EDUARDO ARCANJO URIOSTE

“PARÂMETROS BIOINFORMÁTICOS DO CONTEXTO GENÔMICO COMO PREDITORES DO EFEITO FUNCIONAL DE SUBSTITUIÇÕES PONTUAIS NA SEQUÊNCIA 5’ UTR EM GENES HUMANOS”

PIRACICABA
2013

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ODONTOLOGIA DE PIRACICABA

EDUARDO ARCANJO URIOSTE

"PARÂMETROS BIOINFORMÁTICOS DO CONTEXTO GENÔMICO COMO PREDITORES DO EFEITO FUNCIONAL DE SUBSTITUIÇÕES PONTUAIS NA SEQUÊNCIA 5' UTR EM GENES HUMANOS"

Orientador:
Sérgio Roberto Peres Line

DISSERTAÇÃO DE MESTRADO APRESENTADA À FACULDADE DE ODONTOLOGIA DE PIRACICABA DA UNICAMP PARA OBTENÇÃO DO TÍTULO DE MESTRE EM BIOLOGIA BUCO-DENTAL NA ÁREA DE HISTOLOGIA E EMBRIOLOGIA

Este exemplar corresponde á versão final
da Dissertação defendida pelo aluno Eduardo Arcanjo Urioste,
e orientada pelo Prof. Dr. Sérgio Roberto Peres Line.

Assinatura do Orientador

_____

PIRACICABA
2013

FICHA CATALOGRÁFICA ELABORADA POR
JOSIDELMA F COSTA DE SOUZA – CRB8/5894 - BIBLIOTECA DA
FACULDADE DE ODONTOLOGIA DE PIRACICABA DA UNICAMP

Urioste, Eduardo Arcanjo, 1989-
Ur3p Parâmetros bioinformáticos do contexto genômico como preditores do efeito funcional de substituições pontuais na sequência 5' UTR em genes humanos / Eduardo Arcanjo Urioste. -- Piracicaba, SP : [s.n.], 2013.

Orientador: Sérgio Roberto Peres Line.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Odontologia de Piracicaba.

1. Regiões 5' não traduzidas. 2. Mutação. 3. Curva ROC. 4. Biologia computacional. I. Line, Sérgio Roberto Peres, 1963- II. Universidade Estadual de Campinas. Faculdade de Odontologia de Piracicaba. III. Título.

Informações para a Biblioteca Digital

Título em Inglês: Bioinformatic parameters of genomic context as predictors of functional impact in point substitutions of human gene 5' UTR
Palavras-chave em Inglês:
5' untranslated regions
Mutation
ROC curve
Computational biology
Área de concentração: Histologia e Embriologia
Titulação: Mestre em Biologia Buco-Dental
Banca examinadora:
Sérgio Roberto Peres Line [Orientador]
Nilva de Karla Cervigne
Danielle Gregorio Gomes Caldas
Data da defesa: 19-04-2013
Programa de Pós-Graduação: Biologia Buco-Dental

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Odontologia de Piracicaba

A Comissão Julgadora dos trabalhos de Defesa de Dissertação de Mestrado, em sessão pública realizada em 19 de Abril de 2013, considerou o candidato EDUARDO ARCANJO URIOSTE aprovado.

Prof. Dr. SERGIO ROBERTO PERES LINE

Profa. Dra. DANIELLE GREGORIO GOMES CALDAS

Profa. Dra. NILVA DE KARLA CERVIGNE

Dedico este trabalho ao meu querido pai José Urioste Gonçalves Jr. a quem sempre tive muito orgulho de ser filho. Nada disso teria sido possível sem seu o sacrifício, sua dedicação e o seu amor por mim e por nossa família. Você foi e sempre será minha inspiração como ser humano. Também dedico à minha mãe Antonia e minha irmã Sarah, minha linda família.

AGRADECIMENTOS

Ao meu orientador Prof. Dr. Sérgio Roberto Peres Line, pela confiança, compreensão e humildade que tornaram a realização desse trabalho não só possível, mas também prazerosa e construtiva. Espero seguir seu exemplo e ter a oportunidade de realizar muitos outros trabalhos da mesma forma ao longo de minha vida acadêmica.

Às professoras participantes da banca Nilva de Karla Cervigne e Danielle Gregório Gomes Caldas, pelo interesse e pelas contribuições a este trabalho.

À Profa. Ana Paula de Souza Pardo, pela paciência e pelas oportunidades apresentadas ao longo de tantas mudanças e contratempos ocorridos no decorrer de meu mestrado.

Ao Prof. Dr. Marcelo Rocha Marques pela grande humildade e dedicação exemplares ao longo das disciplinas, sempre preocupado em nos fazer dar o nosso melhor.

A todos amigos e colegas da FOP pela perguntas, respostas, ajudas, brincadeiras e momentos simples, por terem tornado este período de formação bastante agradável.

Às agencias de fomento CAPES e CNPQ por terem financiado minha formação acadêmica.

E finalmente a todos os demais professores e funcionários da FOP, que contribuíram mesmo que de forma indireta e desconhecida para a elaboração deste trabalho.

RESUMO

Estima-se que cada indivíduo carregue cerca de 120 a 430 variantes raras em regiões UTRs (Abecasis *et al*, 2012). Apesar da tolerância a variação na região 5' UTR, a pato-fisiologia de várias doenças está ligada a mutações na mesma (Cazzola & Skoda, 2000; Reynolds, 2002; Chatterjee & Pal, 2009; Wethmar *et al* 2010), sendo necessário o entendimento a determinação dos mecanismos regulatórios. O objetivo deste trabalho é descobrir assinaturas genéticas encontradas no contexto genômico de mutações pontuais de região 5' UTR que permitam prever o impacto funcional de outras variações pontuais na mesma região. As mutações, causadora de doença, foram selecionadas do banco de dados do Human Gene Mutation Database (HGMD) (Stenson *et al*, 2008); e os polimorfismos, de impacto funcional desconhecido, foram obtidos no banco de dados NHLBI Grand Opportunity Exome Sequencing Project (ESP), sendo originados do trabalho de Tenessen *et al* (2012). No total foram utilizadas 235 mutações e 21.542 polimorfismos. Para as variações foram calculados parâmetros de variação da estabilidade da estrutura secundária do contexto das variações (ΔΔG*folding*), presença de sítios de ligação de fatores de transcrição (JASPAR), tipo de variação (transição/transversão, tipoV), distância do início da sequência codificante (DiSC), distância do início de transcrição (DiTr) e conservação filogenética por distância de Levenshtein do contexto (Lev). A estatística foi calculada pelos testes de Wilcoxon e Binomial. A partir destes foram gerados modelos de regressão logística analisados através de curva ROC. Os parâmetros ΔΔG*folding* máximo, tipoV, DiSC, e Lev permitiram a distinção significativa ($\alpha = 0,05$) entres os os polimorfismos e as mutações permitindo modelos explicativos mas incompletos (área da Curva ROC 0,772). ΔΔG*folding* max. indicou uma relação entre as mutações e entre estruturas secundárias mais estáveis geradas pelas mesmas. Os parâmetros Lev e tipoV sugerem a origem das mutações como resultantes de *hotspots*. O parâmetro DiSC indicou regiões com provável funcionalidade. Apesar de não ter sido possível estabelecer relação causal entre os parâmetros e o impacto funcional das variações, encontraram-se correlações importantes.

Palavras chave: 5' UTR, mutações, polimorfismos, curva ROC, bioinformática.

