanos (Figura 20).

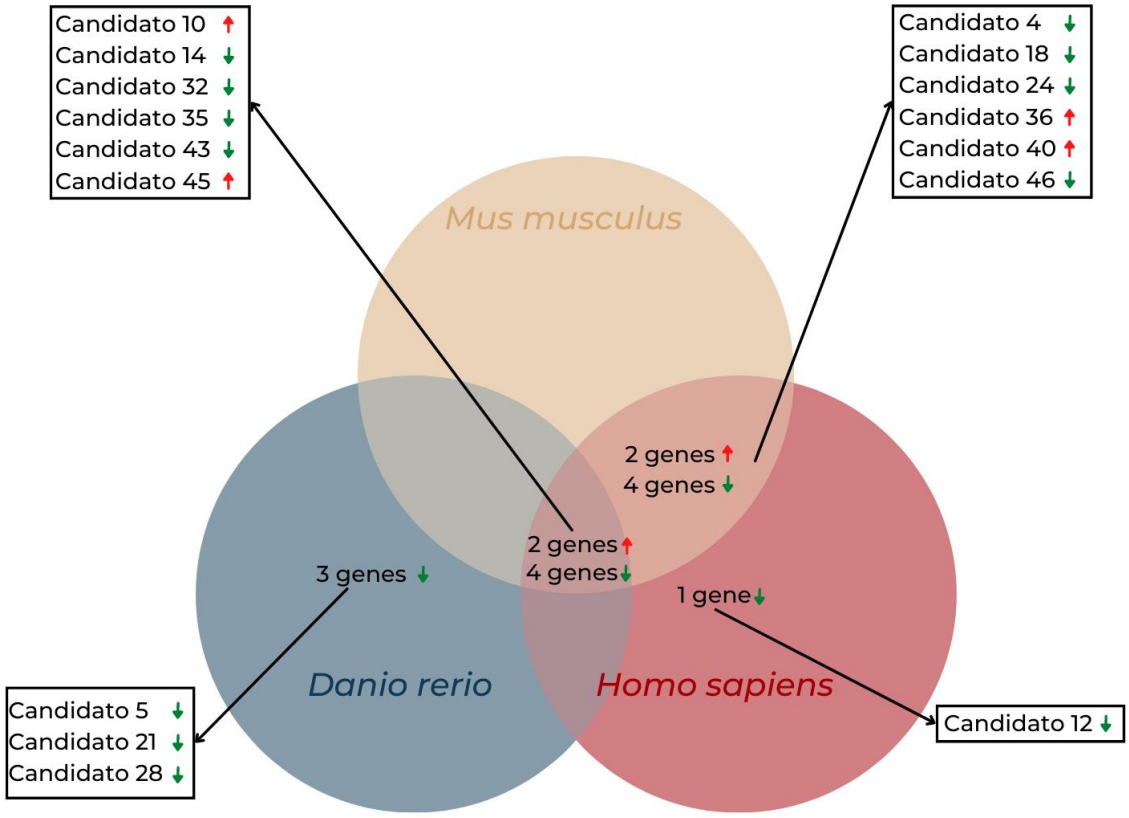

Figura 20 - Diagrama de Venn para a análise de conservação. O círculo bege representa a espécie *Mus musculus*, o azul *Danio rerio* e o vermelho *Homo sapiens*. Em cada área do círculo estão os genes que cumprem os requisitos e nas intersecções estão os que apresentam mesmo aa e domínio na região em que se encontra a variante nas espécies em questão.

4.2.2. Análise de predição de dano à estrutura proteica

Considerando as 46 proteínas codificadas pelos genes candidatos identificados em camundongo com alteração genética não sinônima, quatro não foram avaliadas por se tratar de proteínas cujos variantes causa um stop códon, não permitindo a análise pelo programa utilizado. Das 42 proteínas restantes, duas foram depositadas no banco de dados do PDBe cumprindo os parâmetros metodológicos, sendo que nenhuma apresentou dano estrutural quando submetida ao Missense3d (Figura 21). Das demais proteínas, 34 tiveram estruturas preditas disponíveis no banco de dados do AlphaFold, sendo que 25 cumpriram o requisito para serem submetidas ao preditor de dano estrutural e nesta análise, 11 apresentaram dano estrutural ocasionado pela alteração missense obtida na análise fenotípica. Por fim, as sequências de aminoácidos de 13 proteínas foram submetidas ao Phyre² sendo que apenas duas cumpriram os critérios para seguir a etapa seguinte e nenhuma estrutura apresentou dano estrutural. De todas as sequências proteicas codificadas pelos genes com variantes missense 13 não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na estrutura tridimensional da proteína.

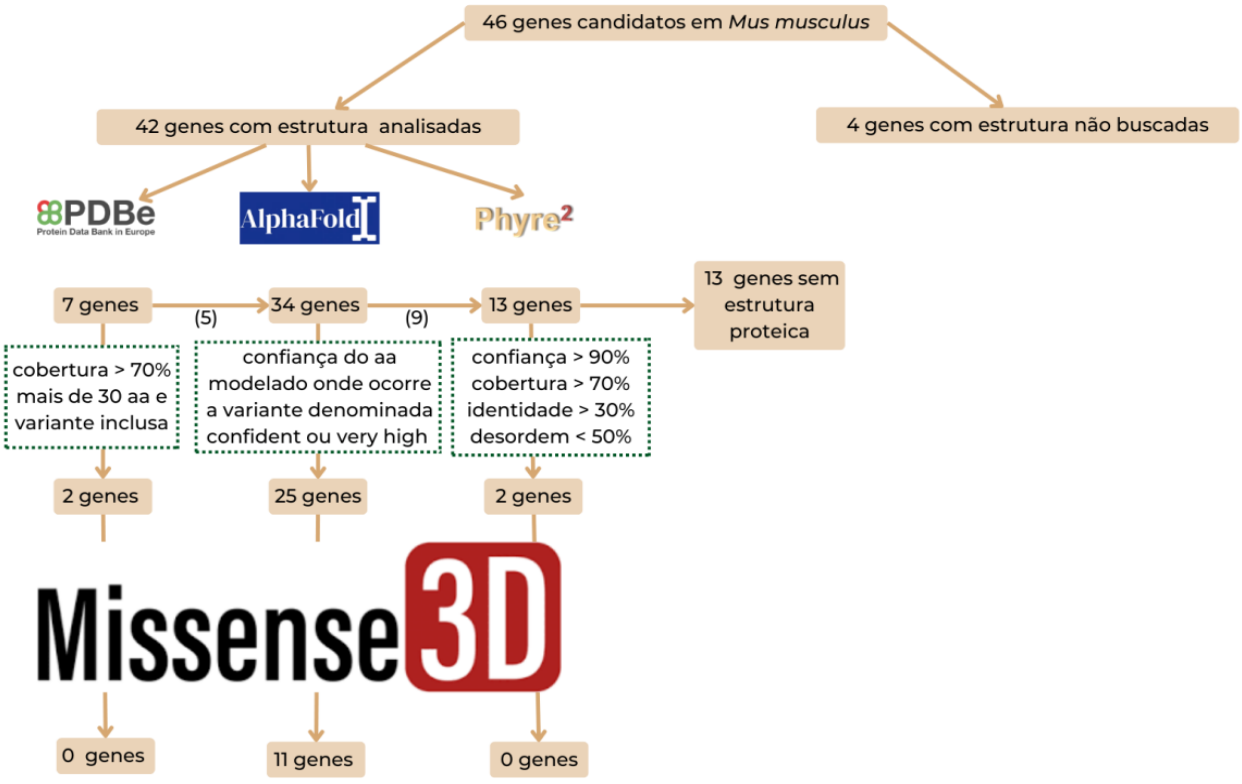

Figura 21 - Árvore de resultados da análise de predição de dano estrutural para os genes em camundongo. O esquema mostra o número de estruturas proteicas codificadas pelos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína codificada.

Dos 35 genes com variantes missense cujo ortólogo foi identificado em peixe paulistinha, 29 codificam proteínas cuja estruturas preditas estão disponíveis no banco de dados do AlphaFold. Destas, 22 foram submetidas ao Missense3d para verificar o possível dano à estrutura causado pela variante, sendo que sete apresentaram dano estrutural (Figura 22). Das sequências proteicas codificadas pelos genes com variantes missense seis não foram analisadas por não possuir estrutura proteica com parâmetros adequados para prever o impacto da variante nela.

Figura 22 - Árvore de resultados da análise de predição de dano estrutural para os genes em peixe paulistinha. O esquema mostra o número de estruturas proteicas codificadas pelos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína.

A mesma análise foi realizada para proteínas em humanos como demonstrada na Figura 23. Das 40 proteínas codificadas por genes ortólogos com variantes missense avaliadas, 16 apresentaram estrutura obtida experimentalmente depositadas no PDBe, sendo que apenas cinco puderam

ser submetidas ao Missense3d, obtendo como resultado uma proteína que apresenta dano estrutural quando sujeita à variante. Das demais proteínas, 33 tiveram estruturas preditas disponíveis no banco de dados do AlphaFold, sendo que 24 cumpriram o requisito para serem submetidas ao preditor de dano estrutural, retornando sete com dano estrutural causado pela alteração obtida no screening fenotípico. As sequências proteicas de nove proteínas foram submetidas ao Phyre² sendo que apenas uma cumpriu os requisitos para seguir para a análise no Missense3d e nenhuma estrutura apresentou dano estrutural. Das sequências proteicas codificadas pelos genes com variantes missense 10 não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na mesma.

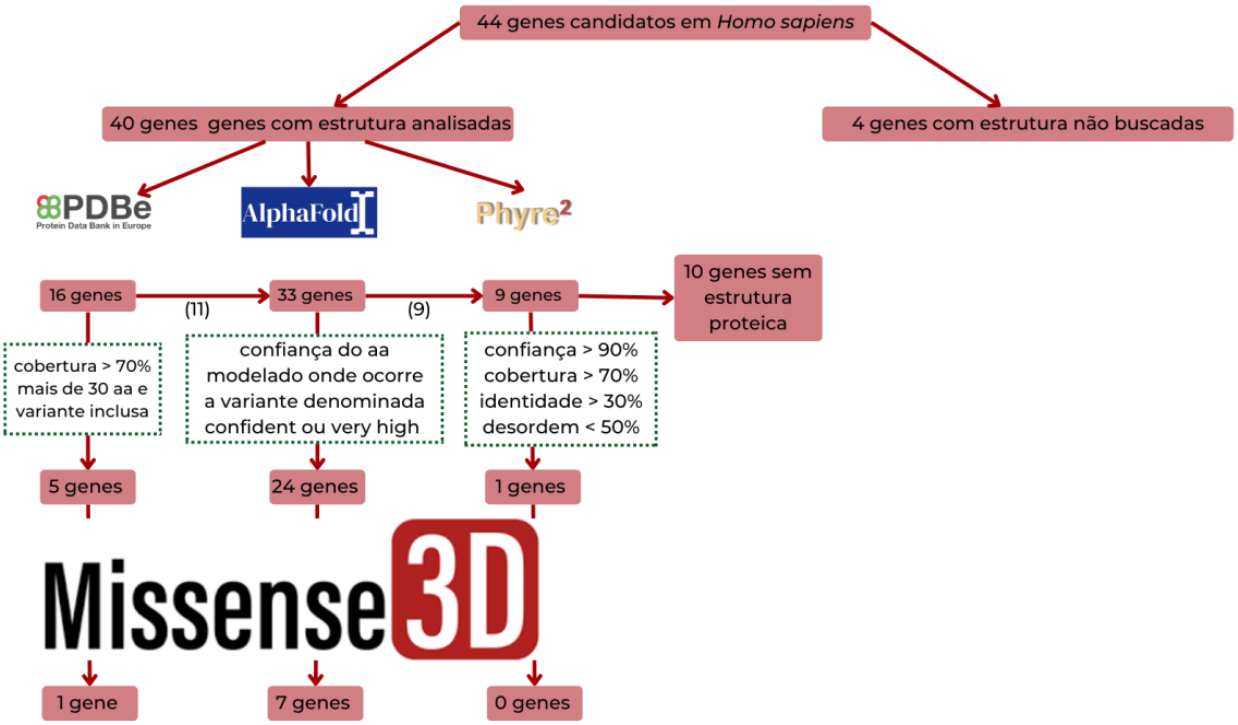

Figura 23 - Árvore de resultados da análise de predição de dano estrutural para os genes em humano. O esquema mostra o número de estruturas proteicas codificadas pelos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína.

Finalmente, uma análise comparativa entre as espécies foi realizada (Figura 24), de modo que as estruturas proteicas com alterações missense identificadas nos camundongos codificadas pelos genes Candidato 6, Candidato 16, Candidato 35 e Candidato 47 mostraram potencial dano estrutural em todas as espécies analisadas. As estruturas relacionadas aos

genes Candidato 29, Candidato 31 e Candidato 46 mostraram dano em camundongo e humano, enquanto o Candidato 3 mostrou dano em camundongo e peixe paulistinha. Também foram identificadas estruturas proteicas com dano pela variante em apenas uma espécie sendo: três para camundongo, duas para peixe paulistinha e uma em humano.

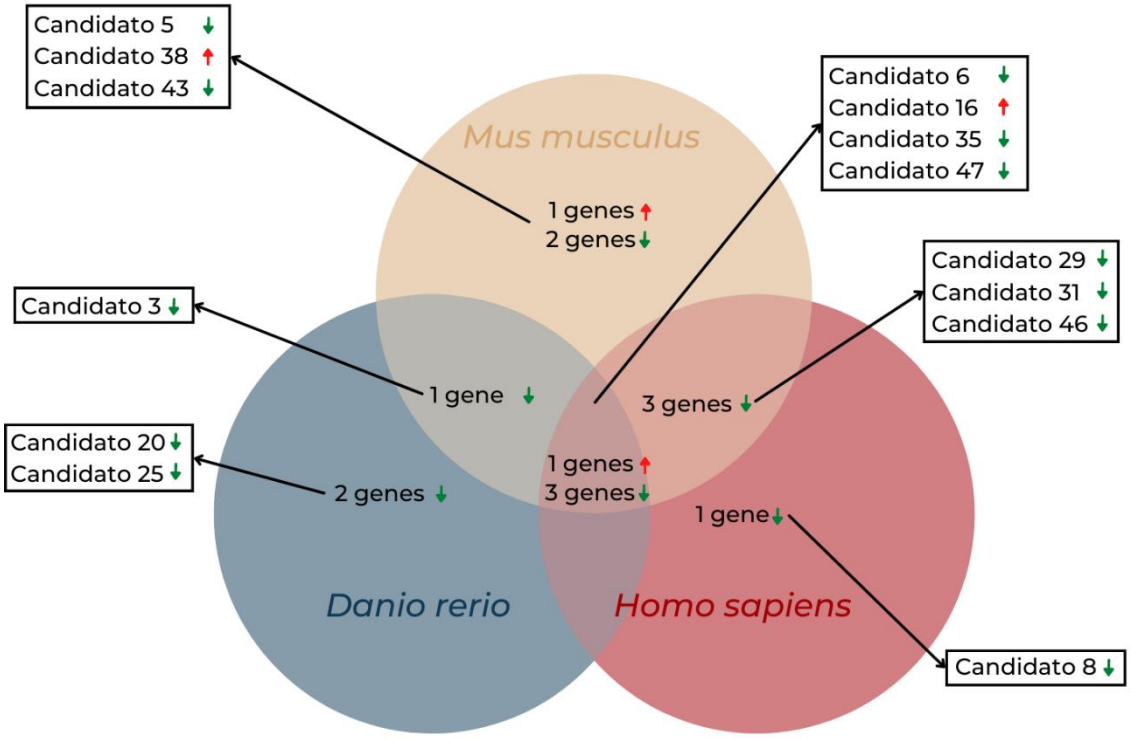

Figura 24 - Diagrama de Venn de apresentação de resultados de dano estrutural. Temos no diagrama os genes cuja proteína apresenta dano estrutural ocasionado pela variante por espécie. As setas em verde mostram a diminuição de frequência cardíaca enquanto as vermelhas correspondem a aumento de FC.

4.2.3. Comparação entre análises

Comparamos, então, os resultados da análise de conservação e de dano estrutural à proteína considerando as variantes genéticas associadas diretamente com a alteração do fenótipo (Figura 25). Dentre todos os candidatos, aqueles que apresentam maiores evidências biológicas de afetar funcional e estruturalmente a proteína foram os Candidatos 35 e 46. O primeiro por apresentar variante em mesmo aa e região de domínio, além da mesma ser predita por ocasionar alteração na estrutura proteica nas três espécies. Já o segundo candidato, por apresentar as características mencionadas acima, mas em camundongo e humano. O último não apresentou as características de interesse nas três espécies, pois não possui ortólogo em peixe paulistinha. Os

candidatos 35 e 46 serão analisados mais profundamente nas seções seguintes.