ABSTRACT

It is estimated that each individual carries about 120 to 430 rare variante in the UTR regions (Abecasis *et al*, 2012). Despite the increased tolerance towards variations in 5' UTR region, the patho-phisiology of several disease are linked to its mutations (Cazzola & Skoda, 2000; Reynolds, 2002; Chatterjee & Pal, 2009; Wethmar *et al* 2010). Therefore it is necessary the understanding and the determination of the regulatory elements. The objective of this study is the discovery of genetic signatures found in the genomic context of disease causing point mutations in 5' UTR, thus allowing the prediction of the functional impact of other point variations in the same region. The disease causing mutations were selected from Human Gene Mutation Database (HGMD) (Stenson *et al*, 2008). The polymorphisms of unknown functional impact were obtained from the NHLBI Grand Opportunity Exome Sequencing Project (ESP), originated from the work of Tenessen *et al* (2012). A total of 235 mutations and 21,542 polymorphisms were used. For each variation, parameters related with the differences of the variation's context folding stability (ΔΔG*folding*), presence of transcription factor binding sites (JASPAR), type of variation (transition/transversion, tipoV), distance from coding sequence start (DiSC), distance from transcription start site (DiTr) and phylogenetic conservations by distance of Levenshtein from wild type to variant context (Lev). The statistical test was done by Wilcoxon and Binomial. Logistical regressions models were generated from the parameters and its performance was evaluated by a ROC curve. The parameters maximal ΔΔGfolding, tipoV, logarithm of DiSC and Lev allowed a significant distinction ($\alpha = 0,05$) between the groups, generating models of reasonable explanation but incomplete (area under the ROC curve 0,772). Maximal ΔΔGfolding showed a relationship between mutations and stable secondary structures generated by them. Lev and tipoV suggested the origin of the mutation from hotspots. The DiSC parameter identified regions with possible functionality. While it was not possible to establish any clear causal relationship between the parameters and the functional impact of the variations, important correlations were found.

Key words: 5' UTR, mutations, polymorphisms, ROC curve, bioinformatics.

SUMÁRIO

1 INTRODUÇÃO

Estima-se que cada indivíduo carregue cerca de 120 a 430 variantes raras em UTRs (*Untranslated Regions,* regiões não traduzidas) (Abecasis *et al*, 2012). Verifica-se também um enriquecimento de variantes raras com provável perda de função próximo às extremidades 5' dos genes, indicando que tais variantes possam ser compensadas através de outros mecanismos regulatórios como sítios distintos de inicialização da transcrição (MacArthur *et al*, 2012). Estudos de ortologia de éxons de sequências entre humanos, camundongo e primatas revelam que estas regiões estão sob pressão de seleção menores do que as sequências codificantes, sofrendo evolução mais rápida. Tais regiões abrigam éxons sem ortologia entre diferentes espécies, indicando uma tolerância maior a variações e participação em mecanismos regulatórios específicos de cada organismo (Resch *et al*, 2009; Fu & Lin, 2012). Apesar da aparente maior tolerância a variação na região 5' UTR, a pato-fisiologia de várias doenças está ligada a mutações na mesma (Cazzola & Skoda, 2000; Reynolds, 2002; Chatterjee & Pal, 2009; Wethmar *et al* 2010), o que ressalta a importância de um maior entendimento não só dos elementos regulatórios conhecidos mas também da busca de outros mecanismos ainda não bem entendidos. O objetivo deste trabalho é a descoberta de assinaturas genéticas encontradas no contexto genômico de mutações pontuais de região 5' UTR, causadoras de doenças, que permitam prever o impacto funcional de outras variações pontuais na mesma região.

2 REVISÃO DE LITERATURA

2.1 O material genético

O entendimento da natureza do material genético teve suas raízes na descoberta da transformação pelo médico militar inglês Frederick Griffith (1928) ao descobrir que a suspensão celular de bactérias *Streptococcus pneumoniae* virulentas, mortas por calor, era capaz de transformar bactérias de colônias não virulentas da mesmas espécie, conferindo-lhes virulência. Ao material das bactérias capaz de realizar a transformação foi dado o nome de "princípio transformante". Posteriormente, Avery *et al* (1944), ao trabalhar com extratos purificados livres de células, determinou a natureza química do "princípio transformante" como sendo moléculas de ácido desoxirribonucleico (DNA). Entretanto o entendimento dos mecanismos genéticos só seria possível com o trabalho de Watson & Crick (1953), Nobel de fisiologia ou medicina em 1962, que estabeleceu o modelo estrutural de dupla hélice do DNA, descrito a seguir.

2.1.1 Fundamentos bioquímicos da genética

Em sua estrutura primária o código genético contendo as informações que determinam as características de um organismo, se armazena em longas cadeias poliméricas de ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA), sendo a unidade estrutural do mesmos denominada nucleotídeos. Cada nucleotídeo é composto por um grupo fosfato ligado a uma pentose (monossacarídeo com cadeia orgânica de 5 carbonos, sendo 2-desoxirribose em cadeias de DNA, e ribose em cadeias de RNA), ligada a uma base nitrogenada, que é formada por um ou dois anéis aromáticos contendo nitrogênio, sendo classificadas como pirimidinas ou purinas respectivamente. A base nitrogenada mais sua respectiva pentose é chamada de nucleosídeo (Figura 1). Os nucleotídeos formam cadeias através da ligação do grupo hidroxil, ligado ao carbono

3' da pentose, com o grupo fosfato do nucleotídeo seguinte, ligado ao carbono 5' da pentose, constituindo uma ligação fosfodiéster, restando apenas um grupo hidroxil 3' e um grupo fosfato 5' livres nas extremidades da cadeia de ácido nucleico. O sentido convencional de leitura de cadeias de ácido nucleico se dá da extremidade 5' em direção à extremidade 3'.

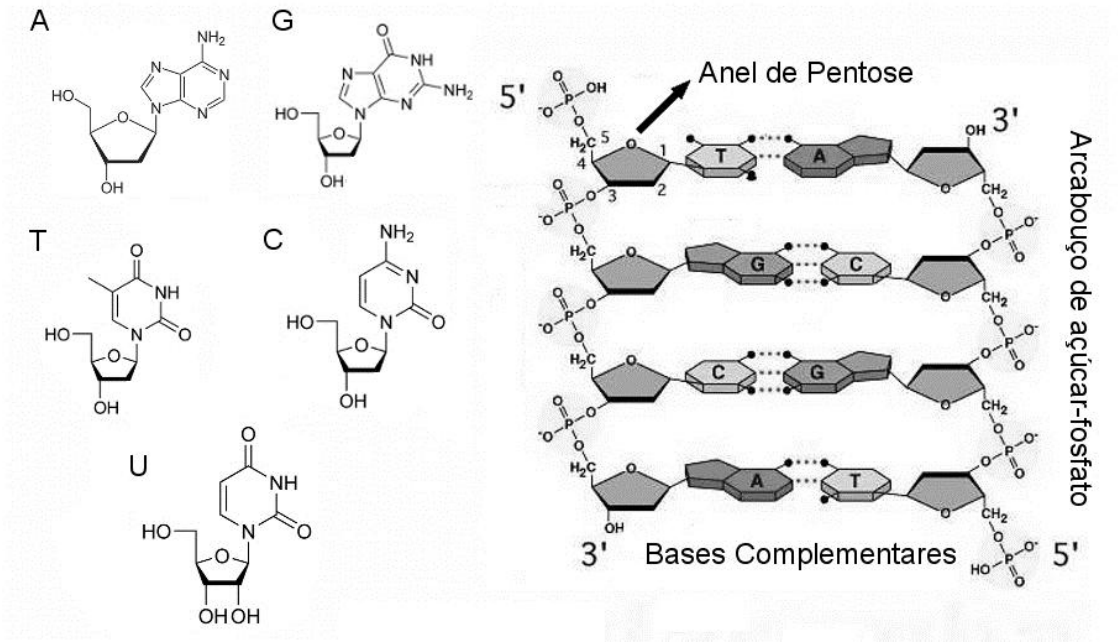

Figura 1 – Formula estrutural dos nucleosídeos (esquerda), formados pelas bases nitrogenadas adenina (A), guanina (G), timina (T), citosina (C), mais a pentose desoxirribose (encontrada no DNA) e uracila (U), mais a pentose ribose (encontrada no RNA). Esquema de cadeia complementar anti-paralela de DNA (direita). Os nucleosídeos apresentam-se unidos entre si por resíduos de fosfato, formando o arcabouço de açúcar fosfato. Notar a numeração convencional do anel de pentose (seta).