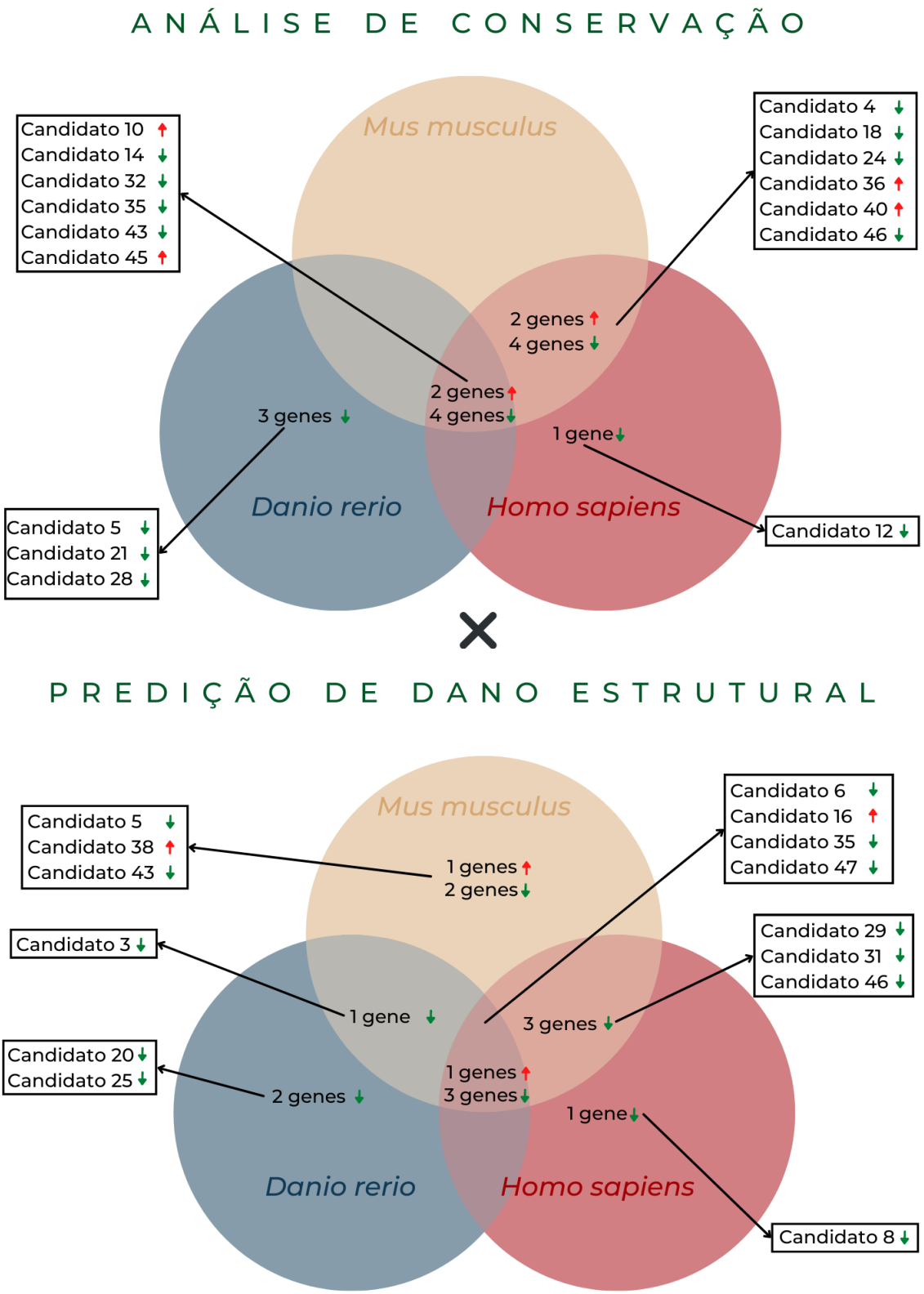

Figura 25 - Esquema comparativo dos diagramas das metodologias utilizadas.

4.2.4. Candidato 35

O gene candidato 35 foi selecionado por apresentar a variante genética associada à redução da frequência cardíaca em domínio proteico nas três espécies analisadas, *Mus musculus*, *Danio rerio* e *Homo Sapiens*, bem como ocasionar danos na estrutura proteica em ambas às espécies.

4.2.4.1. Análise fenotípica em camundongo

A obtenção do Candidato 35 como gene candidato associado à alteração de FC ocorreu durante o screening fenotípico da família R7096 de camundongos, cuja mutação ocasionada pelo agente mutagênico ENU levou a troca de uma arginina (R) por uma histidina (H) na posição 638 da cadeia polipeptídica e foi associada com uma redução de 16,32% na FC do camundongo homozigoto mutado quando comparado aos camundongos homozigoto selvagem e heterozigoto (Tabela 1 - Figura 26).

Tabela 1 - Apresentação das diferenças fenotípicas entre homozigoto selvagem (homo ref), heterozigoto (het) e homozigoto mutado para mutação no Candidato 35 no pedigree R7096.

Screen Name	Add	Ref Mean	Ref SD	Het Mean	Het SD	% vs homo ref	dif vs homo ref	Var Mean	Var SD	% vs homo ref	dif vs homo ref
HR Average	4.72e-04	649.98	41.27	627.30	51.32	-3.49	-22.67	543.9	65.31	-16.32	-106.08

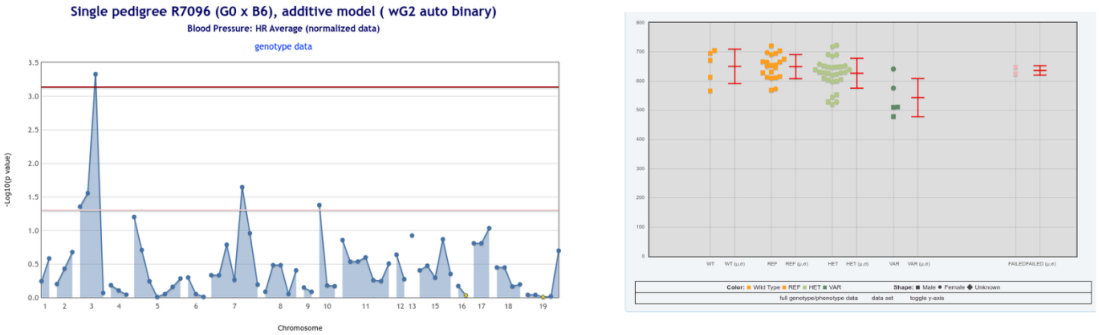

Figura 26 - Manhattan Plot de variantes identificadas no pedigree R7096 (esquerda) e gráficos de FC (direita) de dados da média dos dois diasEm detalhe, no Manhattan

plot, no eixo x a posição cromossômica em que se encontram as variantes ocasionadas pelo ENU e testadas quanto sua associação no fenótipo de FC e no eixo y é possível ver o valor de –log10(p valor), mostrando o impacto da variante no fenótipo. O ponto acima da linha em vermelho mostra que a variante no cromossomo 11 está associada à redução de frequência cardíaca. Nos gráficos de relação fenótipo x genótipo, são mostrados da esquerda para direita, animais controle C57(WT - laranja), homozigoto selvagem (REF - laranja), heterozigoto (HET - verde claro), e homozigoto mutado (VAR - verde escuro).

4.2.4.2. Resultados das análises de predição de dano à estrutural

Para analisar o efeito da variante na proteína codificada pelo Candidato 35 foram utilizadas as estruturas preditas disponíveis no AlphaFold DB para cada uma das espécies avaliadas, sendo que os aminoácidos nos quais ocorre a variante apresentaram um pLLDTs superior a 90, que corresponde a uma confiança “very high”. A análise de predição de dano à estrutura proteica em camundongo indica que a variante R638H leva à quebra de uma ponte salina formada, bem como para as posições análogas nas demais espécies.

Considerando a estrutura de camundongo utilizada para fazer as predições, esta possui 756 aminoácidos e um peso de 82,99 KDa. A estrutura foi submetida à uma análise no PDBsum, que fornece informações como diagrama de estrutura secundária (Figura 27 - A), mapa topológico (Figura 27 - B), e diagrama de Ramachandran (Figura 27 - C).

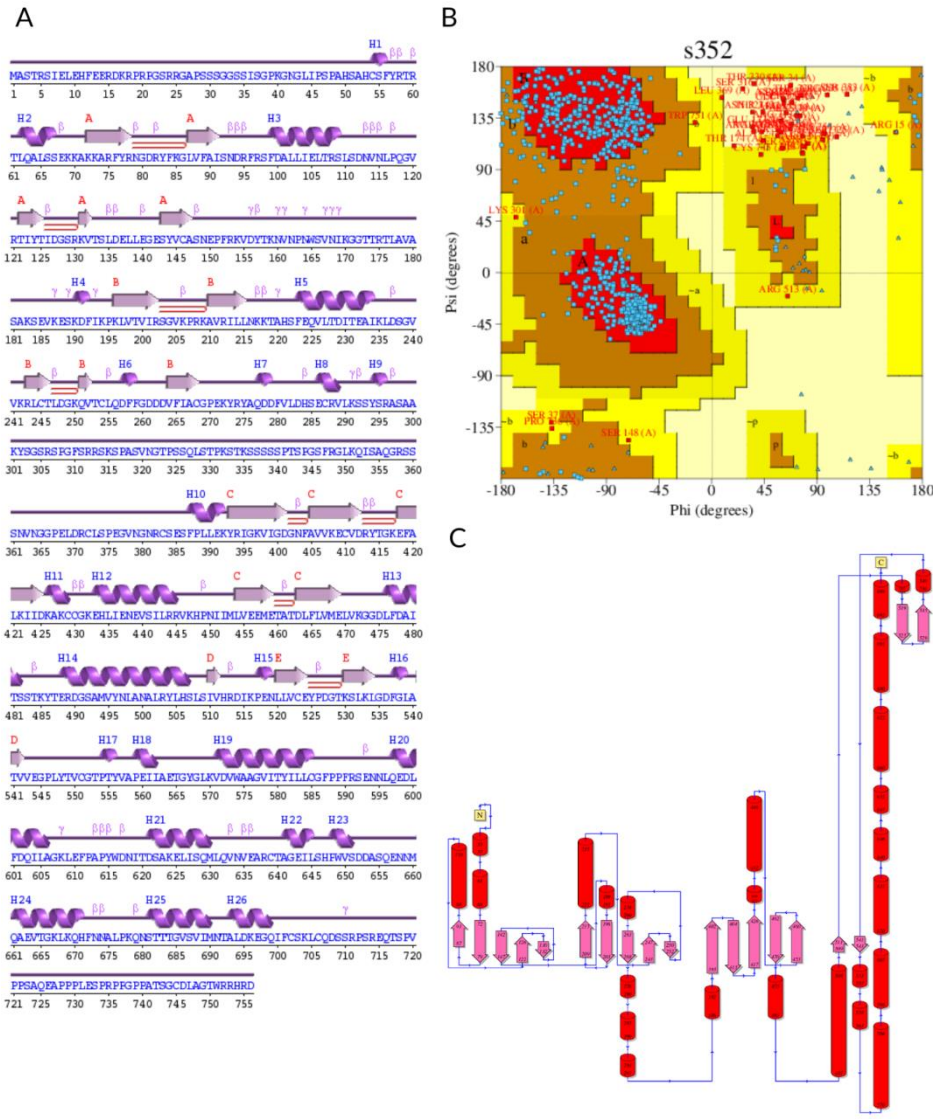

Figura 27 – Análises PDBsum *Candidato 35*.(A) Mapa de estrutura secundária, mostrando os motivos estruturais. As setas em lilás representam as fita beta, as helicoidais em roxo representam as hélices e as linhas curvadas em vermelho os *hairpins*. (B) Diagrama de Ramachandran. O diagrama apresenta a conformação dos ângulos phi e psi para cada resíduo. Os resíduos são representados em azul claro, as regiões vermelhas são as mais favoráveis, seguidas pelas em marrom que são adicionalmente permitidas, as em amarelo são permitidas e em amarelo claro, regiões não permitidas. (C) Mapa topológico. Iniciando do N terminal para o C terminal, mostra as fitas em rosa e as hélices em vermelho.

A análise de estrutura secundária indica a presença de cinco folhas beta (conjuntos de setas em rosa na Figura 27 - C), composta por 19 fitas (setas lilás na Figura 27 - A e rosas na Figura 27 - C) formando oito beta “bulges”; oito beta “hairpin (seta em vermelho Figura 27 - A), vinte e seis hélices (hélice roxa na figura 26A e cilindro vermelho na Figura 27 - C) formando vinte e três interações hélice-hélice, além de quarenta e nove voltas beta e 15 voltas gama. Já a análise do diagrama de Ramachandran representada na Tabela 2, indica que a

estrutura possui 80,2% de seus resíduos em regiões permitidas, 13,1% em regiões adicionalmente permitidas e 20,6% em regiões generosamente permitidas (Figura 27-C). Estes dados não possuem excelente qualidade em termos de disposição pelas regiões do gráfico, mas cabe lembrar que estamos tratando de uma estrutura predita e algumas estruturas não estão bem definidas. Os aminoácidos nas regiões não permitidas provavelmente correspondem às regiões que o AlphaFold classifica como *very low confidence* e regiões que são classificadas como desordem. Enquanto a primeira análise apresenta os motivos proteicos e ligações químicas que compõe o domínio, a segunda apresenta a localização destes aminoácidos no gráfico mediante sua conformação angular, de modo a indicar os domínios e a estabilidade da proteína. Pelo diagrama, juntamente à Tabela 2, vemos que o domínio apresenta estabilidade ao possuir todos seus resíduos em regiões permitidas.

Tabela 2 - Dados estatísticos do Diagrama de Ramachandran da estrutura proteica obtida para o *Candidato 35* pelo AlphaFoldDB.

	N°. Of residues	**%-tage**
Most favoured regions [A,B,L]	525	80,2%**
Additional allowed regions [a,b,l,p]	86	13,1%
Generously allowed regions [~a,~b,~l,~p]	17	2,6%
Disallowed regions [XX]	27	4,1%
Non-glycine and non-proline residues	655	100.0%
End-residues (excl. Gly and Pro)	2	
Glycine residues	54	
Proline residues	45	
Total number of residues	756	

4.2.4.3. Resultados das análises de domínio

O código Uniprot Q6PGN3 fornece informações revisadas pela Swiss-Prot da proteína de estudo - pertencente a família serina/treonina quinase, em camundongo. A mutação, quando analisada nas três espécies, se localiza no mesmo domínio, o domínio proteína quinase. Variantes neste domínio estão associadas a doenças como

câncer e Parkinson[57,58]. No contexto da proteína estudada, quando o domínio quinase se fusiona ao domínio duplocortina (DC), este se torna funcionalmente ativo, de modo a permitir a interação com microtúbulos, atuando em sua estabilização[59].

4.2.4.4. Simulação de Dinâmica Molecular para investigar efeito da variante no comportamento do domínio

Na simulação de dinâmica molecular de 30ns com a estrutura da proteína de camundongo, foi possível visualizar uma alteração na conformação da proteína ao longo do tempo resultante da quebra de uma ponte salina (cuja distância normalmente é de 0,47nm) envolvendo o aminoácido afetado pela variante. A ponte salina é uma ligação química do tipo não covalente entre resíduos de cargas opostas que estão localizados suficientemente próximos ao ponto de serem submetidos à atração eletrostática [60]. Esta ligação contribui para a estabilidade do dobramento proteico e é dependente do pH do sistema. Uma alteração na ponte comumente leva a desestabilização da proteína, uma vez que não é possível recuperar a perda de energia livre[60].

No presente caso, apresentamos a distância entre os átomos constituintes da ponte salina ao longo da simulação: o nitrogênio NE pertencente a arginina 638 no domínio quinase e o oxigênio OE1 do glutamato 560 presente no domínio DC (Figura 28). O gráfico mostra em verde a distância atômica entre esses dois resíduos em sua forma nativa, que fica em torno de 0,47nm durante toda a simulação. Em contrapartida, como podemos observar na linha vermelha, quando submetida a troca da arginina por uma histidina na posição 638, a distância entre o nitrogênio da histidina 638 e a oxigênio do glutamato 560 apresenta não só um valor superior, como sofre uma grande variabilidade ao longo da simulação, com uma distância mínima de 0,47 nm no início da simulação e máxima de aproximadamente 1,2nm próximo ao momento 25ns da simulação (Figura 28).

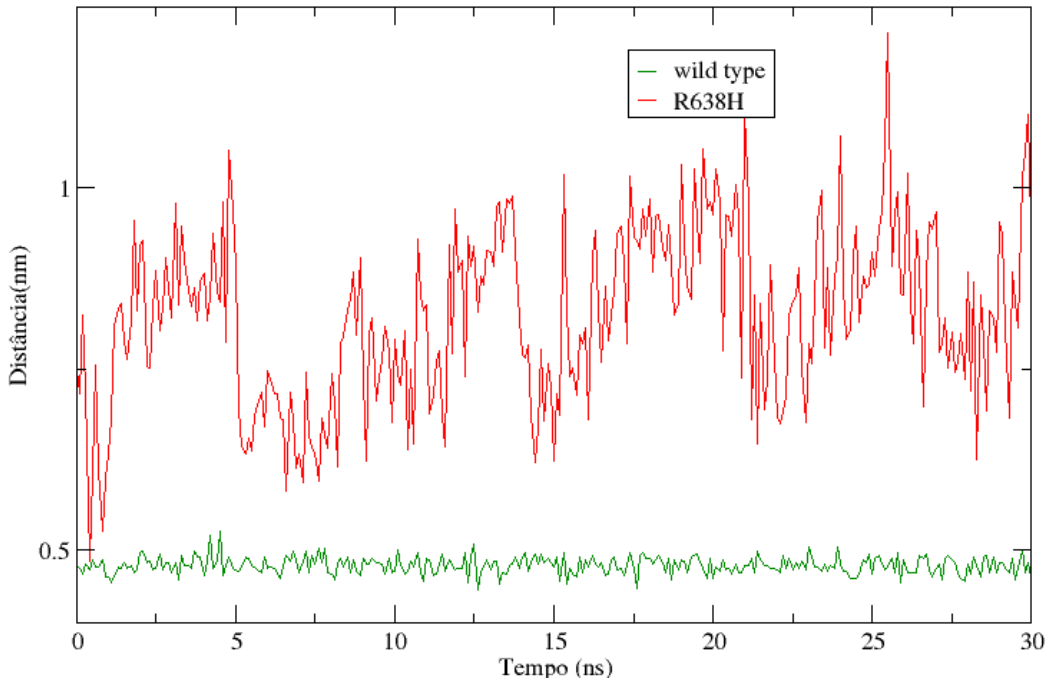

Figura 28 - Gráfico da distância entre o nitrogênio NE da arginina (selvagem) e histidina (mutante) 638 e o oxigênio OE1 glutamato 560, em nanômetros, ao longo do tempo de simulação de dinâmica molecular, em nanosegundos. Em verde temos a linha da proteína selvagem (*wild type*), mostrando a distância referente a uma ponte salina. Em vermelho, encontra-se a representação para a proteína resultante da variante R638H.