O DNA é o ácido nucleico responsável pelo armazenamento da informação genética, formado de longas cadeias estáveis. Seus nucleotídeos são caraterizados pelas bases nitrogenadas timina (T), citosina (C), adenina (A) e guanina (G). Duas cadeias de DNA se acoplam formando uma dupla-hélice antiparalela (orientação 5'-3' inversas, Figura 1). Tal acoplamento é determinado pelas ligações de pontes de hidrogênio das bases nitrogenadas das cadeias, que são ditas complementares. Timina é complementar à adenina, enquanto que a citosina é complementar à guanina. A sequência de bases

nitrogenadas da cadeia é que determina a informação genética, que é traduzida a partir do código genético. O RNA tem como função transcrever a informação da molécula de DNA para que possa ser traduzida e expressa em proteínas. Apresenta como diferença em relação ao DNA, a presença da base nitrogenada uracila (U) em vez da base timina, que apresenta complementariedade à adenina. É encontrado em cadeias simples na maioria dos organismos, mais curtas que às de DNA, contendo o mesmo código genético.

2.1.2 Estrutura do gene eucariótico

A um fragmento de DNA capaz de codificar um polipeptídio dá se o nome de gene. O gene eucariótico entretanto, normalmente apresenta outras sequências ao redor e dentro da sua sequência expressa em polipeptídio (a sequência codificante) podendo esta, desta forma, ocorrer de maneira não contínua. Os elementos presentes na maioria dos genes eucarióticos, seguindo o sentido 5'-3', são:

Promotor Basal: Nome dado a região sequência contendo os sítios de ligação para a enzima RNA polimerase e para todas as proteínas fatores de transcrição necessárias e suficientes para que o início da transcrição, isto é, da síntese de uma molécula de RNA a partir da molécula de DNA do gene, gerando uma cadeia complementar à mesma. Encontra-se *upstream* ("rio acima", "antes" no sentido 5'-3') ao sítio de início de transcrição, que delimita sua extremidade 3'.

Região não traduzida 5' (*Five Prime Unstranslated Region*, 5' UTR): Delimitada *upstream* pelo sítio de início de transcrição e *downstream* ("rio abaixo", "depois" no sentido 5'-3') pelo sítio de início da tradução. É uma região transcrita porém não traduzida, na extremidade 5' do pré-RNA mensageiro (pré-mRNA).

Região Codificante: É a região compreendida entre os sítios de início e terminação do processo de tradução do código genético em uma cadeia polipeptídica. Tais sítios são delimitados pelos códons de iniciação (*start codon*, trinca de bases ATG no DNA, e AUG no RNA) e pelos códons de terminação (*stop codon*, trincas de base TAG, TAA ou TGA no DNA e UAG, UAA ou UGA no RNA), sendo este intervalo também denominado fase de leitura aberta ou *open reading frame* (ORF). Dentro dessa região pode se distinguir as regiões de *exons*, que são traduzidas em uma cadeia polipeptídica, intercaladas por regiões de *introns*, que são excisadas durante o processamento do pré-mRNA, chamado de *splicing*, não estando presentes no RNA mensageiro maduro (mRNA), e portanto não sendo traduzidas no polipeptídio.

Região não traduzida 3' (*Three Prime Unstranslated Region*, 3' UTR): Delimitada *upstream* pelo sítio de terminação de tradução (*stop codon*) e *downstream* pelo sítio de terminação da transcrição. É uma região transcrita porém não traduzida, na extremidade 3' do pré-RNA mensageiro (pré-mRNA).

Figura 2 – Esquema da estrutura de um gene eucariótico. Sequências flanqueadoras do gene, não transcritas estão representadas pelo tracejado. *Exon*s estão representados pelas caixas. *Introns* estão representados pelas linhas angulares. Caixas destacadas pelas faixas representam a sequência codificante. As caixas brancas representam as UTRs.

2.1.3 Expressão gênica

A expressão de um gene se dá em várias etapas: primeiramente, a sequência de DNA entre o promotor basal e o sítio de terminação da transcrição é transcrita em uma

cadeia de RNA reversa e complementar. O RNA assim sintetizado é denominado pré-RNA mensageiro. O pré-mRNA passa pelo processamento ou *splicing*: seus *introns* removidos, seus *exons* ligados uns aos outros, e por fim há a adição do *cap* 5' (base nitrogenada 7-metil guanosina) na extremidade 5' e de uma cauda de poli-adenosina (cauda poli-A) na extremidade 3', constituindo assim o RNA mensageiro maduro (mRNA). Este é então reconhecido pelas subunidades ribossomais, que catalisam a síntese de uma cadeia polipeptídica a partir da sequência da região codificante, através do processo denominado tradução. A expressão gênica também é influenciada pelo balanço entre síntese e degradação dos mRNAs e das proteínas.

2.2 A Região 5' UTR

O tamanho médio das 5' UTRs humanas é de 210 nucleotídeos, podendo variando entre 18 a 2803 nucleotídeos (Pesole *et al*, 2001). Esta região apresenta função na regulação da tradução do mRNA. Trabalhos mostram que a regulação traducional da expressão gênica é tão importante quanto a regulação transcricional para o funcionamento correto das células, e que sua disfunção está ligada à pato-fisiologia de várias doenças (Cazzola & Skoda, 2000; Reynolds, 2002; Scheper *et al*, 2007; Chatterjee & Pal, 2009; Wethmar *et al* 2010). Estima-se que 1% da abundância de uma proteína seja explicada pelo comprimento e composição da região 5' UTR de seu mRNA, enquanto que 31% seja explicado pelas características de sequência codificante, 27% seja explicado pela abundância do mRNA, 8% pelas características da região 3' UTR e 33% seja explicado por fatores ainda desconhecidos (Vogel *et al* 2010). A eficiência da tradução depende do recrutamento de subunidades ribossomais livres e de fatores de inicialização, seguido da inicialização propriamente dita. A inicialização do processo de tradução normalmente depende do reconhecimento do *cap* 5’ m7G pelo complexo de inicialização 43S, composto pela subunidade ribossomal 40S mais os Fatores de Inicialização Eucarióticos (*eukaryotic Initiation Factor,* eIF) 1, 1A, 2, 3 e 4. O reconhecimento e acoplamento do complexo 43S é auxiliado pela ação dos fatores eIF4F (formado por eIF 4A, 4G e 4E) e eIF4B ou eIF4H, que atuam desenovelando

estruturas secundárias do mRNA na 5' UTR próximas ao *cap* 5'. Tal processo de inicialização constitui a via canônica e é denominado de *cap* dependente. Após o acoplamento, o complexo 43S "varre" o mRNA no sentido 5'-3', até encontrar um códon de inicialização, AUG, no qual há o recrutamento do complexo ternário eIF2–GTP–Met-tRNA^{Met}, a liberação dos fatores de inicialização, o acoplamento da subunidade ribossomal 60S e o início da tradução do mRNA (Jackson *et al*, 2010). O sítio de inicialização da tradução geralmente consiste no primeiro AUG no contexto ótimo GCC(A/G)CCAUGG (Kozak, 1991). Este é considerado o passo limitante e mais regulado da tradução.

Acredita-se que as características da região 5' UTR tenham papel fundamental em prever a eficiência de tradução de um determinado mRNA. São tidos como mRNAs "fracos" em relação a eficiência de tradução aqueles que possuem região 5'UTR muito estruturada, com energia livre de Gibbs (ΔG) próxima de -50 kcal/mol ou menos, mais de 100 nucleotídeos de comprimento, fases de leitura abertas *upstream* (*upstream reading frames*, uORFs) e *codons* de inicialização *upstream* (uAUGs). Em contraste, mRNAs "fortes" teriam 5' UTRs relativamente curtas e pouco estruturadas (Spruill & McDermott, 2009).

2.2.1 Estruturas secundárias e RNA *Binding Proteins*

Acredita-se que estruturas secundárias possam inibir a tradução, sendo necessário apenas um *hairpin* com energia livre inferior a -30 kcal/mol para bloquear o acesso do complexo de pré-iniciação ao mRNA. Quando localizados mais adiante, *hairpins* requerem uma energia livre inferior a -50 kcal/mol para serem capazes de impedir a tradução, por inibição estérica (Araujo *et al*, 2012). O trabalho de Vassilenko *et al* (2011), entretanto, demonstraram que o tempo de varredura da subunidade 43S até o primeiro AUG em contexto favorável está correlacionado apenas com o comprimento da região 5' UTR, ocorrendo de maneira contínua, linear e unidirecional, independentemente da presença de *stem loops* com energia igual a -30 kcal/mol. A

correlação de estruturas secundárias de 5’ UTR com a função gênica tem sido sugerida pela sua prevalência em mRNA codificantes de fatores de transcrição, de proto-oncogenes, de fatores de crescimento e seus receptores, e de proteínas pouco expressas em condições normais. Mais de 90% dos transcritos de genes destes tipos possuem 5' UTR contendo estruturas secundárias com energia livre de Gibbs inferiores a -50 kcal/mol, sendo que 60% destas estruturas secundárias se posicionam próximas às estruturas do *cap* 5' m7G (Gray & Hentze, 1994; Pickering & Wilis, 2005; Araujo *et al*, 2012)

Um exemplo de estrutura secundária com função inibitória traducional em região 5' UTR ocorre no mRNA do fator de crescimento transformante beta 1 (*Transforming Growth Factor* – TGF-β1), no qual um *motif* conservado tem papel na formação de uma estrutura secundária de *stem loop*. Uma mutação neste *motif* desestrutura o mRNA e resulta no aumento da eficiência traducional. Interessantemente esta estrutura é prevista como apenas moderadamente estável, com energia de -24 kcal/mol, sendo considerada portanto insuficiente por si só no impedimento da varredura da subunidade ribossomal 43S na 5' UTR (Jenkins *et al,* 2010).

Estruturas secundárias também podem facilitar a tradução, como no exemplo de um *stem loop* na região 5' UTR do mRNA da cadeia α1 do colágeno tipo I. Neste caso foi mostrado *in vivo* que uma mutação abolindo a estrutura secundária reduzia significantemente a expressão da cadeia α1 (Parsons *et al*, 2010). Andreev *et al* (2009) mostram que três mRNA com 5' UTRs longas e estruturadas, dos genes Apaf-1, c-Myc e LINE-1, possuem eficiência de tradução alta e comparável a dos mRNAs de tradução *cap* dependente canônica beta-globina e beta-actina, sugerindo que estrutura secundárias não necessariamente tem efeito inibitório.

Estes três trabalhos (Jenkins *et al,* 2010; Parsons *et al*, 2010; Andreev *et al,* 2009) sugerem que as estruturas secundárias na região 5' UTR devem agir em conjunto com proteínas ligantes de RNA (RNA *Binding Proteins*, RBPs), sendo seus efeitos regulatórios positivos ou negativos resultados da interação de ambos elementos com a maquinaria de inicialização da transcrição. O cálculo da estabilidade da estrutura secundária da região 5' UTR por si só não necessariamente se correlaciona com sua

eficiência de inicialização de tradução. No caso das 5' UTR dos genes de alfa e beta-globina, os resultados para eficiência de tradução obtidos *in vitro*, em lizado de reticulócito de coelho, foram contraditórios aos resultados obtidos em cultura de células, contradizendo as previsões baseadas na estrutura secundária prevista, sendo o mRNA da alfa-globina traduzida mais in células vivas e menos nos lizados in vitro, ao contrário da beta-globina (Babendure *et al*, 2006).

Estima-se que o genoma humano codifique em torno de 1000 RBPs, tendo uma grande parte destas papel na regulação traducional. As RBPs podem ser assim categorizadas em dois grupos: RBPs que são parte da maquinaria básica de tradução de todos os mRNAs, como PABPI e elF4E; e RBPs que atuam seletivamente modulando os níveis de expressão de mRNAs específicos, como HuR e Musashi1. Em relação a este último grupo, tem sido observado que RBPs podem usar mecanismos distintos para aumentar ou diminuir a tradução. As RBPs geralmente reconhecem *motifs* específicos nas UTRs e interagem com a maquinaria de tradução para modular a expressão proteica, geralmente interferindo no passo de inicialização da tradução (Abaza & Gebauer, 2008; Araujo *et al*, 2012).

Um exemplo de regulação traducional mediada por estruturas secundárias e RBPs são os *Iron Responsive Elements* (IREs) caracterizados por um *stem loop* altamente conservado de cerca de 30 nucleotídeos. Os IREs são caracterizados por uma alça de 6 nucleotídeos (hexanucleotídeo) com a sequência CAGYCX (Y = U ou A; X = U, C, ou A), e uma haste superior de 5 nucleotídeos, separada da haste inferior, de tamanho variável, por uma citosina não pareada. Em condições de baixas concentração de ferro, as proteínas *iron regulatory protein* (IRP) 1 e 2 se ligam aos IREs próximos ao *cap* 5', impedindo a inicialização da tradução, ou impedindo a varredura da subunidade 43S, por inibição estérica (Goss & Theil, 2011). Estes elementos são importantes na manutenção da homeostase de ferro, agindo nas 5' UTR dos mRNAs das cadeias H e L de ferritina , ferroportina, ALAS2, aconitase mitocondrial (ACO2) e *hypoxia-inducible factor* 2α (HIF2α/EPAS1) (Hentze *et al*, 2010).

Uma outra forma de estrutura secundária em regiões 5' UTR de mRNAs são os G-quadruplex. Estas estruturas são formadas por interações distintas das ligações

canônicas previstas por Watson & Crick, de complementariedade purina-pirimidina. G-quadruplexes são formados por quartetos de guaninas, em que cada G é doador e aceptor de duas pontes de hidrogênio, se arranjando de maneira planar, formando um quadrado, que pode ser estabilizado por cátions no seu centro. Estes quartetos podem se formar intramolecularmente em sequências ricas em G, com 4 ou mais regiões de Gs contíguos próximas, de forma que os quartetos se empilham uns sobre os outros, formando estruturas secundárias tridimensionais estáveis, estabilizadas mais ainda pelas concentrações fisiológicas de K^+ e Na^+ (Bugaut & Balasubramanian, 2012). Um alto percentual de C+G é uma característica conservada da região 5' UTR, 60% no caso de vertebrados de sangue quente (Babendure *et al*, 2006), o que torna bastante susceptíveis estas regiões não só à formação de estruturas secundárias baseadas nas interações canônicas, mas também a formação de outros tipos de estruturas como os G-quadruplex, que podem atuar de maneira inibitória nos processos de inicialização e varredura da tradução.