A análise preliminar do docking, realizado para avaliar o efeito da variante sob a ótica da interação entre a proteína e o microtúbulo, representado pela tubulina, indica que, ao longo do tempo, a estrutura com a variante apresenta um enovelamento distinto da nativa (Figura 29). No instante inicial (t=0), as estruturas apresentaram conformação e interação com a tubulina similar. Após 30 ns de simulação de dinâmica molecular (t=30), é possível observar que enquanto a proteína nativa apresenta a interação entre os domínios quinase, em verde escuro, e DC, em azul, sendo que este último interage e está mais próximo da tubulina, e o mesmo não ocorre para a estrutura proteica com a variante R638H. É possível observar que a quebra da ponte salina levou à alteração na conformação proteica de modo que o domínio DC não mais interage com o domínio quinase e este último se localize mais próximo da tubulina.

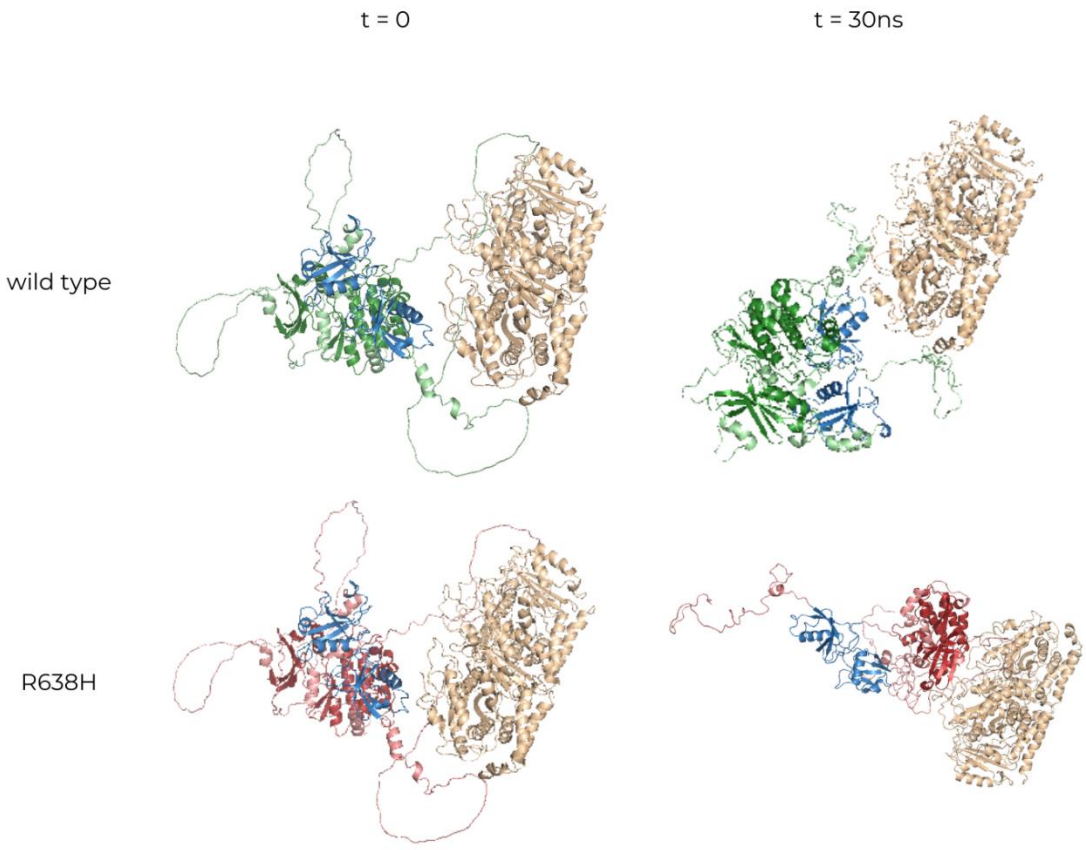

Figura 29 – Análise de docking entre a proteína *Candidato 35* selvagem e com a mutação R638H. Apresentamos a interação entre a estrutura da proteína de estudo e a molécula de tubulina (em bege)(código PDB 1TUB) nos instantes 0 e 30ns para as proteínas em seu estado nativo (em verde) e resultante da virante (em vermelho). Nas representações temos os domínios DCs em azul enquanto o domínio quinase é verde no wild type/nativa e vermelho no mutado.

4.2.4.5. Hipótese da contribuição do Candidato 35 na alteração de FC

O Candidato 35 é um gene que codifica uma proteína pertencente à família serina/treonina quinase, uma das maiores de que se tem conhecimento , sendo seus membros codificados por aproximadamente 2% dos genes eucariotos. A proteína é constituída por 756 aminoácidos, agrupados em dois domínios duplocortina, três regiões de desordem, um conjunto de resíduos ácidos e básicos, e quatro conjuntos de resíduos polares.

Edelman e colaboradores demonstraram que esta proteína possui uma associação com microtúbulos[59]. É de conhecimento pela literatura que a proteína atua na estabilização dos microtúbulos[59,61], bem como é utilizado como um marcador para eles em estudos com peixe paulistinha[62]

A proteína tem seu domínio duplocortina (DC) funcionalmente ativado quando unido ao domínio quinase. O domínio quinase é independente do DC quanto a manifestação de sua atividade catalítica, que, quando fosforilado, a reduz a afinidade da proteína com o microtúbulo. Assim, sugere-se que a autofosforilação desta proteína leva a regulação da afinidade com o microtúbulo[59].

Os microtúbulos são estruturas compostas por um arranjo de tubulinas em um tubo rígido de aproximadamente 25nm de diâmetro. Eles compõem, juntamente com filamentos intermediários e microfilamentos de actina, o citoesqueleto, responsável por manter o formato e locomoção da célula[63]. Em cardiomiócitos, os microtúbulos juntamente com os filamentos intermediários compõem o citoesqueleto não sarcomérico. No remodelamento patológico, a proliferação de microtúbulos e de filamentos intermediários aumentam a viscoelasticidade e impedem fisicamente a contração e o relaxamento de miócitos isolados, podendo afetar o ritmo cardíaco[64].

Alicerçado nas informações extraídas dos resultados das análises *in silico*, conjuntamente aos conteúdos adquiridos pela busca na literatura, hipostenizamos que a alteração de uma arginina por uma histidina na posição 638 da proteína, sitiada no domínio quinase, rompe uma ponte salina desestabilizado o dobramento proteico, de modo a mudar a conformação dos domínios. Esta mudança por suposição alterará a interação da proteína com o microtúbulo, afetando a estrutura e funcionalidade do citoesqueleto não sarcomérico de cardiomiócitos e, consequentemente, sua contração e relaxamento. Finalmente, essas alterações podem levar a uma alteração no ritmo cardíaco.

4.2.5. Candidato 46

O gene candidato 46 foi selecionado por apresentar a variante genética associada à redução da frequência cardíaca em domínio proteico em duas espécies analisadas, *Mus musculus* e *Homo Sapiens*, bem como ocasionar danos na estrutura proteica em ambas às espécies.

4.2.5.1. *Análise fenotípica em camundongo*

A obtenção do Candidato 46 como gene candidato associado à alteração de FC ocorreu durante a análise fenotípica da família R7750 de camundongos, cuja mutação ocasionada pelo agente mutagênico ENU levou a troca de uma metionina (M), um resíduo apolar, por uma lisina (K), um resíduo polar carregado, na posição 52 da cadeia polipeptídica e foi associada com uma redução de 20,9% na FC do camundongo homozigoto mutado quando comparado aos camundongos homozigoto selvagem (Tabela 3 e Figura 30).

Tabela 3 – Apresentação das diferenças fenotípicas entre homozigoto selvagem (homo ref), heterozigoto (het) e homozigoto mutado para mutação no Candidato 46 no pedigree R7750.

Screen Name	Rec	Ref Mean	Ref SD	Het Mean	Het SD	% vs homo ref	dif vs homo ref	Var Mean	Var SD	% vs homo ref	dif vs homo ref
HR Average	7.64e-4	678.1	48	664.6	49.8	-2	-13.50	536.3	61.2	-20.9	-141.80

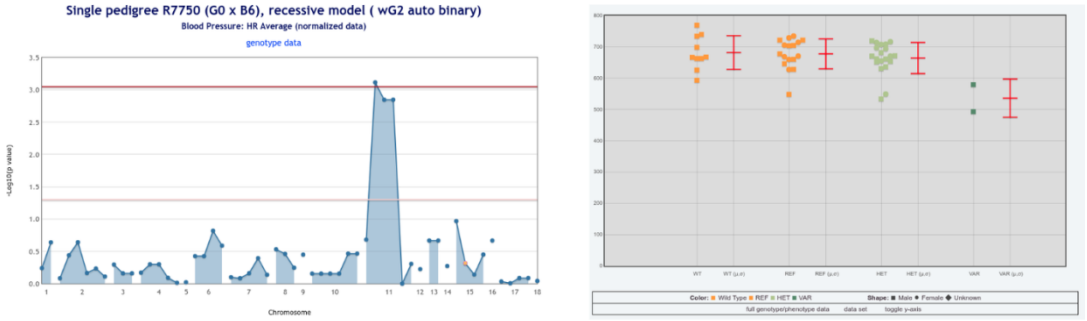

Figura 30 - Manhattan Plot de variantes identificadas no pedigree R7750 (esquerda) e gráficos de FC (direita) de dados da média dos dois dias. Em detalhe, no Manhattan plot, no eixo x a posição cromossômica em que se encontram as variantes ocasionadas pelo ENU e testadas quanto sua associação no fenótipo de FC e no eixo y é possível ver o valor de –log10(p valor), mostrando o impacto da variante no fenótipo. O ponto acima da linha em vermelho mostra que a variante no cromossomo 11 está associada à redução de frequência cardíaca. Nos gráficos de relação fenótipo x genótipo, são mostrados da esquerda para direita, animais controle C57(WT - laranja), homozigoto selvagem (REF - laranja), heterozigoto (HET - verde claro), e homozigoto mutado (VAR - verde escuro).

4.2.5.2. *Resultados das análises de predição de dano à estrutura*

Segundo a análise realizada na ferramenta Missense3d, a variante M52K danifica à estrutura proteica estudada promovendo a troca de um resíduo hidrofóbico interiorizado por um resíduo hidrofílico, apresentando um

alerta de “clash”. Clash é uma ferramenta do MolProbity, que no Missense3d sugere estrutura danificada quando apresenta um score superior a 30.

A estrutura de camundongo utilizada para fazer as predições foi obtida pelo banco de dados do AlphaFold por meio do Uniprot. Esta estrutura possui 343 aminoácidos e um peso de 37,5 KDa. A estrutura foi submetida à uma análise no PDBsum, que fornece informações como diagrama de estrutura secundária (Figura 31- A), mapa topológico (Figura 31 - B), e diagrama de Ramachandran (Figura 31 - C).

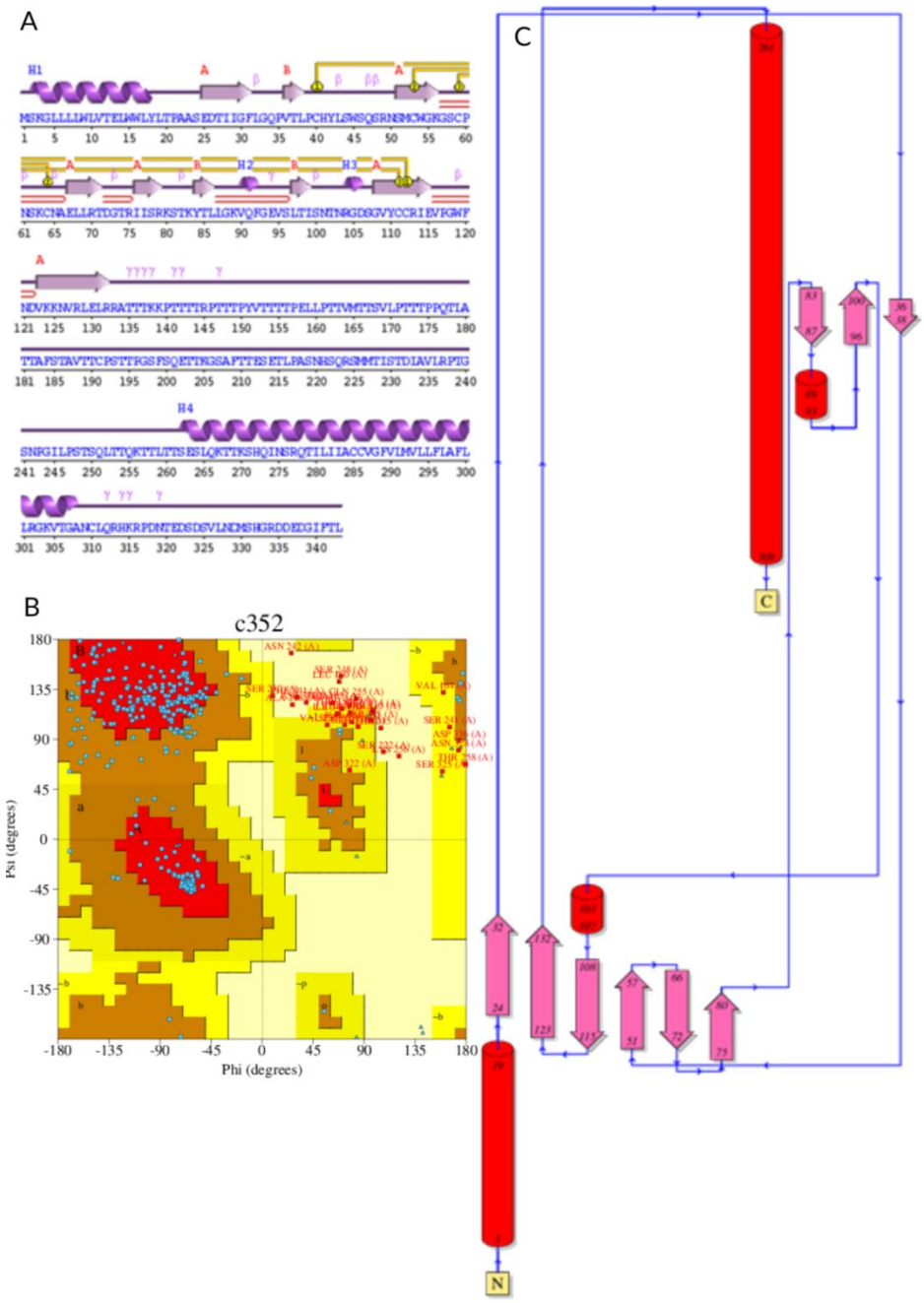

Figura 31 – Análises PDBsum Candidato 46.(A) .(A) Mapa de estrutura secundária, mostrando os motivos estruturais. As setas em lilás representam as fitas betas, as helicoidais em roxo representam as hélices, as ligações em amarelo representam as pontes dissulfeto e as linhas curvadas em vermelho os *hairpins*. (B) Diagrama de Ramachandran. O diagrama apresenta a conformação dos ângulos phi e psi para cada resíduo. Os resíduos são representados em azul claro, as regiões vermelhas são as mais favoráveis, seguidas pelas em marrom que são adicionalmente permitidas, as em amarelo são permitidas e em amarelo claro, regiões não permitidas. (C) Mapa topológico. Iniciando do N terminal para o C terminal, mostra as fitas em rosa e as hélices em vermelho.

A análise de estrutura secundária indica a presença de duas folhas beta (conjuntos de setas em rosa na Figura 31-C), composta por ”, nove fitas (setas

lilás na Figura 31-A e rosas na Figura 31-C) formando dois beta “bulges”,; quatro beta “hairpin” (seta em vermelho Figura 31-A),, quatro hélices (hélice roxa na Figura 31-A e cilindro vermelho na Figura 31-C), uma interação hélice-hélice, dez voltas beta, doze voltas gama e três ligações dissulfeto (Em amarelo na Figura 31-A). Já a análise do diagrama de Ramachandran, representada na Tabela 4 e na Figura 31-B, indica que a estrutura possui 69,4% de seus resíduos em regiões permitidas, 20,6% em regiões adicionalmente permitidas, 6% em regiões generosamente permitidas e, por fim, 4% em regiões não permitidas, indicando que em sua maioria ela é estável. Estes 4%, representam 12 resíduos (sem serem glicina ou prolina) da proteína, sendo que ao avaliar a confiabilidade de cada um desses resíduos foi verificado que estes apresentam baixa confiança, o que pode justificar a instabilidade e falta de definição.

Tabela 4 - Dados estatísticos do Diagrama de Ramachandran do candidato 46 predito pelo AlphaFold

	N°. Of residues	**%-tage**
Most favoured regions [A,B,L]	209	69,4%**
Additional allowed regions [a,b,l,p]	62	20,6%
Generously allowed regions [~a,~b,~l,~p]	18	6,0%
Disallowed regions [XX]	12	4,0%
Non-glycine and non-proline residues	301	100.0%
End-residues (excl. Gly and Pro)	2	
Glycine residues	20	
Proline residues	20	
Total number of residues	343	

4.2.5.3. Resultados das análises de domínio

A proteína codificada pelo gene de interesse é composta por um domínio N-terminal imunoglobulina V like, um domínio mucina, um domínio transmembrana e uma curta cauda citoplasmática na porção C-terminal [65]. A variante genética se encontra em sítio de domínio proteico N-terminal imunoglobulina V like, composto por 116 aminoácidos. Estudos de caracterização por cristalografia de raio-X do domínio em questão mostram que

nesta região há uma cavidade onde se liga à fosfatildilserina por meio de um íon metálico [66]. Este fato leva a crer que uma mutação em qualquer aminoácido desta região pode ocasionar má ligação da fosfatildilserina e um mau funcionamento da proteína de modo a afetar sua função.

4.2.5.4. Estudo das estruturas proteicas codificadas pelos genes escolhidos

Para seguir a análise de uma dinâmica de maior duração e inserir a variante, optamos pela estrutura do domínio com o fosfatildilserina (PSF) obtido experimentalmente por Santiago e coloboradores[67]. Tal escolha é justificada por dois fatores: primeiramente, devido à precisão do encaixe ligante-sítio de ligação obtido experimentalmente com resolução de 2,5Å; e em segundo lugar, a estrutura de domínio é mais estável quando comparada à estrutura inteira da proteína, que como mostrado no Diagrama de Ramachandran na Figura 31apresenta regiões fora da região permitida, indicando possível instabilidade ou indefinição de estrutura secundária Enquanto isso, o gráfico na Figura 32 juntamente à tabela indica uma melhor definição e estabilidade conformacional.