No mRNA do proto-oncogene NRAS, um *motif* conservado, tanto em sequência quanto em posição, em humano, camundongo, rato, chimpanzé e cachorro, forma um G-quadruplex, cuja presença resulta em inibição de 80% da tradução do mRNA *in vitro*, sugerindo uma igual função *in vivo.* A posterior análise computacional revelou a presença de *motifs* de prováveis G-quadruplexes na 5' UTR de 2,922 genes humanos, incluindo outros proto-oncogenes como BCL2, JUN e FGR, revelando uma densidade de *motifs* por pares de base 4,8 vezes maior nas 5' UTRs do que a densidade encontrada no restante do genoma humano (Kumari *et al*, 2007). Resultados similares de regulação traducional foram obtidos nos mRNAs dos genes Zic-1, MT3-MMP, BCL-2 e ADAM 10 (Arora *et al*, 2008; Shahid *et al* 2011, Lammich *et al* 2010).

2.2.2 Fases de leitura e AUGs *upstream*

Os uORFs e uAUGs são considerados as principais formas de regulação traducional pela região 5' UTR. As uORFs são sequências definidas por um códon de

inicialização e um códon de terminação *upstream* da região codificante principal, enquanto que os uAUGs são condons de inicialização sem um códon de terminação, na mesma fase de leitura, *upstream* à região codificante principal. Um grande percentual do transcriptoma humano contém uORFs e/ou uAUGs, em torno de 44 a 49%. (Iacono *et al,* 2005; Crowe *et al*, 2006; Calvo *et al*, 2008; Araujo *et al*, 2012). Dados de conservação em um subconjunto de transcritos humanos, de camundongo e de rato indicam que ambos elementos são moderadamente conservados, sendo 38% dos uORFs e 24% dos uAUGs conservados entre as espécies. A razoável conservação de uORFs, o fato de que seu tamanho médio, de 20 nucleotídeos, ser o esperado aleatoriamente, e o fato de que uAUGs atuam como supressores mais fortes da expressão, indica que muitos uAUGs provavelmente foram neutralizados ao longo da evolução pela aquisição de um códon de parada *downstream* (Iacono *et al* 2005, Crowe *et al, 2006;* Araujo *et al*, 2012).

Um exemplo de regulação por uAUG importante é o encontrado no mRNA do gene BRCA1, um supressor de tumor, frequentemente mutado no câncer de mama, com funções envolvidas no ciclo celular, apoptose (morte celular programada), e reparo de dano ao DNA. O BRCA1 produz dois transcritos distintos derivados de dois promotores diferentes, e, portanto, com diferentes 5' UTR. O transcrito mais curto é eficientemente traduzido e é expresso tanto em células cancerígenas quanto em células normais de tecido mamário, enquanto que o transcrito mais longo é predominantemente expresso em câncer de mama. A presença de vários uAUG e de estruturas mais complexas na região 5’ UTR inibe a tradução do transcrito mais longo, causando redução nos níveis de expressão de BRCA1 em células tumorais, promovendo o crescimento celular (Sobczak & Krzyzosiak, 2002).

As uORFs podem inibir a expressão da ORF *downstream*, ao pararem a varredura da subunidade 43S para a tradução das mesmas. Isto implicaria na parada das mesmas no códon de terminação da uORF, aguardando o recrutamento do complexo ternário eIF2–GTP–Met-tRNA^{Met} para o reinício da varredura. A “varredura vazada” da subunidade 43S permite a mesma ignorar uORFs e traduzir a ORF principal de maneira mais eficiente (Ait Ghezala *et al*, 2012). Esse tipo de regulação possui importância na regulação positiva das ORFs principais de genes contendo uORFs e uAUGs em

situações de estresse celular como privação de glicose, níveis alterados de Ca_2^+, e hipoxia. Tais condições levam ao acúmulo de proteína com arranjo estrutural alterado dentro do retículo endoplasmático rugoso, o que resulta indiretamente no aumento da fosforilação pós-traducional de eIF2α. Tal evento inibe de maneira global a síntese proteica em células estressadas, ao impedir a reação de renovação de eIF2–GTP a partir de eIF2–GDP necessário para inicialização da tradução (Ait Ghezala *et al*, 2012). Em condições de disponibilidade de eIF2–GTP reduzida, mais tempo é dado para a varredura da região 5' UTR pela subunidade ribossomal 43S que assim acaba tendo uma maior chance de ignorar os condons de inicialização das uORFs. Esse mecanismo faz parte da regulação dos genes ATF4 e CHOP, fatores de transcrição relacionados a respostas ao estresse celular (Palam *et al*, 2011). No caso do gene ATF4, existem duas uORFs, sendo que a menos *upstream* (uORF2) sobrepõem-se a ORF principal, impedindo sua expressão. Para que esta uORF seja traduzida é necessária a terminação da tradução no códon de parada da outra uORF (uORF1), mais *upstream*, seguida da reinicialização da tradução no AUG seguinte. Em situações de estresse celular a reinicialização ocorre predominantemente no AUG da ORF principal, ignorando o AUG da uORF2 (Ait Ghezala *et al*, 2012). A distância entre os códons de terminação de uma uORF e de inicialização de outra ORF *downstream* influencia a capacidade de reinicialização da subunidade 43S após a terminação da uORF. Distâncias curtas, como de 45 nucleotídeos, reduzem grandemente a eficiência de reinicialização na ORF *downstream* (Kozak, 1987), sugerindo o recrutamento gradativo do complexo ternário ao longo da varredura antes da reinicialização. No caso do gene ATF4, em situações de estresse celular, a distância de aproximadamente 190 nucleotídeos entre o códon de parada da uORF1 e o AUG da ORF principal mostrou-se ótima para a reinicialização da tradução para a síntese de ATF4 (Ait Ghezala *et al*, 2012). Entretanto, Calvo *et al* (2008) não encontraram correlação entre a distância da uORF da sequência codificante principal e o quão deletério é seu efeito na redução dos níveis da proteína codificada pela ORF principal, sugerindo que na maioria dos mRNA contendo uORFs, a tradução da ORF principal se dá por subunidades 43S que ignoraram o AUG do uORF, iniciando a tradução apenas na sua sequencia codificante canônica.

A eficiência de tradução da uORF também parece ter um papel na regulação de ORFs *downstream.* Crowe *et al* (2006) demonstraram que, dentre as uORFs de 20 a 99 códons de comprimento, conservadas entre humanos e não roedores/primatas, aproximadamente 70% possuíam um contexto de inicialização de tradução (sequência de Kozak) considerado ótimo, com um desvio para a ocorrência de polimorfismos sinônimos, sugerindo a importância da tradução destas uORFs. A presença de códons raros na uORF, pouco eficientes tem possivelmente efeito deletério na tradução da ORF principal, podendo atrasar ou parar os ribossomos, impedindo a acessibilidade dos ribossomos para sua tradução. O mesmo pode ocorrer quando a terminação da uORF ocorre de maneira ineficiente (Meijer & Adri, 2003).

Peptídeos sintetizados pela uORF podem também influenciar a tradução, como o peptídeo atenuador de arginina (*arginine attenuator peptide*, AAP), que negativamente controla a tradução de proteínas envolvidas na biosíntese *de novo* de arginina em fungos. Em concentrações altas de arginina, o APP muda de conformação causando parada do ribossomo no códon de terminação da uORF codificante do APP (Wang *et al*, 1999; Gaba *et al*, 2001; Wu *et al*, 2012; Araujo *et al*, 2012).

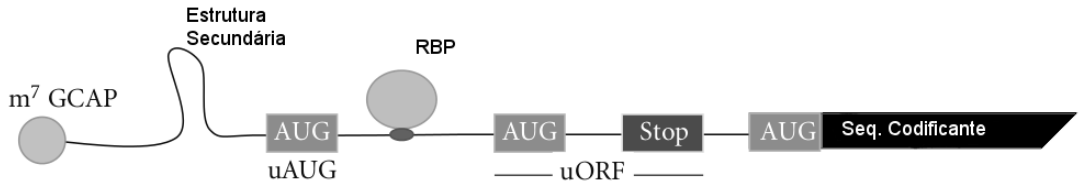

Figura 3 – Esquema representando os elementos regulatórios da região 5' UTR, antecedendo o códon de início da sequência codificante principal do gene. Adaptado de Araujo *et al,* 2012.

2.2.3 Formas de inicialização de tradução não canônicas

Outras formas de inicialização da tradução podem ocorrer de maneira independente do *cap* 5' e até mesmo independentemente de uma extremidade 5'

acessível. Trabalhos mostram que alguns vírus, cujos mRNAs não possuem *cap* 5', são capazes de inicializar a tradução diretamente no códon de inicialização, através de sítios internos de entrada de ribossomo *(Internal Ribossome Entry Sites* – IRES) (Jackson *et al*, 2010). Os IRES virais são estruturas secundárias capazes de recrutar elementos de inicialização da tradução e as subunidades ribossomais, com o auxílio de fatores de trans-ativação de IRES (IRES *trans-activating factors*, ITAFs), próximos ao início de tradução. São exemplos já bem caracterizados os picornavirus, flaviviruses, pestiviruses e HCV. Os IRES celulares ainda não foram completamente caracterizados, apesar de haverem diversos trabalhos mostrando inicialização da tradução independente do reconhecimento de *cap* 5' em mRNAs humanos. Diferentes IRES mostram pouca similaridade estrutural entre si, tendo mecanismos de funcionamento distintos. (Andreev *et al*, 2009; Jackson *et al*, 2010; Terenin *et al,* 2012). Um trabalho com *Saccharomyces cerevisae* e *Drosophila melanogaster* mostrou que a única característica de prováveis IRES celulares em comum nestes organismos é a ausência de estrutura secundária forte (Xia & Holcik, 2009). Ao contrário dos IRES virais, que dependem de *motifs* específicos para o recrutamento dos fatores de inicialização, no IRES celulares, tais *motifs* ainda não foram caracterizados, indicando que seu funcionamento talvez se baseie em interações pouco específicas com os ITAFs, aumentando a probabilidade de inicialização próxima ao códon de início (Andreev *et al*, 2009; Terenin *et al*, 2012).

2.2.4 Polimorfismos na 5’ UTR

Estima-se que cada indivíduo saudável possua em média 10.000 a 11.000 polimorfismos não-sinônimos e 10.000 a 12.000 polimorfismos sinônimos em relação à sequencia de referência do genoma humano. Cada indivíduo carrega mais de 2.500 variantes não sinônimas em posições conservadas, sendo 20-40 variantes consideradas como deletérias em sítios conservados e em torno de 150 variantes de perda de função do gene como ganho de códons de parada, mudanças de fase de leitura nas sequências codificantes e perda de sítios de *splicing* essenciais. Entretanto apenas uma pequena parte destas variações são consideradas raras, entre 130-400, sendo 10-20 variantes de

perda de função. Variantes raras são tidas como prováveis causadoras de doenças, uma vez tais variantes sofrem uma seleção natural negativa, que as impedem de alcançar maiores frequências na população. Acredita-se que tais variantes tenham efeitos deletérios fracos na regulação e função dos genes. (Abecasis *et al*, 2012). Variantes raras são tidas também como eventos mais recentes, sendo o principal diferencial entre populações humanas distintas, correspondendo a 82% das variantes específicas de populações distintas e a 95,7% das variantes com impacto funcional previsto (Tennessen *et al*, 2012). Verifica-se também um enriquecimento de variantes raras com provável perda de função próximo às extremidades 5' dos genes, indicando que tais variantes possam ser compensadas através de outros mecanismos regulatórios como sítios distintos de inicialização da transcrição (MacArthur *et al*, 2012)

Estima-se que cada indivíduo carregue cerca de 120 a 430 variantes raras em regiões UTRs (Abecasis *et al*, 2012). Estudos de Ortologia de éxons de sequências entre humanos, camundongo e primatas revelam que as regiões 5’ UTR estão sob pressão de seleção menores do que as sequências codificantes, sofrendo evolução mais rápida. Tais regiões abrigam éxons sem ortologia entre diferentes espécies, indicando uma tolerância maior a variações e participação em mecanismos regulatórios específicos de cada organismo (Resch *et al*, 2009; Fu & Lin, 2012). Apesar da aparente maior tolerância a variação na região 5' UTR, a pato-fisiologia de várias doenças está ligada a mutações na mesma (Cazzola & Skoda, 2000; Reynolds, 2002; Chatterjee & Pal, 2009; Wethmar *et al* 2010), o que ressalta a importância não só de um maior entendimento dos elementos regulatórios conhecidos, mas também da determinação de outros mecanismos ainda não bem entendidos.

3 PROPOSIÇÃO

O objetivo deste trabalho é a descoberta de assinaturas genéticas encontradas no contexto genômico de mutações pontuais de região 5' UTR, causadoras de doenças, que permitam prever o impacto funcional de outras variações pontuais na mesma região. Para isso foram utilizadas sequências dos contextos genômicos flanqueadores de mutações, causadoras de doença, e de polimorfismos, de efeito neutro ou desconhecido, buscando a correlação de características destas sequências com seu efeito funcional. A partir de tais parâmetros, se propõe a geração de modelos matemáticos com o objetivo de predizer o comportamento funcional de outras variações genéticas na região 5' UTR.

4 MATERIAL E MÉTODOS

4.1 Mutações e Polimorfismos

Para este estudo foram utilizadas mutações e polimorfismos pontuais presentes na região 5'UTR, portanto *upstream* ao local de início de tradução do gene. As mutações foram selecionadas do banco de dados do *Human Gene Mutation Database* (HGMD) (Stenson *et al*, 2008, http://www.hgmd.cf.ac.uk/, profissional *trial version*, ver Anexo). Neste estudo entendeu-se como mutações, as variações pontuais para as quais existem publicações reportando-as como causadoras de doenças, conforme anotado no HGMD.

Os polimorfismos foram obtidos do banco de dados NHLBI *Grand Opportunity Exome Sequencing Project* (ESP, https://esp.gs.washington.edu/drupal/), originados do trabalho de Tenessen *et al* (2012). Neste estudo entendeu-se como polimorfismos variações pontuais de caráter funcional desconhecido, mas que provavelmente possuem efeito deletério pequeno ou nulo.

No total foram utilizadas 235 mutações distintas e 21.542 polimorfismos para as análises. Foram utilizadas as posições genômicas das mutações e polimorfismos e as respectivas sequências das formas comum e variante, incluindo 30 pares de base *downstream* e 30 pares de base *upstream* do ponto mutado (montagem genômica GRCh37/hg19), obtidas nos respectivos bancos de dados de mutação e polimorfismos.

4.2 Cálculo de parâmetros bioinformáticos caracterizadores das sequências mutadas e polimórficas

Exceto quando afirmado o contrário, todos os cálculos foram realizados através de programas escritos em linguagem de programação Ruby, desenvolvidos especificamente para este estudo.

4.2.1 Designação dos valores de tipo de variação (tipoV)

A cada mutação e polimorfismo foi designado um valor binário arbitrário tendo como critério a mudança do tipo de base nitrogenada. Consideram-se como transições as mutações nas quais não há mudança do tipo de base (pirimidina ou purina) da base original para a base variante, e transversões aquelas em que o tipo da base variante é diferente do tipo da base original. As transversões alteram mais o caráter físico-químico do ponto mutado, sendo portanto esperado um maior impacto funcional para este tipo de mutação. Para os resultados deste estudo, para as transições foi designado o valor 0 e para as transversões foi designado o valor de 1 para a posterior análise.