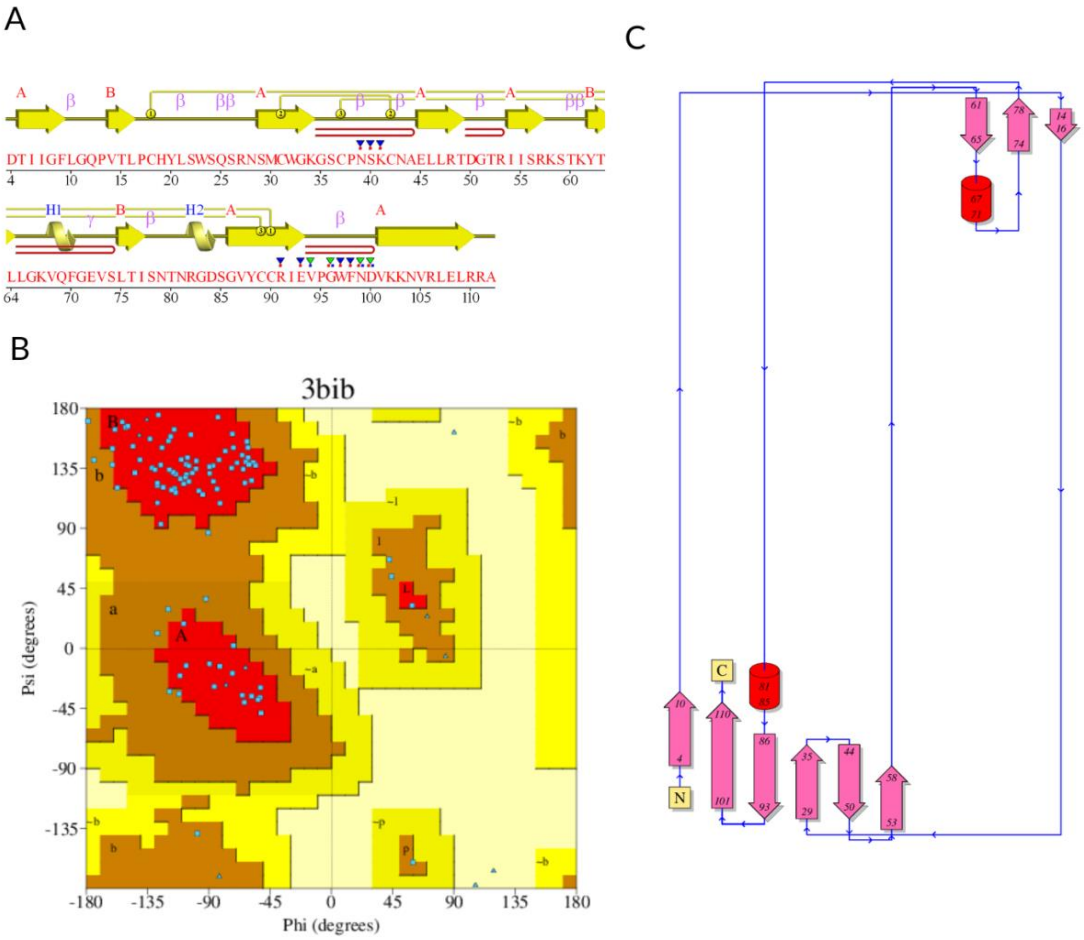

Figura 32 – Análises PDBsum da estrutura retirada do PDBe com código 3bib referente ao domínio IgV-like da *Candidato 46*.(A)) Mapa de estrutura secundária, mostrando os motivos estruturais. As setas em amarelo representam as fita beta, as helicoidais em amarelo representam as hélices, as ligações em amarelo representam as pontes dissulfeto e as linhas curvadas em vermelho os *hairpins*. (B) Diagrama de Ramachandran. O diagrama apresenta a conformação dos ângulos phi e psi para cada resíduo .Os resíduos são representados em azul claro, as regiões vermelhas são as mais favoráveis, seguidas pelas em marrom que são adicionalmente permitidas, as em amarelo são permitidas e em amarelo claro, regiões não permitidas. (C) Mapa topológico. Inicicando do N terminal para o C terminal, mostra as fitas em rosa e as hélices em vermelho.

A análise de estrutura secundária do domínio indica a presença de duas folhas beta, quatro beta “hairpin”(em vermelho na Figura 32 – A), dois beta “bulges” (em vermelho na Figura 32 – C), nove fitas ”(em amarelo na Figura 32 – A e em rosa na Figura 32 – C), duas hélices, uma volta gama e três ligações dissulfeto (Figura 32 – A e B). Já a análise do diagrama de Ramachandran representada na Tabela 5, indica que a estrutura possui 84,9% de seus resíduos em regiões permitidas e 15,1% em regiões adicionalmente permitidas, diferencialmente de quando analisamos a estrutura proteica inteira obtida do AlphaFold, onde encontramos apenas 69,4% dos resíduos em regiões

permitidas (Tabela 5) (Figura 32-C). Enquanto a primeira análise apresenta os motivos proteicos e ligações químicas que compõe o domínio, a segunda apresenta a localização destes aminoácidos no gráfico mediante sua conformação angular, de modo a indicar os domínios e a estabilidade da proteína. Pelo diagrama, juntamente à Tabela 5, vemos que o domínio apresenta estabilidade ao possuir todos seus resíduos em regiões permitidas.

Tabela 5 - Dados estatísticos do Diagrama de Ramachandran da estrutura do domínio IgV-like (3bib) da *Candidato 46* disponibilizado no PDBe.

	N°. Of residues	**%-tage**
Most favoured regions [A,B,L]	79	84,9%**
Additional allowed regions [a,b,l,p]	14	15,1%
Generously allowed regions [~a,~b,~l,~p]	0	0%
Disallowed regions [XX]	0	%
Non-glycine and non-proline residues	93	100.0%
End-residues (excl. Gly and Pro)	2	
Glycine residues	10	
Proline residues	4	
Total number of residues	109	

4.2.5.5. Simulação de Dinâmica Molecular para investigar efeito da variante no comportamento do domínio com o ligante

Como salientado anteriormente, a mutação M52K no Candidato 46 associada com a diminuição da frequência cardíaca no screening fenotípico em camundongo se encontra no domínio IgV-like da proteína. Casanovas e colaboradores apresentaram a estrutura resolvida deste domínio com o ligante PSF e o íon metálico Na^+ e esta foi utilizada na simulação de dinâmica molecular, como explicitado da seção de metodologia. Vale salientar que nesta estrutura, a mutação associada ao fenótipo se encontra no resíduo 30 do domínio, passando a ser denominada M30K.

Ao realizar a SDM longa para os domínios contendo a sequência selvagem (WT) e mutada (M30K), foi verificado por meio do vídeo gerado que a

ligação do ligante com o domínio é mais fraca na presença da mutação, fazendo com que o ligante permaneça menos tempo ligado ao domínio em comparação com a sequência selvagem. Para exemplificar o que é observado no vídeo, listamos os resíduos do domínio proteico que interagem diretamente com o ligante e calculamos a distância entre seus centros de massa (Figura 33 e Figura 34). No modelo do domínio selvagem, selecionamos quatro momentos para tal análise: o último passo da equilibração e os três primeiros nanosegundos da dinâmica de produção (Figura 33). Já para o modelo contendo a mutação M30K, utilizamos o último passo da equilibração e o primeiro nanosegundo da dinâmica de produção, uma vez que logo em seguida o ligante se desprende do domínio (Figura 34). Assim, calculamos a distância entre os seguintes resíduos e o ligante: Asp100, Trp97, Ser40, Phe98, Lys41 e por fim o resíduo afetado pela mutação, que apesar de não estar diretamente ligado ao ligante, é o objeto deste estudo.

Primeiramente, conseguimos observar que as interações entre a Asp100 e Ser40 e a PSF observada no modelo selvagem (Figura 33) não ocorrem da mesma forma na proteína contendo a mutação (Figura 34), tanto no passo da equilibração, quanto no primeiro passo da dinâmica de produção. Essas observações são confirmadas pelos gráficos de distância entre os centros de massa desses resíduos e a molécula de PSF (Figura 35 B e D), já que nos primeiros nanosegundos da dinâmica, ambos resíduos apresentam grande distância com o PSF no modelo contendo mutação (linha vermelha). Já os resíduos Lys41, Trp97 e Phe98, apesar de se ligarem da mesma forma a molécula de PSF no modelo selvagem e mutado (Figura 33e 34), estes apresentam maior distância com o ligante desde os primeiros nanosegundos da dinâmica no modelo mutado em comparação com o modelo selvagem (Figura 35C, E e F). Finalmente, apesar de não haver ligação direta do resíduo afetado pela mutação (M30K) com o ligante, podemos observar que sua distância com o ligante assume mesmo padrão que os demais resíduos, apresentando maior distância no modelo mutado quando comparado com o modelo selvagem nos primeiros nanosegundos da dinâmica (Figura 35A).

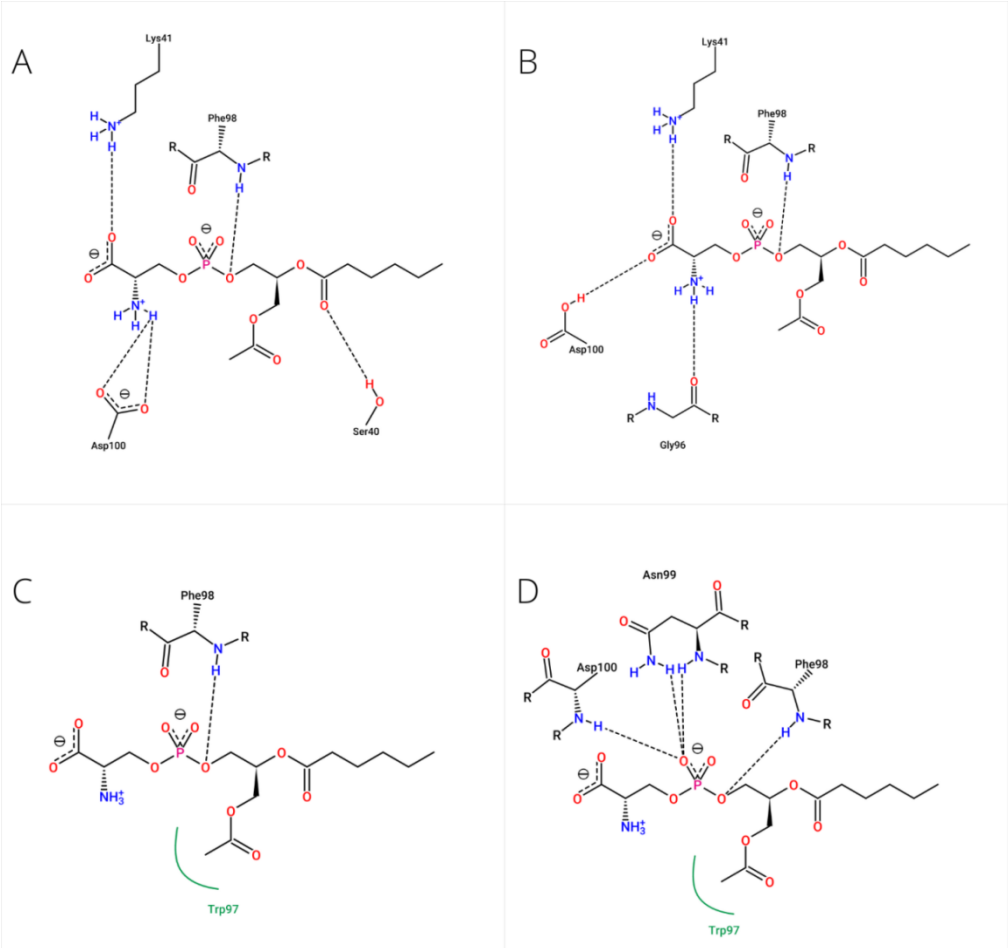

Figura 33 - Mapa de interação entre resíduos do domínio IgV-like da proteína *Candidato 46* e o PSF em sua forma selvagem. (A) Interação no final da equilibração. (B) Interação no primeiro nanosegundo da dinâmica de produção. (C) Interação no segundo nanosegundo da dinâmica de produção. (D) Interação no terceiro nanosegundo da dinâmica de produção.

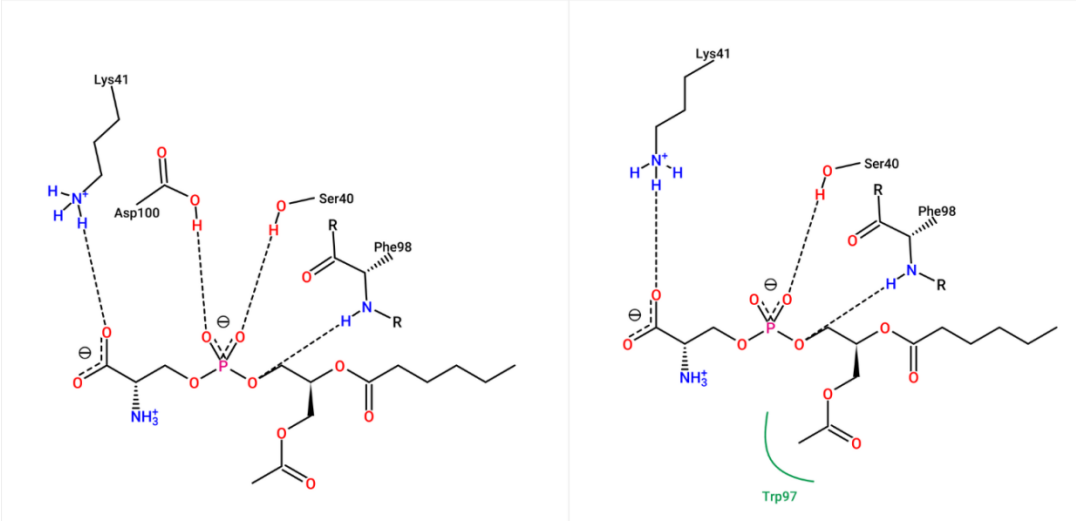

Figura 34 - Mapa de interação entre resídos do domínio da proteína *Candidato 46* em sua forma gerada pela variante. À direita é representada a interação no final da equilibração. À esquerda se encontra a Interação no primeiro nanosegundo da dinâmica de produção.

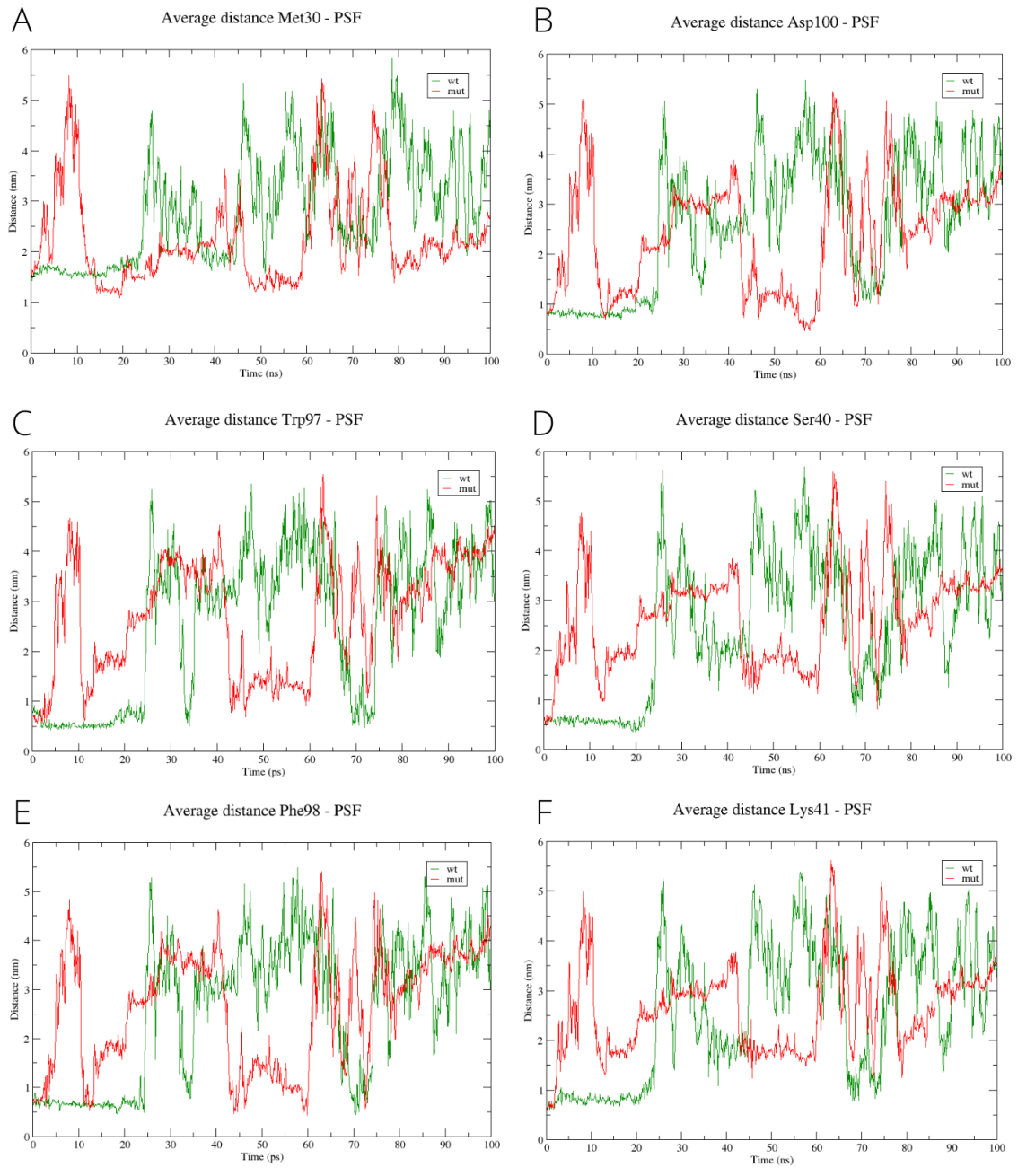

Figura 35 - Gráficos de distância entre os centros de massa dos dados resíduos e a molécula de PSF. (A) Gráfico mostrando a distância ao logo do tempo entre a metionina 30 e o PSF.(B) Gráfico mostrando a distância ao logo do tempo entre a asparagina 100 e o PSF.(C) Gráfico mostrando a distância ao logo do tempo entre o triptofano 97 e o PSF.(D) Gráfico mostrando a distância ao logo do tempo entre a serina 40 e o PSF.(E) Gráfico mostrando a distância ao logo do tempo entre a fenilalanina 98 e o PSF.(F) Gráfico mostrando a distância ao logo do tempo entre a lisina 41 e o PSF.

Dessa forma, as análises da interação dos diferentes resíduos e o ligante e a quantificação da distância entre os centros de massa dos resíduos e a molécula de PSF corroboram as observações visualizadas no vídeo. Sendo

assim, nos dados sugerem que a troca de uma metionina por uma lisina na posição 30 da proteína *Candidato 46* no domínio IgV-like afeta a ligação do domínio e o ligante PSF, permitindo uma ligação fraca que se mantém por pouco tempo.

4.2.5.6. Hipótese da contribuição do Candidato 46 *na alteração de FC*

O Candidato 46 é um gene pertencente à família de genes célula T imunoglobulina e mucina [65]. Esta família é composta por oito genes que se encontram no cromossomo 11B1.1 em camundongo, sendo quatro genes expressos e quatro genes preditos [65,68]. A proteína codificada pelo gene Candidato 46 é composta por 343 resíduos, tendo massa de aproximadamente 37,5 kDa, distribuídos em um domínio imunoglobulina N-terminal, um domínio mucina altamente glicosilado, uma região transmembrana e uma curta região citoplasmática [66].

Segundo o “THE HUMAN PROTEIN ATLAS”, a proteína estudada é expressa em testículo, placenta, apêndice, baço, linfonodos, amídalas e medula óssea. Na literatura é revelado ainda que a expressão se deve primariamente em células apresentadoras de antígenos (do inglês antigen-presenting cells), dentre elas células dendríticas, células B e, principalmente, macrófagos[68–70].

No coração, os macrófagos são as células do sistema imune mais abundante, correspondendo a aproximadamente 7% da população não-miócito em corações humanos saudáveis[71]. Atualmente, sabe-se que além dos macrófagos derivados de monócitos, existem populações de macrófagos residentes cardíacos e que parte destes se desenvolvem previamente a hematopoese. Os macrófagos residentes são definidos como macrófagos que exercem funções homeostáticas e que residem no tecido na ausência de injúria e inflamação e que podem surgir tanto linhagens embrionárias independentes de monócitos e hematopoese, como serem originárias de monócitos circulantes. No estado estacionário, os macrófagos residentes atuam nos processos de desenvolvimento coronariano, fagocitose de corpos apoptóticos, angiogenese, manutenção da homeostase, condução elétrica por meio de

conexina 43, ação anti inflamatória entre outras[72–76]. Quando ocorre um insulto cardíaco, como o estresse agudo do ventrículo direito, estudos indicam que os macrófagos cardíacos atuam na regulação dinâmica e contínua das junções gap evitando morte súbita[77]. Ao tratar-se de infarto do miocárdio, estas células imune atuam como terapêuticas no processo de reparo uma vez que os monócitos migram para a área afetada, diferenciando-se em macrófagos e atuando na retirada de células mortas, seja por necrose ou apoptose, além de secretar moléculas angiogênicas, entre outras[73,78,79].

Nos macrófagos cardíacos, a expressão da proteína *Candidato 46* se dá especificamente em um tipo de macrófago residente cardíaco que não expressa a proteína “C-C motif chemokine receptor 2 (CCR2)” e que pouco expressam a proteína “major histocompatibility complex II”[72]. Esta classe de macrófagos cardíacos residentes possui origem embrionária e se prolifera localmente, estando, aparentemente, envolvido com as funções de homeostase, fagocitose, desenvolvimento coronariano e linfoangiogenese [72]. É interessante notar que após o infarto, os macrófagos residentes que expressam o *Candidato 46* reduzem significativamente no tecido cardíaco, sendo substituídos por macrófagos derivados de monócitos circulantes que adquirem padrão de expressão gênica semelhantes, mas que não expressam o *Candidato 46*[26].

Homeostase é o nome dado ao processo auto regulatório que o corpo executa a fim de manter o sistema estável mediante alterações externas [80]. Um modo de manter a homeostase é através da remoção programada de células em um processo chamado de eferocitose. O processo de eferocitose ocorre por várias razões fisiológicas e patológicas, incluindo a remoção do excesso de células geradas durante o desenvolvimento, substituição de células velhas e eliminação de células danificadas e mortas, assim como de organelas com mau funcionamento do organismo[81]. Na maioria dos tecidos, esse processo é mediado por fagócitos professionais e não profissionais, sendo os macrófagos os fagócitos profissionais mais comuns, sendo capazes de ingerir rapidamente e processar vários corpos apoptóticos sucessivos[82].

Para permitir uma alta eficiência e impedir a eliminação de células saudáveis, o processo de eferocitose é rigidamente regulado e orquestrado por meio de vários programas de sinalização: sinalização de “encontre-me”, que proporcionam o recrutamento de fagócitos pelas células que estão morrendo ou que já estão mortas; a sinalização de “coma-me”, que medeia a captação de células apoptóticas e necróticas por receptores; e, finalmente, a sinalização da digestão, que permite o processamento do material celular pós fagocitose, geralmente por meio da degradação fagolisossomal. *O Candidato 46* é um dos receptores de eferocitose que identifica a fosfatidilserina (PSF), o sinal “coma-me” mais bem caracterizado.

Nish e colaboradores apontam que a fagocitose de corpos apoptóticos por macrófagos peritoneais ocorre de forma eficiente apenas na presença de duas proteínas: *Candidato 46* e MerTK, uma molécula “ponte”. Eles demonstraram que a ausência de *Candidato 46* promove uma redução na ligação e fagocitose de células apoptóticas, enquanto que a ausência de MerTK permite a ligação dos macrófagos às células apoptóticas, reduzindo sua capacidade de fagocitose. Dessa forma, eles sugerem que a fagocitose de células apoptóticas por macrófagos residentes peritoneais ocorrem em duas etapas: ligação do *Candidato 46* ao receptor PS, seguido da fosforilação do MerTK pelo *Candidato 46*, levando a fagocitose celular [83].

A realização errônea da eferocitose pode implicar em alterações de homeostase, processos inflamatórios e modulação do desenvolvimento cardíaco [76]. Recentemente, Ávila e colaboradores demonstraram que a depleção de macrófagos cardíacos ou a deficiência do receptor fagocítico MerTK em macrófagos cardíacos resulta na eliminação defeituosa de mitocôndria do tecido cardíaco, levando a ativação do inflamasoma, prejuízo da autofagia, acúmulo de mitocôndrias anômalas nos cardiomiócitos, alterações metabólicas e disfunção ventricular[67].

No processo de desenvolvimento cardíaco, há evidências de que uma falha na fagocitose pode levar ao bloqueio congênito cardíaco (CHB, do inglês congenital heart block). Clancy e colaboradores demonstraram que fetos portadores de CHB apresentam apoptose exagerada, 30 vezes maior, no

tecido septal de condução, conjuntamente com aumento de IgG[84]. Em estudo subsequente, eles avaliaram a contribuição de cardiomiócitos na remoção de células apoptóticas e observaram que IgG anti-SSA/Ro e anti-SSB/La produzidos pelas mães de crianças CHB bloqueiam a captação fagocitária dessas células, sugerindo que a não captação de células apoptóticas pelos cardiomiócitos promoveria sua eliminação por meio um processo pró-inflamatório de infiltração de macrófagos, que culminaria com a formação de fibrose e o bloqueio cardíaco permanente[8586].

Com base nas informações obtidas pelos resultados *in silico*, juntamente com a revisão bibliográfica, acreditamos que a troca de uma metionina por uma lisina na posição 52 da proteína, localizada no domínio IgV, causa uma alteração estrutural no sítio de ligação do íon Na^+ e, por consequência, na ligação com a molécula de fosfatildilserina. Esta alteração supostamente altera a função proteica ao desfavorecer a ligação com a proteína MertK, impedindo a sua fosforilação e a fagocitose dos corpos apoptóticos. Com uma função fagocitária ineficiente, haveria um processo inflamatório que conduziria à uma fibrose, afetando a condução elétrica e, consequentemente, a FC. Como este gene é expresso apenas em macrófagos residentes de origem embrionária, acreditamos que o dano tenha início no desenvolvimento, assim como observado em pacientes com CHB, podendo ter consequências após o nascimento ou promover remodelamento cardíaco e agravamento da condição até a fase adulta.

4.3. Genes dubiamente associados ao fenótipo

4.3.1. Análise de padrões conservativos

A partir da análise fenotípica realizada em camundongo, foram identificados 65 genes cuja associação da variante genética com a alteração de frequência cardíaca era dúbia, ou seja variantes genéticas em dois ou mais genes ocorreram em homozigose nos mesmos animais do mesmo pedigree, impedindo a identificação direta do gene relacionado ao fenótipo. Destes genes, 61 apresentaram variantes não sinônimas, sendo que, em 25 deles, essas variantes afetam aminoácidos que se localizam em região de domínio proteico (Figura 36).

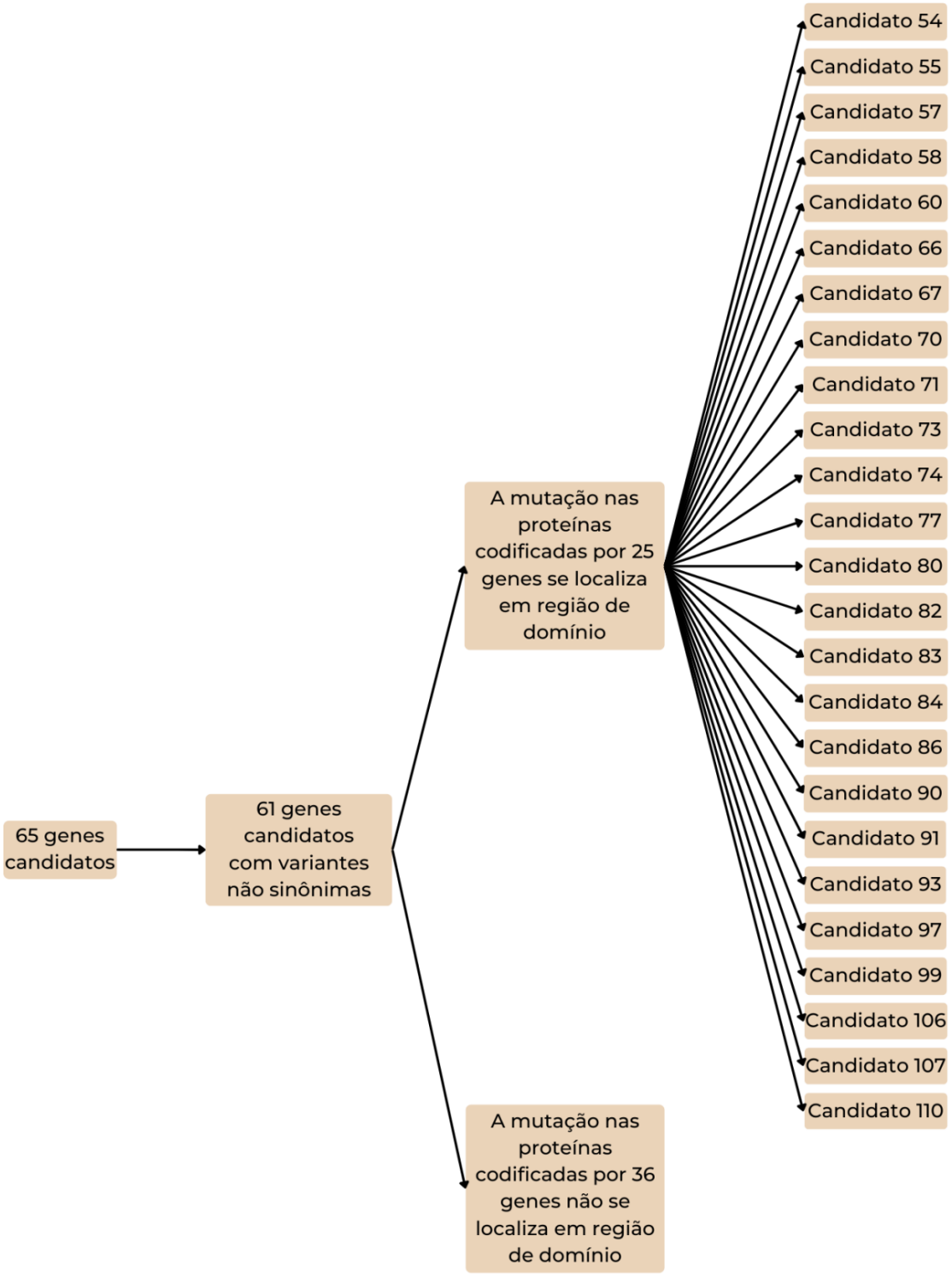

Figura 36 - Árvore de resultados da análise de conservação para *Mus musculus*. Dos 65 genes identificados no screening de camundongo, quatro apresentam variante intrônica enquanto 61 possuem variantes não sinônimas, das quais 25 afetam região de domínio proteico.

A análise seguiu para a etapa de busca de ortólogo em peixe paulistinha (Figura 37). Identificamos 44 genes ortólogos, dos quais 40 apresentaram sequência de cDNA correspondente em peixe paulistinha daquele onde se encontra a variante em camundongo. Nove genes apresentaram variante em códon que codifica aminoácido distinto do encontrado em camundongo, sendo que apenas um se localiza em sítio de domínio da proteína codificada. A respeito dos 29 genes que possuem alteração em códon relacionado ao mesmo aminoácido que na espécie referência, nove genes apresentam possível mutação em aa localizada em domínio proteico.

Figura 37 - Árvore de resultados da análise de conservação para *Danio rerio*. O esquema mostra a quantidade de genes obtidos em cada etapa da análise, e por fim a lista de genes que cumprem os requisitos estabelecidos.

Em termos comparativos com a espécie humana, 56 genes apresentaram ortólogos com variantes não sinônimas e 55 apresentam variante em sítio em que o cDNA se alinha ao equivalente cDNA em camundongo. Destes, nove genes possuem a variante localizada em códon que codifica aminoácido distinto do conhecido em camundongo, sendo que em três, o aminoácido afetado pela alteração genética está situado em região de domínio proteico. As demais 46 sequências proteicas codificadas pelos genes candidatos possuem aa igual ao identificado na espécie referência e, em 18 delas, o aa se localiza em domínio proteico (Figura 38).

Figura 38 - Árvore de resultados da análise de conservação para *Homo sapiens*. O esquema mostra a quantidade de genes obtidos em cada etapa da análise, e por fim a lista de genes que cumprem os requisitos estabelecidos.

Por fim, os resultados obtidos nas análises das três espécies foram comparados e observamos que seis genes apresentaram as características buscadas nas três espécies, sendo todos relacionados à redução da frequência

cardíaca (Candidato 55, Candidato 67, Candidato 73, Candidato 82, Candidato 91 e Candidato 99) (Figura 39). Também é possível observar que outros nove genes apresentaram as características buscadas tanto em camundongo quanto na espécie humana. Destes, um gene é associado ao aumento de frequência cardíaca (Candidato 74) e oito genes são associados à redução do fenótipo.

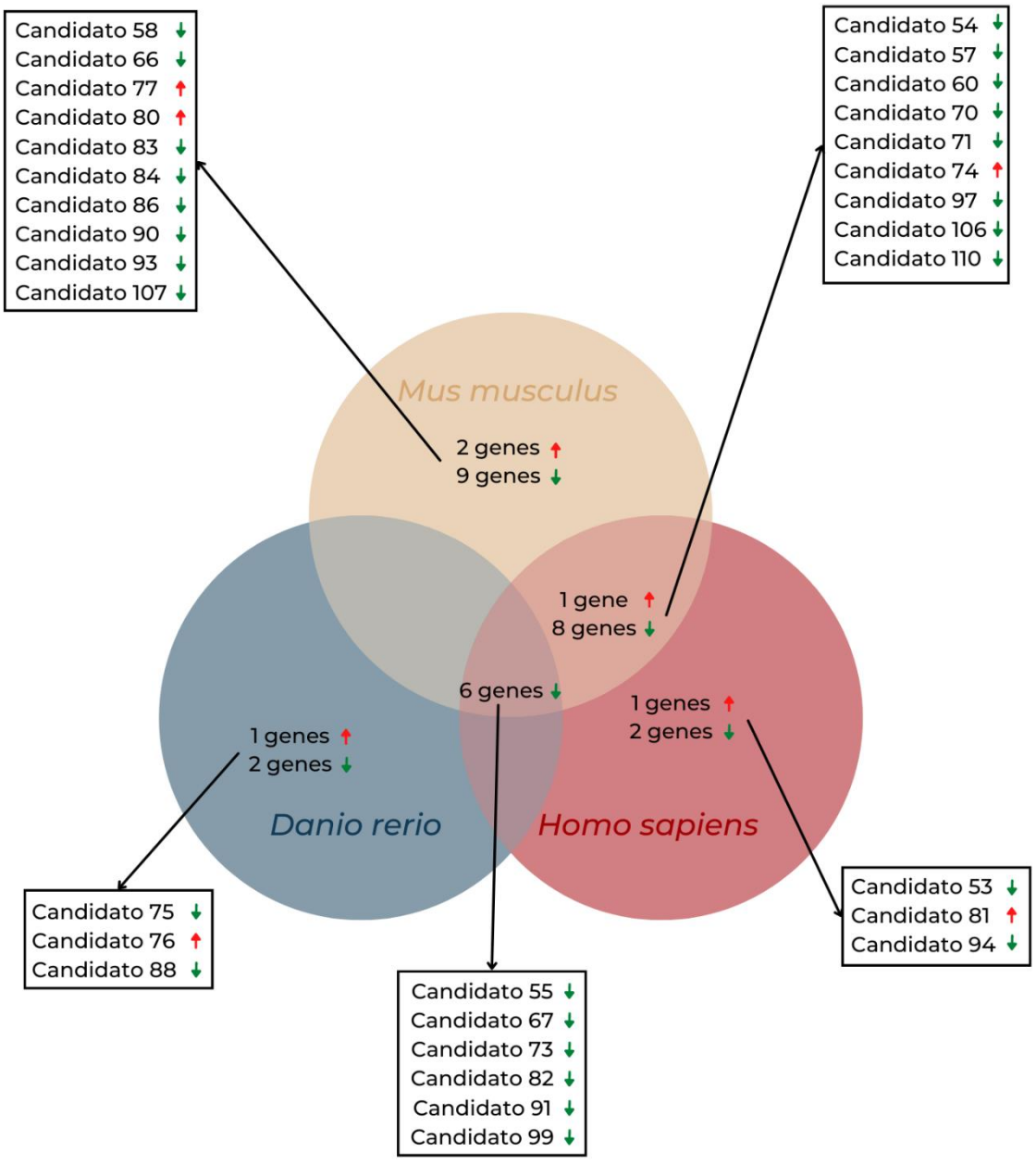

Figura 39 - Diagrama de Venn para a análise de conservação. Os genes representados são aqueles que se alinham com a variante em camundongo, possuem mesmo aminoácido entre as três espécies e estão em região de domínio proteico. Em cada área do círculo estão os genes que cumprem os requisitos e nas intersecções estão os que apresentam domínios nas espécies em questão.