4.2.2 Cálculo da diferença da variação da energia livre de Gibbs de *folding* entre as sequências contexto *wild type* e variante (ΔΔG*folding*)

O valor da variação da energia livre de Gibbs (ΔG) entre dois estados distintos de uma mesma molécula é utilizado para inferir o quão mais ou menos estável o estado final é em relação ao estado inicial. Sendo assim o ΔG foi utilizado para inferir a estabilidade termodinâmica de uma determinada sequência de DNA ou RNA, comparando os estados de fita simples com o estado de auto complementariedade, formando uma estrutura secundária ou "*folding"*. Quanto mais negativa a variação de energia livre, menor a energia da forma final, e portanto, maior a espontaneidade da reação e maior a estabilidade do produto, a estrutura secundária. A busca de estruturas secundárias com menores valores possíveis de ΔG constitui a principal forma de previsão das conformações adotadas *in vivo* por sequências de RNA e DNA, assim como com sua estabilidade (Seetin & Mathews, 2012). Com o objetivo de verificar se as variações pontuais modificavam o "enovelamento" *(folding*) do RNA equivalente à variação mais seu contexto genômico, calculou-se a variação do ΔG*folding* entre as sequências do "tipo selvagem" (*wild type*) e as sequências variantes (ΔΔG*folding).*

Para o cálculo do ΔΔG*folding* de mutações e polimorfismos, foi incluída a base variante e 30 bases flanqueadoras de cada lado (totalizando 61 bases). O contexto do genômico (30 nucleotídeos *upstream*, base mutada, 30 nucleotídeos *downstream*) de cada mutação e polimorfismo foi analisado em relação à formação de estrutura secundária. Os valores de ΔG*folding* foram calculadas pelo programa RNAFold (Bompfünewerer *et a*l, 2008). Para isto as sequências, *wild type* e respectiva variante, de 61 pares de bases, foram subdivididas em múltiplas sequências com tamanho variando entre 10 e 30 bases (janelas) contendo todas as sequências contínuas dos fragmentos (como uma janela deslizante, Figura 5) incluindo a base variante. Para cada tamanho de janela *n* foram geradas dessa forma *n* janelas. O valor de ΔG*folding* então foi calculado para cada janela individualmente, obtendo-se assim os valores para as janelas "*wild type"* (ΔG*folding wt*) e para suas respectivas janelas variantes (ΔG*folding var*). O valor de ΔΔG*folding* para cada janela foi calculado segundo a fórmula *ΔΔGfolding = ΔGfolding var - ΔGfolding wt*. Para cada substituição pontual foram analisados os valores máximo (ΔΔGf max), mínimo (ΔΔGf min) e médio (ΔΔGf med) dos valores de ΔΔG*folding* das janelas.

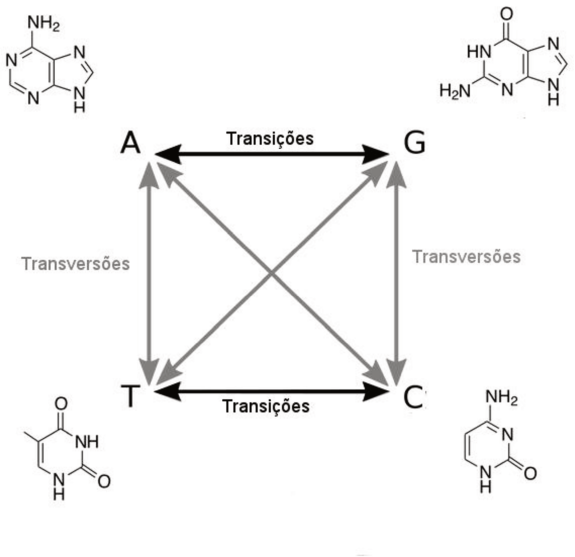

Figura 4 – Esquema de todas as substituições de bases nitrogenadas possíveis no genoma. Transições estão representadas pelas setas escuras. Transversões estão representadas pelas setas claras.

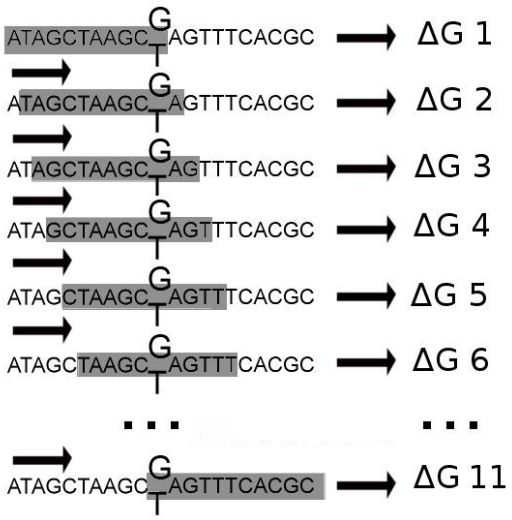

Figura 5 – Esquema representando o método de janelas deslizantes, exemplificando um tamanho de janela de 11 nucleotídeos de comprimento do contexto genômico. Para cada janela (sombreado) calculou-se um valor de energia livre de Gibbs (ΔG).

4.2.3 Designação dos valores para os sítios de ligação de fatores de transcrição alterados (Jaspar_alt) ou criados (Jaspar_crd) pelas variações

Para determinar se as variações se encontravam em sítios (*motifs*) de ligação de fatores de transcrição e os alteravam ou criavam novos sítios de ligação, utilizou-se o banco de dados de perfis de sítios de ligação JASPAR (http://jaspar.genereg.net/) para fatores de transcrição de vertebrados. A partir das matrizes de peso para os sítios de ligação, descritas por Stormo (2006), foram geradas sequências curtas de tamanho variável, formadas pelas bases cujos pesos em cada posição apresentavam um valor mínimo de 70% do valor da base de maior peso (quanto maior o peso, maior a conservação da base naquela posição do *motif*); gerando assim várias combinações possíveis das bases de peso alto. Estas sequências foram então alinhadas com os contextos *wild type* para a identificação de sítios alterados pelas variações pontuais, e com os contextos variantes para a identificação de sítios gerados pelas variações,

determinando-se assim a existência de sequências de possíveis sítios de ligação de transcrição com possível importância funcional decorrente destas variações.

A	0	1	2	2	5	**5**
T	1	2	1	2	**6**	4
C	**5**	2	**8**	3	3	2
G	2	**7**	4	**8**	3	1

Figura 6 – Representação de uma matriz de peso para um *motif* de 6 nucleotídeos de comprimento. Cada linha apresenta valores de peso estatístico da frequência da respectiva base em cada uma das posições do *motif*. O *motif* mais frequentemente encontrado é formado pelas bases de maior valor (sombreadas), CGCGTA neste exemplo.

4.2.4 Cálculo da Distância do Início da Sequência Codificante (DiSC) e do logaritmo natural da Distância do Início da Sequência Codificante (DiSClog)

Para o cálculo da distância da substituição em relação ao sítio de início da sequência codificante (códon de início, ATG) foram utilizadas sequências de DNA complementar dos genes estudados (cDNA) obtidas dos bancos de dados do Biomart Ensembl (Flicek *et al,* 2012, www.ensembl.org/biomart). Além disso sequências 5' UTR adcionais foram obtidas subtraindo-se sequências codificantes, do banco de dados *Consensus Coding Sequences* (Pruitt *et al*, 2009) de transcritos do Biomart do mesmo gene. O contexto genômico ao redor da variação pontual foi alinhado às de 5' UTR obtidas e a distância até a extremidade 3', *downstream,* da sequência UTR foi calculada. Calculou-se também o valor do logaritmo natural do valor de DiSC (DiSClog).