4.3.2. Análise de predição de dano à estrutura proteica

A análise para prever se a variante afeta a estrutura proteica codificada pelo gene nas espécies estudadas foi realizada conforme descrito na metodologia. Das 61 proteínas codificadas pelos genes candidatos identificados em camundongo com variantes não sinônimas, duas não foram avaliadas por tratar-se de proteínas cuja variante causa um códon de parada, não permitindo a análise pelo programa utilizado (Figura 40). Das 59 proteínas restantes, seis apresentam estrutura depositada no banco de dados do PDBe, mas nenhuma estrutura cumpriu os quesitos para serem submetidas ao Missense3d estabelecidos na seção metodológica. Das demais proteínas, 47 tiveram estruturas preditas disponíveis no banco de dados do AlphaFold, sendo que 35 cumpriram o requisito para serem submetidas ao preditor de dano estrutural e 9 apresentaram provável dano estrutural ocasionado pela alteração missense obtida na análise fenotípica. Por fim, as sequências de aminoácidos de 23 proteínas foram submetidas ao Phyre², sendo que seis cumpriram os critérios para seguir a etapa seguinte e três apresentaram dano estrutural. Das proteínas que apresentam variantes missense 18 não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na mesma.

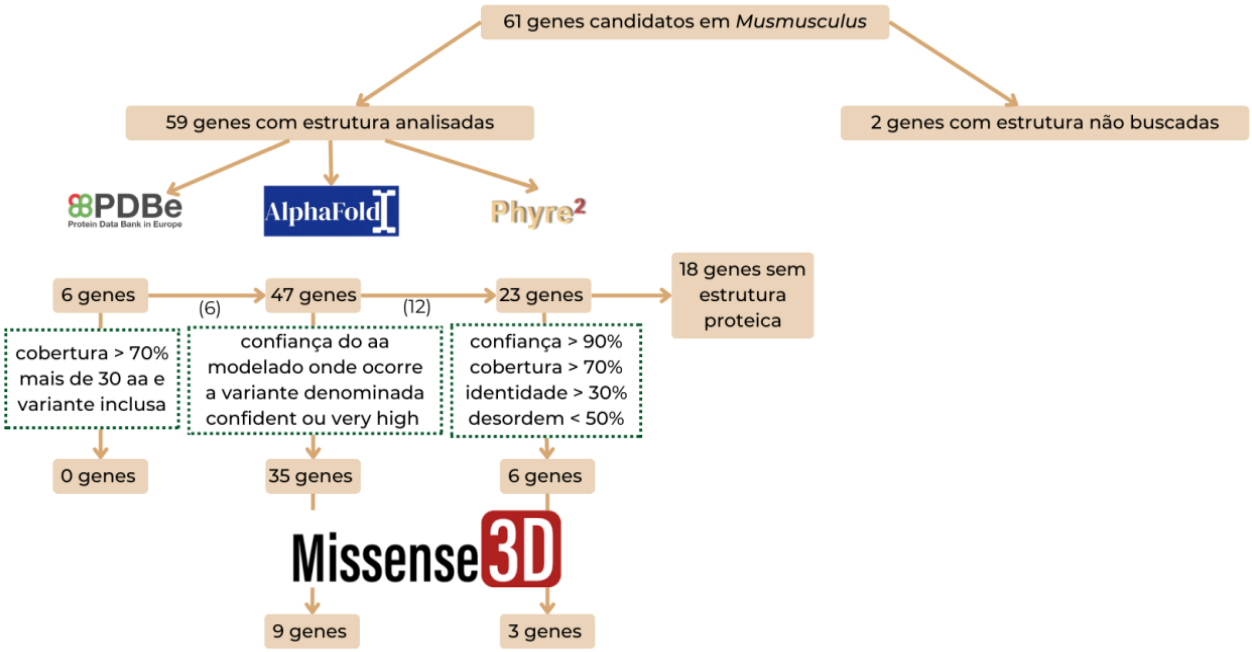

Figura 40 - Árvore de resultados da análise de predição de dano estrutural para os genes em camundongo. O esquema mostra o número de estruturas relacionadas aos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína associada.

Dos 40 genes que tem região com variante alinhada entre peixe paulistinha e camundongo, 38 apresentaram variantes missense identificados em peixe paulistinha, 31 possuem pelo menos um ortólogo cuja variante genética identificada não sinônima em camundongo codifica uma proteína que possui estrutura da proteína codificada predita disponível no banco de dados do AlphaFold (Figura 41). Destas, 23 apresentaram os critérios necessários e foram submetidas ao Missense3d para verificar o possível dano à estrutura causado pela variante, sendo que oito apresentaram dano estrutural. Das proteínas que apresentam variantes missense sete não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na mesma.

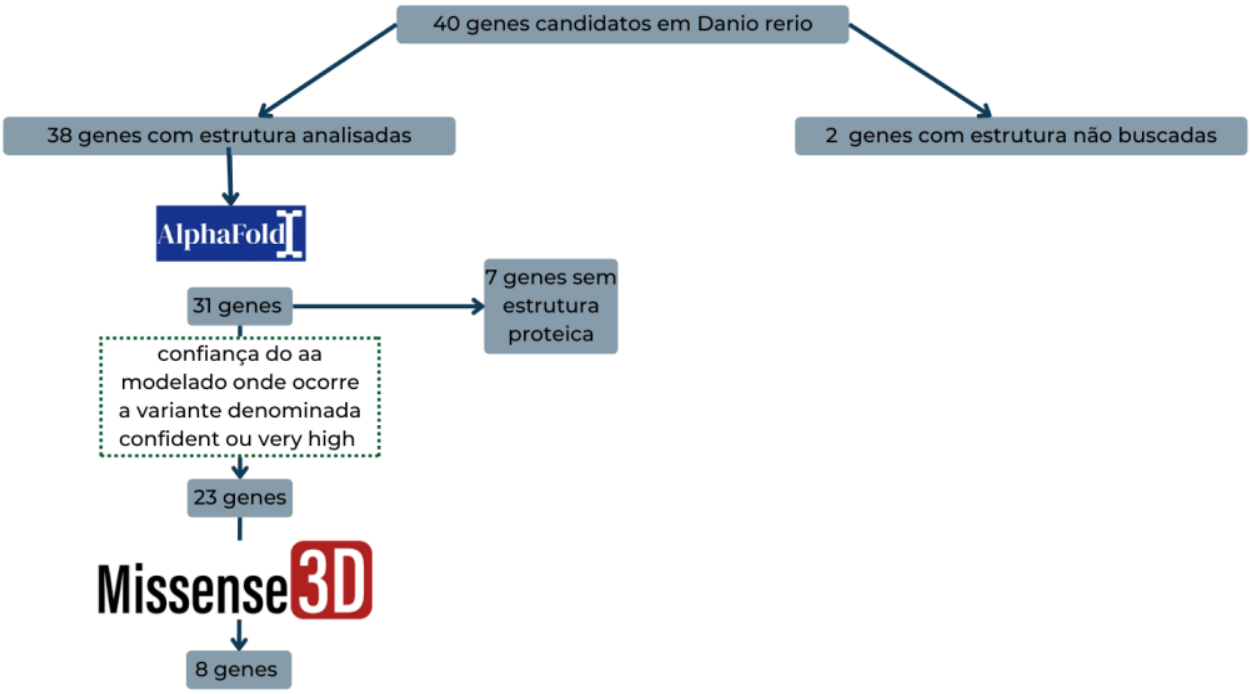

Figura 41 - Árvore de resultados da análise de predição de dano estrutural para os genes em peixe paulistinha. O esquema mostra o número de estruturas relacionadas aos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína associada.

A mesma análise foi realizada para proteínas em humanos (Figura 42). Das 53 proteínas com variantes missense avaliadas, 18 apresentaram estrutura obtida experimentalmente depositadas no PDBe, sendo que quatro cumpriram os requisitos para serem submetidas ao Missense3d, obtendo como resultado uma proteína que apresenta possível dano estrutural quando sujeita à variante. Das demais proteínas, 48 apresentaram estruturas preditas disponíveis no banco de dados do AlphaFold, sendo que 31 cumpriram o requisito para serem submetidas ao preditor de dano estrutural, retornando oito proteínas com possível dano estrutural causado pela alteração obtida no screening fenotípico. Por fim, as sequências proteicas de 18 sequências foram submetidas ao Phyre² sendo que quatros dos modelos gerados apresentaram os requisitos para seguir para a análise no Missense3d e uma proteína apresentou dano estrutural. Das proteínas que apresentam variantes missense

15 não foram analisadas por não haver estrutura com parâmetros adequados para prever o impacto da variante na mesma.

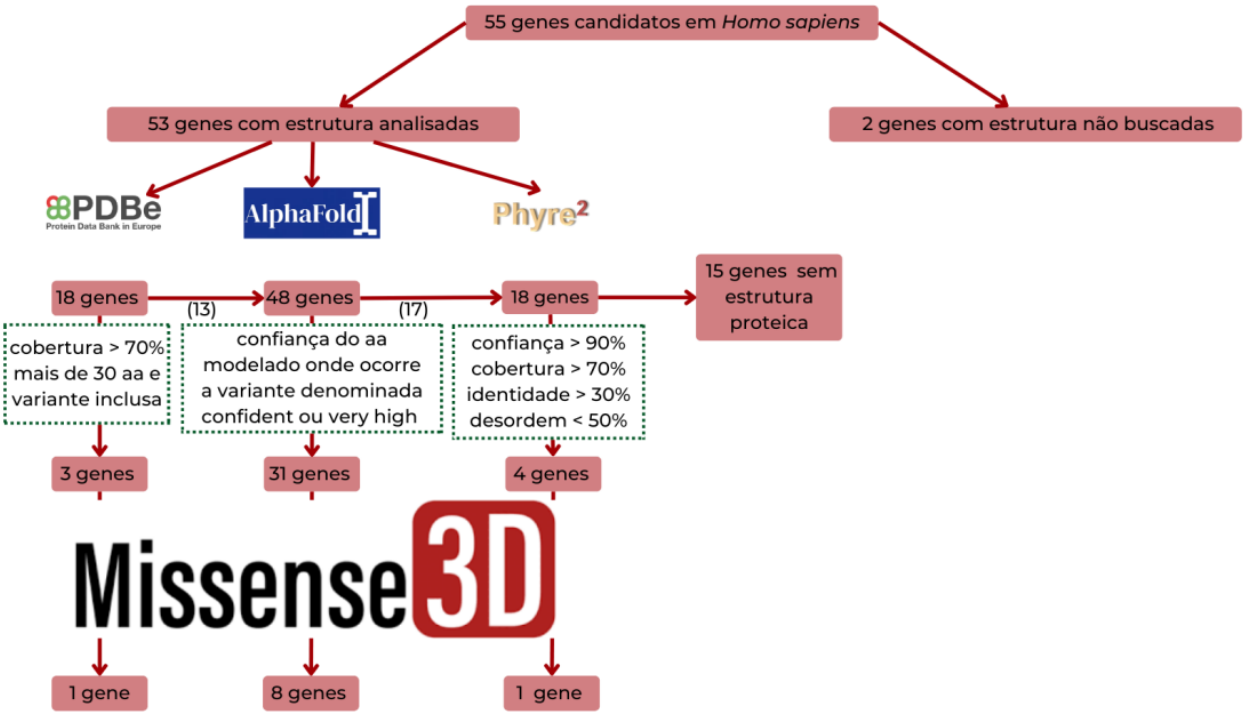

Figura 42 - Árvore de resultados da análise de predição de dano estrutural para os genes em humano. O esquema mostra o número de estruturas relacionadas aos genes a cada etapa com detalhe para os genes em que a variante ocasiona dano estrutural na proteína associada.

Finalmente, a análise comparativa entre as espécies foi realizada (Figura 43) de modo que as estruturas proteicas com alterações missense identificadas nos camundongos codificadas pelos genes Candidato 61, Candidato 79, Candidato 91 e Candidato 99 mostraram potencial dano estrutural em todas as espécies analisadas. As estruturas relacionadas aos genes Candidato 81, Candidato 106 e Candidato 11 mostraram dano em camundongo e humanos, enquanto o Candidato 2 mostrou-se comum em camundongo e peixe paulistinha. Também foram identificadas estruturas proteicas com dano pela variante em apenas uma espécie sendo: três para camundongo, duas para peixe paulistinha e três em humano.

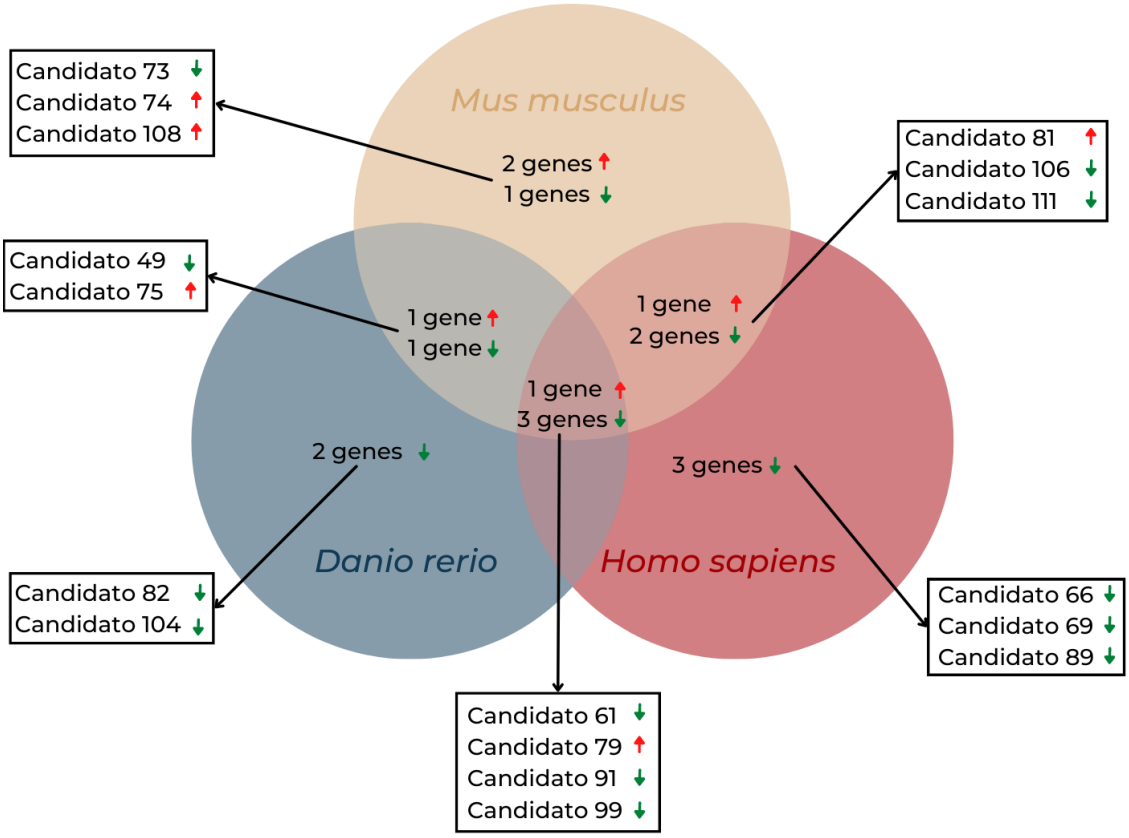

Figura 43 - Diagrama de Venn de apresentação de resultados de dano estrutural. Temos no diagrama os genes cuja proteína apresenta dano estrutural ocasionado pela variante por espécie. As setas em verde mostram a diminuição de frequência cardíaca enquanto as vermelhas referem-se ao aumento da frequência.

4.3.3. Comparação entre análises

Quando comparadas as análises de padrões de conservação e de predição de dano à estrutura da proteína codificada pelos genes candidatos, observamos que o Candidato 91 e o Candidato 99 apresentam maiores evidências biológicas de afetar funcional e estruturalmente a proteína, já que em ambos a variante afeta o mesmo aa situado em região de domínio, além da mesma ser predita por ocasionar alteração na estrutura proteica nas três espécies. Da mesma forma, o Candidato 106 apresentou as características mencionadas acima, mas apenas nas espécies *Homo sapiens* e *Mus musculus* (Figura 44).

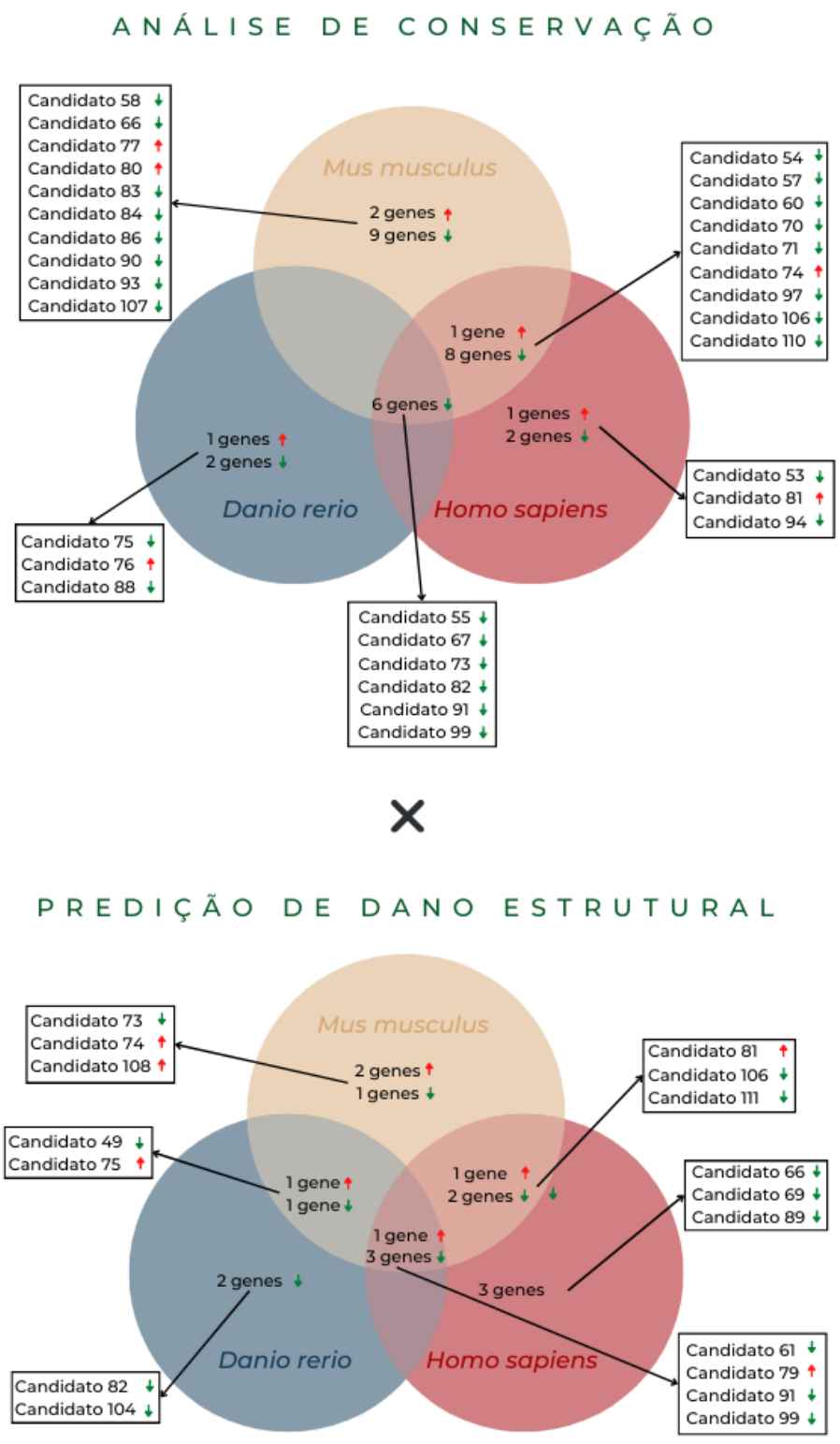

Figura 44 - Esquema comparativo dos diagramas das metodologias utilizadas.

Levando-se em consideração os achados considerando os genes dúbios, podemos concluir que a utilização das análises *in silico* conseguiram priorizar genes a serem futuramente estudados. A seguir, os três candidatos que apresentaram maior evidência biológica de afetar funcional e estruturalmente as proteínas terão o fenótipo observado no camundongo descrito em detalhe.

4.3.4. Candidato 91

O gene Candidato 91 codifica uma proteína pertencente à família pro proteína convertase do tipo subtilisina [87]. O gene é expresso em maiores níveis no sistema neuroendócrino, como hipotálamo e pancreas[87,88]. Ele foi identificado pela análise fenotípica por associar a mutação I584N e a redução de 12% na frequência cardíaca em animais homozigoto mutado quando comparado aos camundongos homozigotos selvagem pertencentes à família R6567 (Tabela 6 Figura 45).

Tabela 6 - Apresentação das diferenças fenotípicas entre homozigoto selvagem (homo ref), heterozigoto (het) e homozigoto mutado para mutação no Candidato 91 no pedigree R6567.

Screen Name	Rec	Ref Mean	Ref SD	Het Mean	Het SD	% vs homo ref	dif vs homo ref	Var Mean	Var SD	% vs homo ref	dif vs homo ref
HR Average	5.37e-05	692.2	26.4	683	42.7	-1.3	-9.20	609.1	53.9	-12	-83.10

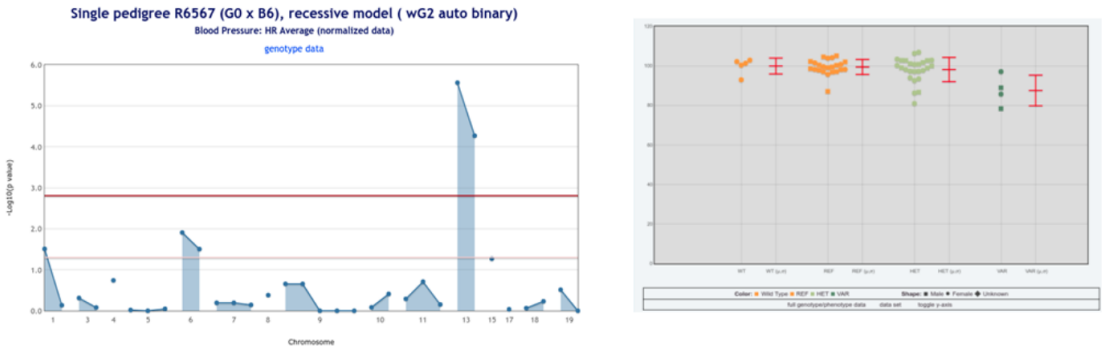

Figura 45 - Manhattan Plot de variantes identificadas no pedigree R7750 (esquerda) e gráficos de FC (direita) de dados da média dos dois dias. Em detalhe, no Manhattan plot, no eixo x a posição cromossômica em que se encontram as variantes ocasionadas pelo ENU e testadas quanto sua associação no fenótipo de FC e no eixo y é possível ver o valor de –log10(p valor), mostrando o impacto da variante no fenótipo. O ponto acima da linha em vermelho mostra que a variante no cromossomo 11 está associada à redução de frequência cardíaca. Nos gráficos de relação fenótipo x genótipo, são mostrados da esquerda para direita, animais controle C57(WT - laranja), homozigoto selvagem (REF - laranja), heterozigoto (HET - verde claro), e homozigoto mutado (VAR - verde escuro).

4.3.5. Candidato 99

O gene Candidato 99 codifica uma proteína que compõe junto a outras o proteossomo 26S [89]. Este complexo atua na degradação de proteínas ubiquitinadas, além de estar envolvido no processo de reparo de DNA, diferenciação de células tronco embrionárias, proliferação e estabilidade proteica, entre outras [89]. Ele foi identificado pelo screening fenotípico por demonstrar uma associação entre a mutação V53A e uma redução de 18,7% na frequência cardíaca em animais homozigotos mutado quando comparado aos camundongos homozigotos selvagens pertencentes à família R7271 (Tabela 7 Figura 46).

Tabela 7 - Apresentação das diferenças fenotípicas entre homozigoto selvagem (homo ref), heterozigoto (het) e homozigoto mutado para mutação no Candidato 99 no pedigree R7271.

Screen Name	Add	Ref Mean	Ref SD	Het Mean	Het SD	% vs homo ref	dif vs homo ref	Var Mean	Var SD	% vs homo ref	dif vs homo ref
HR Average	1.22e-04	660	62.7	601.6	71.9	-8.9	-58.40	536.8	31.3	-18.7	-123.20

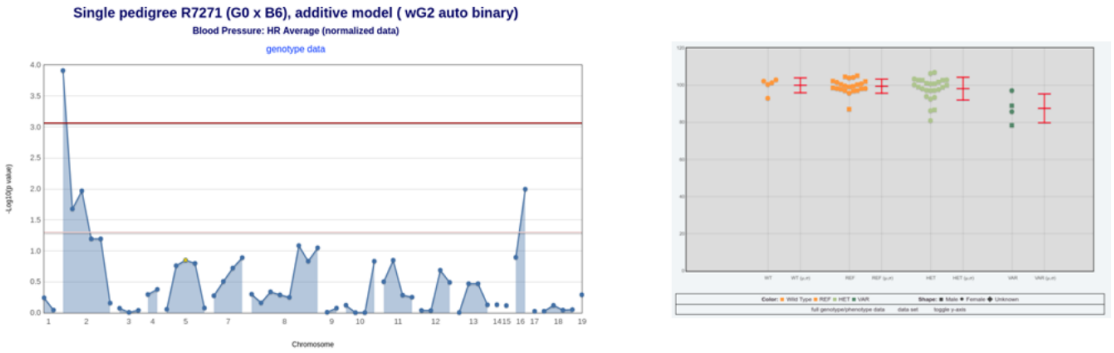

Figura 46- Manhattan Plot de variantes identificadas no pedigree R7271 (esquerda) e gráficos de FC (direita) de dados da média dos dois dias. Em detalhe, no Manhattan plot, no eixo x a posição cromossômica em que se encontram as variantes ocasionadas pelo ENU e testadas quanto sua associação no fenótipo de FC e no eixo y é possível ver o valor de –log10(p valor), mostrando o impacto da variante no fenótipo. O ponto acima da linha em vermelho mostra que a variante no cromossomo 11 está associada à redução de frequência cardíaca. Nos gráficos de relação fenótipo x genótipo, são mostrados da esquerda para direita, animais controle C57(WT - laranja), homozigoto selvagem (REF - laranja), heterozigoto (HET - verde claro), e homozigoto mutado (VAR - verde escuro).

4.3.6. Candidato 106

O gene Candidato 106 codifica uma proteína que pertence a família TRIM (do inglês “Tripartite motif” – “motivo tripartido” em português). As proteínas pertencentes à esta família comumente estão associadas com actividade ligase ubiquitina E3 [90]. A proteína codificada pelo gene em questão é envolvida em processos como a regulação de resposta imune, ciclo celular, apoptose, estabilização de p53, entre outros[90]. O Candidato 106 foi identificado pelo screening fenotípico ao apresentar uma associação entre a mutação V309E e uma redução de 16,8% na frequência cardíaca em animais homozigotos mutado quando comparado aos camundongos homozigoto selvagem pertencentes à família R7563 (Figura 47 - Tabela 8)

Tabela 8 - Apresentação das diferenças fenotípicas entre homozigoto selvagem (homo ref), heterozigoto (het) e homozigoto mutado para mutação no Candidato 91 no pedigree R7563.

Screen Name	Add	Ref Mean	Ref SD	Het Mean	Het SD	% vs homo ref	dif vs homo ref	Var Mean	Var SD	% vs homo ref	dif vs homo ref
HR Average	6.18e-04	651.3	57.7	615.3	59.5	-5.5	-36.00	541.9	35.8	-16.8	-109.40

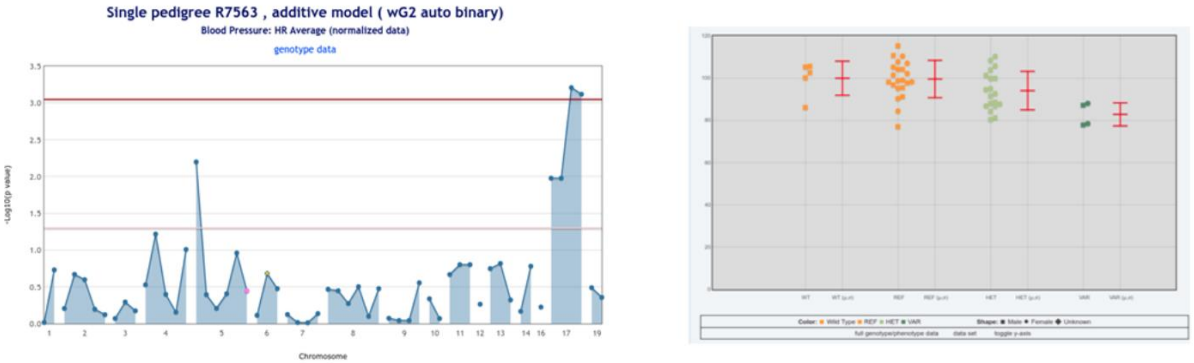

Figura 47 - Manhattan Plot de variantes identificadas no pedigree R7563 (esquerda) e gráficos de FC (direita) de dados da média dos dois dias. Em detalhe, no Manhattan plot, no eixo x a posição cromossômica em que se encontram as variantes ocasionadas pelo ENU e testadas quanto sua associação no fenótipo de FC e no eixo y é possível ver o valor de –log10(p valor), mostrando o impacto da variante no fenótipo. O ponto acima da linha em vermelho mostra que a variante no cromossomo 11 está associada à redução de frequência cardíaca. Nos gráficos de relação fenótipo x genótipo, são mostrados da esquerda para direita, animais controle C57(WT - laranja), homozigoto selvagem (REF - laranja), heterozigoto (HET - verde claro), e homozigoto mutado (VAR - verde escuro).

5. CONCLUSÃO

A estratégia proposta para priorizar os genes candidatos identificados pelo mapeamento fenotípico apresentou resultados promissores. Primeiramente, considerando os resultados obtidos na análise dos genes conhecidos (controles positivos), salientamos que 59% dos genes apresentam região afetada pelas variantes genéticas associadas a alteração da frequência cardíaca conservada por pelo menos duas das três espécies avaliadas. O mesmo resultado não foi observado na predição de dano estrutural devido a limitação da disponibilidade das estruturas obtidas até o fechamento dos resultados desta dissertação. No entanto, esse fato, como apresentado na introdução, foi recentemente resolvido com a publicação de mais de 200 milhões de estruturas preditas de proteínas humanas e de mais 47 organismos. Dessa forma, acreditamos que a reavaliação de estruturas proteicas disponíveis para as proteínas codificadas pelos genes conhecidos para as três espécies avaliadas permitirá a identificação de mais genes cujo dano estrutural está presente.

Em segundo lugar, as análises conduzidas nos genes diretamente associados ao fenótipo selecionaram dois candidatos, Candidato 35 e Candidato 46, que apresentaram evidências robustas para o seu envolvimento em mecanismos que influenciam a frequência cardíaca, além de as simulações de dinâmica molecular demonstrarem resultados interessantes de como as alterações de aminoácidos acarretariam mudanças conformacionais na proteína e, consequentemente, em sua função.

Finalmente, a partir das análises *in sílico* abordadas nesse estudo, selecionamos 3 genes candidatos dentre os genes dúbios que serão priorizados quanto ao estudo de sua contribuição no fenótipo. Os genes dúbios fazem parte de um grupo de genes cuja variante genética foi associada a alterações na frequência cardíaca juntamente com outros genes, impedindo a pronta identificação de qual alteração estava diretamente relacionada ao fenótipo durante a análise sistemática deste fenótipo em camundongo. Os resultados aqui obtidos sugerem que as análises computacionais são úteis na priorização de candidatos que deverão ser validados por apresentarem maiores

chances de comprometimento funcional da proteína candidata, consequentemente, e do fenótipo de interesse.

Dessa forma, em conjunto, os resultados obtidos sugerem que as análise computacionais propostas permitem 1. Priorizar genes candidatos para validação e esclarecimento quanto sua contribuição no fenótipo de frequência cardíaca, inclusive entre aqueles cuja associação com o fenótipo não está clara, e 2. Apresentar possíveis mecanismos pelos quais as alterações genéticas afetam a função proteica e, consequentemente, o fenótipo, direcionando futuros experimentos.

Se validada, esta abordagem será fundamental para aumentar a eficiência da validação *in vivo* dos genes candidatos para RHR.

6. PRÓXIMOS PASSOS

Devido a atualizações recentes no banco de dados AlphaFold DB, sugerimos, como próximos passos, avaliar a contribuição das alterações genéticas em proteínas com estrutura predita recém depositada. Além disso, como apresentado na introdução, novos modelos de predição de estrutura proteica baseados em modelos de linguagem de proteína parecem ser mais acurados na avaliação do efeito de troca de aminoácidos na estrutura proteica[31]. Dessa forma, gostaríamos de avaliar o desempenho dessas novas metodologias em comparação da aqui empregada, utilizando como base as diversas alterações genéticas testadas na metodologia de forward genetic screening. Vale salientar que dispomos de informações de diversas alterações genéticas nos nossos genes candidatos que não foram associadas à alteração no fenótipo de frequência cardíaca. Esse corpo de resultado oferece uma oportunidade única de testar tais algoritmos, já que tem o dado fenotípico associado.

Por fim, as abordagens aqui empregadas selecionaram 5 candidatos, dois deles com evidencias robustas provenientes de SDM, que deverão ser investigados com cuidado experimentalmente, a partir do desenvolvimento e da avaliação de animai knockout para cada um dos genes selecionados.

7. REFERÊNCIAS

1. Palatini P. Elevated resting heart rate is a risk factor for sudden death in middle-aged men. Evid Based Cardiovasc Med. 2001;5(3):86.

2. Greenland P, Daviglus ML, Dyer AR, Liu K, Huang CF, Goldberger JJ, et al. Resting heart rate is a risk factor for cardiovascular and noncardiovascular mortality: The Chicago Heart Association Detection Project in Industry. Am J Epidemiol. 1999;149(9):853–62.

3. Fox K, Borer JS, Camm AJ, Danchin N, Ferrari R, Lopez Sendon JL, et al. Resting Heart Rate in Cardiovascular Disease. J Am Coll Cardiol. 2007;50(9):823–30.

4. Jouven X, Empana JP, Schwartz PJ, Desnos M, Courbon D, Ducimetière P. Heart-Rate Profile during Exercise as a Predictor of Sudden Death. N Engl J Med [Internet]. 2005 May 12;352(19):1951–8. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMoa043012

5. Singh JP, Larson MG, O'Donnell CJ, Tsuji H, Evans JC, Levy D. Heritability of heart rate variability: The Framingham Heart Study. Circulation. 1999;99(17):2251–4.

6. Martin LJ, Comuzzie AG, Sonnenberg GE, Myklebust J, James R, Marks J, et al. Major Quantitative Trait Locus for Resting Heart Rate Maps to a Region on Chromosome 4. Hypertension. 2004;43(5):1146–51.

7. Russell MW, Law I, Sholinsky P, Fabsitz RR. Heritability of ECG measurements in adult male twins. J Electrocardiol. 1998;30(SUPPL.):64–8.

8. Dalageorgou C, Ge D, Jamshidi Y, Nolte IM, Riese H, Savelieva I, et al. Heritability of QT interval: How much is explained by genes for resting heart rate? J Cardiovasc Electrophysiol. 2008;19(4):386–91.

9. Eijgelsheim M, Newton-Cheh C, Sotoodehnia N, de bakker PIW, Müller M, Morrison AC, et al. Genome-wide association analysis identifies multiple loci related to resting heart rate. Hum Mol Genet. 2010;19(19):3885–94.

10. Eppinga RN, Hagemeijer Y, Burgess S, Hinds DA, Stefansson K, Gudbjartsson DF, et al. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat Genet [Internet]. 2016 Dec 31;48(12):1557–63. Available from: http://www.nature.com/articles/ng.3708

11. Amsterdam A, Burgess S, Golling G, Chen W, Sun Z, Townsend K, et al. A large-scale insertional mutagenesis screen in zebrafish. Genes Dev [Internet]. 1999 Oct 15;13(20):2713–24. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.13.20.2713

12. Simon MM, Moresco EMY, Bull KR, Kumar S, Mallon AM, Beutler B, et al. Current strategies for mutation detection in phenotype-driven screens

utilising next generation sequencing. Mamm Genome. 2015;26(9–10):486–500.

13. Wang T, Zhan X, Bua CH, Lyona S, Pratta D, Hildebrand S, et al. Real-time resolution of point mutations that cause phenovariance in mice. Proc Natl Acad Sci U S A. 2015;112(5):E440–9.

14. Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Biol [Internet]. 1970 Jun 1;19(2):99–113. Available from: https://doi.org/10.2307/2412448

15. Sivashankari S, Shanmughavel P. Comparative genomics - A perspective. Bioinformation. 2007;1(9):376–8.

16. Bertram L, Tanzi RE. Genome-wide association studies in Alzheimer's disease. Hum Mol Genet. 2009;18(R2):137–45.

17. Schmidt EE, Davies CJ. The origins of polypeptide domains. BioEssays. 2007;29(3):262–70.

18. Ortiz FW, Sergeev Y V. Global computational mutagenesis of domain structures associated with inherited eye disease. Sci Rep [Internet]. 2019;9(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-019-39905-9

19. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? Curr Opin Struct Biol. 2007;17(3):342–6.

20. Gromiha MM, Nagarajan R, Selvaraj S. Protein structural bioinformatics: An overview. Encycl Bioinforma Comput Biol ABC Bioinforma. 2018;1–3:445–59.

21. Agnihotry S, Pathak RK, Singh DB, Tiwari A, Hussain I. Protein structure prediction. In: Bioinformatics [Internet]. Elsevier; 2022. p. 177–88. Available from: http://dx.doi.org/10.1016/B978-0-323-89775-4.00023-7

22. Sanjeevi M, Hebbar PN, Aiswarya N, Rashmi S, Rahul CN, Mohan A, et al. Methods and applications of machine learning in structure-based drug discovery [Internet]. Advances in Protein Molecular and Structural Biology Methods. Elsevier Inc.; 2022. 405–437 p. Available from: https://doi.org/10.1016/B978-0-323-90264-9.00025-8

23. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol. 1997;270(3):471–80.

24. Yuan X, Shao Y, Bystroff C. Ab initio protein structure prediction using pathway models. Comp Funct Genomics. 2003;4(4):397–401.

25. Hardin C, Pogorelov T V, Luthey-Schulten Z. Ab initio protein structure prediction. Curr Opin Struct Biol [Internet]. 2002 Apr;12(2):176–81. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959440X02003068

26. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al.

Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.

27. Marcu ŞB, Tăbîrcă S, Tangney M. An Overview of Alphafold's Breakthrough. Front Artif Intell. 2022;5(June):1–6.

28. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439–44.

29. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv [Internet]. 2022;2022.07.20.500902. Available from: https://www.biorxiv.org/content/10.1101/2022.07.20.500902v2%0Ahttps://www.biorxiv.org/content/10.1101/2022.07.20.500902v2.abstract

30. Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. Structure [Internet]. 2022;30(8):1169-1177.e4. Available from: https://doi.org/10.1016/j.str.2022.05.001

31. Weissenow K, Heinzinger M, Steinegger M, Rost B. Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. 2022;1–16. Available from: https://doi.org/10.1101/2022.11.14.516473

32. Karplus M, Lavery R. Significance of Molecular Dynamics Simulations for Life Sciences. 2013;1–11.

33. Phys JC, Gur M, Blackburn EA, Ning J, Narayan V, Ball KL, et al. Molecular dynamics simulations of site point mutations in the TPR domain of cyclophilin 40 identify conformational states with distinct dynamic and enzymatic properties. 2018;11919(2004).

34. Piao L, Chen Z, Li Q, Liu R, Song W, Kong R, et al. Molecular dynamics simulations of wild type and mutants of SAPAP in complexed with shank3. Int J Mol Sci [Internet]. 2019 Jan 8;20(1):224. Available from: https://www.mdpi.com/1422-0067/20/1/224

35. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019;47(D1):D745–51.

36. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48(D1):D682–8.

37. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Ridwan Amode M, et al. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–91.

38. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15.

39. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids

Res. 2021;49(D1):D480–9.

40. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature [Internet]. 2021;596(7873):583–9. Available from: http://dx.doi.org/10.1038/s41586-021-03819-2

41. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc [Internet]. 2015 Jun 7;10(6):845–58. Available from: http://www.nature.com/articles/nprot.2015.053

42. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? J Mol Biol. 2019 May 17;431(11):2197–212.

43. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci. 2018;27(1):129–34.

44. Emile Apol, Apostolov R, Berendsen HJC, Buuren A van, Bjelkmar P, Drunen R van, et al. GROMACS Reference Manual 2018.3. 2018;258. Available from: http://books.google.com/books?id=-pn8K53IUqgC&pgis=1

45. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. J Comput Chem. 2005;26(16):1701–18.

46. Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. J Comput Chem [Internet]. 2013 Sep 30;34(25):2135–45. Available from: http://doi.wiley.com/10.1002/jcc.23354

47. Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, et al. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. J Chem Theory Comput. 2016;12(1):405–13.

48. Lee J, Hitzenberger M, Rieger M, Kern NR, Zacharias M, Im W. CHARMM-GUI supports the Amber force fields. J Chem Phys [Internet]. 2020 Jul 21;153(3):035103. Available from: https://doi.org/10.1063/5.0012280

49. Jo S, Cheng X, Lee J, Kim S, Park S jun, Patel DS, et al. CHARMM-GUI 10 years for biomolecular modeling and simulation. J Comput Chem [Internet]. 2017 Jun 5;38(15):1114–24. Available from: http://doi.wiley.com/10.1002/jcc.24660

50. Fährrolfes R, Bietz S, Flachsenberg F, Meyder A, Nittinger E, Otto T, et al. Proteins Plus: A web portal for structure analysis of macromolecules. Nucleic Acids Res. 2017;45(W1):W337–43.

51. Schöning-Stierand K, Diedrich K, Fährrolfes R, Flachsenberg F, Meyder

A, Nittinger E, et al. ProteinsPlus: Interactive analysis of protein–ligand binding interfaces. Nucleic Acids Res. 2020;48(W1):W48–53.

52. Stierand K, Maaß PC, Rarey M. Molecular complexes at a glance: Automated generation of two-dimensional complex diagrams. Bioinformatics. 2006;22(14):1710–6.

53. Desta IT, Porter KA, Xia B, Kozakov D, Vajda S. Performance and Its Limits in Rigid Body Protein-Protein Docking. Structure [Internet]. 2020;28(9):1071-1081.e3. Available from: https://doi.org/10.1016/j.str.2020.06.006

54. Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella SE, Xia B, et al. New additions to the ClusPro server motivated by CAPRI. Proteins Struct Funct Bioinforma. 2017;85(3):435–44.

55. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein–protein docking. Nat Protoc [Internet]. 2017 Feb 12;12(2):255–78. Available from: https://www.nature.com/articles/nprot.2016.169

56. Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, et al. How good is automated protein docking? Proteins Struct Funct Bioinforma [Internet]. 2013 Dec;81(12):2159–66. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf

57. Kobe B, Kemp BE. Principles of kinase regulation [Internet]. Vol. 2, Handbook of Cell Signaling, 2/e. Elsevier Science (USA); 2010. 559–563 p. Available from: http://dx.doi.org/10.1016/B978-0-12-124546-7.50450-2

58. Gispert S, Auburger G, Kuruvilla KP, LeDoux MS. Rodent Models of Autosomal Recessive Parkinson Disease [Internet]. Second Edi. Movement Disorders: Genetics and Models: Second Edition. Elsevier Inc.; 2015. 329–343 p. Available from: http://dx.doi.org/10.1016/B978-0-12-405195-9.00019-6

59. Edelman AM, Kim WY, Higgins D, Goldstein EG, Oberdoerster M, Sigurdson W. Doublecortin kinase-2, a novel doublecortin-related protein kinase associated with terminal segments of axons and dendrites. J Biol Chem [Internet]. 2005;280(9):8531–43. Available from: http://dx.doi.org/10.1074/jbc.M411027200

60. Bosshard HR, Marti DN, Jelesarov I. Protein stabilization by salt bridges: Concepts, experimental approaches and clarification of some misunderstandings. J Mol Recognit. 2004;17(1):1–16.

61. Kerjan G, Koizumi H, Han EB, Dubé CM, Djakovic SN, Patrick GN, et al. Mice lacking doublecortin and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures. Proc Natl Acad Sci U S A. 2009;106(16):6766–71.

62. Marsal M, Bernardello M, Gualda EJ, Loza-Alvarez P. Multiple asters

organize the yolk microtubule network during dclk2-GFP zebrafish epiboly. Sci Rep [Internet]. 2022;12(1):1–15. Available from: https://doi.org/10.1038/s41598-022-07747-7

63. Caporizzo MA, Chen CY, Bedi K, Margulies KB, Prosser BL. Microtubules increase diastolic stiffness in failing human cardiomyocytes and myocardium. Circulation. 2020;902–15.

64. Caporizzo MA, Chen CY, Salomon AK, Margulies KB, Prosser BL. Microtubules Provide a Viscoelastic Resistance to Myocyte Motion. Biophys J [Internet]. 2018;115(9):1796–807. Available from: https://doi.org/10.1016/j.bpj.2018.09.019

65. Freeman GJ, Casasnovas JM, Umetsu DT, DeKruyff RH. TIM genes: a family of cell surface phosphatidylserine receptors that regulate innate and adaptive immunity. Immunol Rev [Internet]. 2010 May;235(1):172–89. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.0105-2896.2010.00903.x

66. Santiago C, Ballesteros A, Martínez-Muñoz L, Mellado M, Kaplan GG, Freeman GJ, et al. Structures of T Cell Immunoglobulin Mucin Protein 4 Show a Metal-Ion-Dependent Ligand Binding Site where Phosphatidylserine Binds. Immunity. 2007;27(6):941–51.

67. Nicolás-Ávila JA, Lechuga-Vieco A V., Esteban-Martínez L, Sánchez-Díaz M, Díaz-García E, Santiago DJ, et al. A Network of Macrophages Supports Mitochondrial Homeostasis in the Heart. Cell. 2020;183(1):94-109.e23.

68. Rodriguez-Manzanet R, Sanjuan MA, Wu HY, Quintana FJ, Xiao S, Anderson AC, et al. T and B cell hyperactivity and autoimmunity associated with niche-specific defects in apoptotic body clearance in TIM-4-deficient mice. Proc Natl Acad Sci U S A. 2010;107(19):8706–11.

69. Albacker LA, Karisola P, Chang YJ, Umetsu SE, Zhou M, Akbari O, et al. TIM-4, a Receptor for Phosphatidylserine, Controls Adaptive Immunity by Regulating the Removal of Antigen-Specific T Cells. J Immunol. 2010;185(11):6839–49.

70. Liu W, Xu L, Liang X, Liu X, Zhao Y, Ma C, et al. Tim-4 in Health and Disease: Friend or Foe? Front Immunol. 2020;11(April):1–10.

71. Pinto AR, Ilinykh A, Ivey MJ, Kuwabara JT, D'antoni ML, Debuque R, et al. Revisiting cardiac cellular composition. Circ Res. 2016;118(3):400–9.

72. Alvarez-Argote S, O'meara CC. The evolving roles of cardiac macrophages in homeostasis, regeneration, and repair. Int J Mol Sci. 2021;22(15).

73. Dick SA, Macklin JA, Nejat S, Momen A, Clemente-Casares X, Althagafi MG, et al. Self-renewing resident cardiac macrophages limit adverse remodeling following myocardial infarction. Nat Immunol [Internet]. 2019 Jan 11;20(1):29–39. Available from:

http://www.nature.com/articles/s41590-018-0272-2

74. Souza DS, Barreto T de O, Santana MNS, Menezes-Filho JER, Cruz JS, de Vasconcelos CML. Resident macrophages orchestrating heart rate. Arq Bras Cardiol. 2019;112(5):588–91.

75. Moon B, Lee J, Lee SA, Min C, Moon H, Kim D, et al. Mertk Interacts with Tim-4 to Enhance Tim-4-Mediated Efferocytosis. Cells. 2020;9(7):1–11.

76. Li Y, Li Q, Fan GC. Macrophage efferocytosis in cardiac pathophysiology and repair. Shock. 2021;55(2):177–88.

77. Sugita J, Fujiu K, Nakayama Y, Matsubara T, Matsuda J, Oshima T, et al. Cardiac macrophages prevent sudden death during heart stress. Nat Commun [Internet]. 2021;12(1):1–8. Available from: http://dx.doi.org/10.1038/s41467-021-22178-0

78. Lambert JM, Lopez EF, Lindsey ML. Macrophage roles following myocardial infarction. Int J Cardiol. 2008;130(2):147–58.

79. Duncan SE, Gao S, Sarhene M, Coffie JW, Linhua D, Bao X, et al. Macrophage Activities in Myocardial Infarction and Heart Failure. Cardiol Res Pract. 2020;2020.

80. Billman GE. Homeostasis: The Underappreciated and Far Too Often Ignored Central Organizing Principle of Physiology. Front Physiol. 2020;11(March):1–12.

81. Morioka S, Maueröder C, Ravichandran KS. Living on the Edge: Efferocytosis at the Interface of Homeostasis and Pathology. Immunity [Internet]. 2019 May;50(5):1149–62. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1074761319301980

82. Cummings RJ, Barbet G, Bongers G, Hartmann BM, Gettler K, Muniz L, et al. Different tissue phagocytes sample apoptotic cells to direct distinct homeostasis programs. Nature. 2016;539(7630):565–9.

83. Nishi C, Toda S, Segawa K, Nagata S. Tim4- and MerTK-Mediated Engulfment of Apoptotic Cells by Mouse Resident Peritoneal Macrophages. Mol Cell Biol. 2014;34(8):1512–20.

84. Clancy RM, Kapur RP, Molad Y, Askanase AD, Buyon JP. Immunohistologic Evidence Supports Apoptosis, IgG Deposition, and Novel Macrophage/Fibroblast Crosstalk in the Pathologic Cascade Leading to Congenital Heart Block. Arthritis Rheum. 2004;50(1):173–82.

85. Clancy RM, Neufing PJ, Zheng P, O'Mahony M, Nimmerjahn F, Gordon TP, et al. Impaired clearance of apoptotic cardiocytes is linked to anti-SSA/Ro and -SSB/La antibodies in the pathogenesis of congenital heart block. J Clin Invest. 2006;116(9):2413–22.

86. Briassouli P, Komissarova E V., Clancy RM, Buyon JP. Role of the Urokinase Plasminogen Activator Receptor in Mediating Impaired Efferocytosis of Anti-SSA/Ro–Bound Apoptotic Cardiocytes. Circ Res

[Internet]. 2010 Aug 6;107(3):374–87. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf

87. Aerts L, Terry NA, Sainath NN, Torres C, Martín MG, Ramos-Molina B, et al. Novel homozygous inactivating mutation in the PCSK1 gene in an infant with congenital malabsorptive diarrhea. Genes (Basel). 2021;12(5).

88. Löffler D, Behrendt S, Creemers JWM, Klammt J, Aust G, Stanik J, et al. Functional and clinical relevance of novel and known PCSK1 variants for childhood obesity and glucose metabolism. Mol Metab. 2017;6(3):295–305.

89. Gong Y, Wei ZR. Identification of PSMD14 as a potential novel prognosis biomarker and therapeutic target for osteosarcoma. Cancer Rep. 2022;5(7):1–15.

90. Hu J, Ding X, Tian S, Chu Y, Liu Z, Li Y, et al. TRIM39 deficiency inhibits tumor progression and autophagic flux in colorectal cancer via suppressing the activity of Rab7. Cell Death Dis [Internet]. 2021;12(4). Available from: http://dx.doi.org/10.1038/s41419-021-03670-3