4.2.5 Cálculo da Distância do Início de Transcrição (DiTr)

Para o cálculo da distância da substituição em relação ao sítio de início de transcrição (extremidade 5' da região 5' UTR) foram utilizadas sequências de DNA complementar dos genes estudados (cDNA) obtidas dos bancos de dados do Biomart Ensembl (Flicek *et al,* 2012). O contexto genômico ao redor da variação pontual foi alinhado às sequências dos bancos de dados de 5' UTR e a distância até a extremidade 5', up*stream*, da sequência UTR foi calculada.

4.2.6 Cálculo de valor de conservação filogenética por distância de edição de Levenshtein (Lev)

A distância de edição, também conhecida como distância de Levenshtein, entre duas sequências de caracteres distintos, escritas em um mesmo código, é dada pelo número mínimo de alterações necessárias para que uma das sequências se torne igual a outra, através de inserções, deleções ou substituições individuais de seus caracteres, sendo utilizada assim como uma medida da dissimilaridade de ambas as sequências (Levenshtein, 1966). Para cada sequência *wild type* das variações foram feitas janelas do contexto genômico, centralizadas na base variante, com igual número de bases *upstream* e *downstream,* com tamanho variando de 11 a 29 nucleotídeos de tamanho total. Para cada janela foi então computada a menor distância de edição encontrada em janelas justapostas, de mesmo tamanho, de sequências de regiões 5' UTR de cDNAs de *Rattus Norvegicus* ou *Mus musculus* (de acordo com a disponibilidade das mesmas para o mesmo gene da substituição) obtidas dos bancos de dados do Biomart Ensembl (Flicek *et al*, 2012), sendo assim obtido um valor de distância de edição para cada variação.

4.3 Análise estatística e modelos de regressão logística

As análises e modelos estatísticos foram feitos através de programas escritos na linguagem R. Às substituições pontuais foi designado arbitrariamente um valor binário, sendo o valor 1 para as mutações e o valor 0 para os polimorfismos; cada valor foi associado aos seus respectivos parâmetros bioinformáticos já calculados (DiSC, DISClog, DiTr, ΔΔGf max, ΔΔGf min, ΔΔGf med, tipoV, Jaspar_alt, Jaspar_crd, Lev). Alguns dos parâmetros só puderam ser calculados para determinadas subamostras de todas as variações obtidas, pois dependeram do alinhamento do contexto genômico à sequências anotadas como região 5' UTR humanas (para DiSC, DISClog e DiTr), ou de rato ou camundongo (Lev. Desta forma, as análises foram realizadas de maneira diferenciada entre a amostra completa, e entre 2 subamostras distintas.

Análise 1: realizada para a amostra completa, com n=235 mutações e n=21.542 polimorfismos distintos. Para este grupo foram realizadas as análises dos parâmetros dependentes apenas do contexto genômico das variantes, ou seja, calculáveis para todas as variações do estudo. Os parâmetros analisados foram ΔΔGf max, ΔΔGf min, ΔΔGf med, tipoV, Jaspar_alt e Jaspar_crd.

Análise 2: composto por um subconjunto do grupo 1, com n=154 mutações e n=19.597 polimorfismos distintos. Para este grupo foram feitas as análises dos parâmetros que dependiam da existência de sequências anotadas como a região 5' UTR humana do gene no qual a variação mais seus contextos poderiam ser encontradas, no banco de dados do Biomart. Os parâmetros analisados foram DiSC, DISClog e DiTr, sem levar em consideração os parâmetros da análise 1.

Análise 3: composto por um subconjunto do grupo 1 e diferente do grupo 2, com n=171 mutações e n=16.436 polimorfismos. Para este grupo foram feitas as

análises dos parâmetros que dependiam da existência de sequências anotadas como região 5' UTR de rato ou camundongo do gene no qual a variação mais seus contextos poderiam ser encontradas, no banco de dados do Biomart. O parâmetro analisado foi o Lev, sem levar em consideração os parâmetros da análise 1.

Além disso, foram realizadas análises dos parâmetros da análise 1 juntamente aos respectivos parâmetros das variações das análise 2 e 3. Estes resultados só tiveram importância para a criação dos modelos de regressão logística multivariáveis.

Análise 1+2: Composto pelas variações do grupo da análise 2, analisando-se os parâmetros considerados para as análises 1 e 2, ou seja: ΔΔGf max, ΔΔGf min, ΔΔGf med, tipoV, Jaspar_alt, Jaspar_crd, DiSC, DISClog e DiTr

Análise 1+3: Composto pelas variações do grupo da análise 3, analisando-se os parâmetros considerados para as análises 1 e 3, ou seja: ΔΔGf max, ΔΔGf min, ΔΔGf med, tipoV, Jaspar_alt, Jaspar_crd, e Lev

Análise 1+2+3: Composta apenas pelas variações para as quais foi possível o cálculo de todos os parâmetros das análises 1, 2 e 3, com n=121 mutações e n=15.319 polimorfismos. Desta forma analisaram-se os parâmetros ΔGf max, ΔΔGf min, ΔΔGf med, tipoV, Jaspar_alt, Jaspar_crd, DiSC, DISClog, DiTr e Lev.

A diferença estatística entre as mutações e polimorfismos dos parâmetros bioinformáticos, em suas respectivas análises 1, 2 e 3, foi verificada pelo teste U de *Man-Whitney,* para os valores de ΔΔGfmax, ΔΔGfmin, ΔΔGfmed (análise 1), DiSC, DiSClog, DiTr (análise 2), e Lev (análise 3), e pelo teste binomial exato para os valores de tipoV. Jaspar_alt e Jaspar_crd (análise 1).

Para todas as análises construíram-se então modelos de regressão logística univariáveis e multivariáveis considerando todos os parâmetros calculados. Para cada modelo logístico uma função logística foi gerada, com os coeficientes dos parâmetros com menores valores P possíveis. Uma função logística é uma curva sigmoide com valor variando entre 0 e 1 no eixo das ordenadas (eixo y), sendo tal valor a probabilidade de um dado evento ocorrer de acordo com os valores das variáveis independentes da função. Como evento, considerou-se a caracterização da variação pontual como uma mutação, atribuindo-se dessa forma valor 1 para as mutações e valor 0 para os polimorfismos para a regressão dos modelos. Assim os modelos gerados tiveram como objetivo prever a probabilidade de uma dada variação pontual ser uma mutação de acordo com o valor dos parâmetros preditores mais estatisticamente significativos. Os parâmetros escolhidos para os melhores modelos de regressão logística múltipla foram aqueles que apresentaram valor p inferior a 0,05 para os testes de diferenças estatísticas (teste de Wilcoxon e teste Binomial, valor p de seu coeficiente inferior a 0,05 nos seus respectivos modelos de regressão logística simples (nas análises 1,2 e 3), e que mantiveram o valor p dos coeficientes de regressão inferior a 0,05 nos modelos de regressão logística múltipla (nas análises 1,2, 3, 1+2, 1+3 e 1+2+3).

O poder explicativo dos modelos foi inferido através do uso de um gráfico de Curvas de Características de Operação do Receptor (ROC- *Receiver Operating Characteristic*). A curva ROC permite a estimativa da performance da regressão logística ao descrever a capacidade de previsão de um modelo de inferir a ocorrência ou a não ocorrência de um determinado evento, plotando a correspondência entre os valores observados e os valores esperados, de forma a gerar uma curva da razão dos valores de verdadeiros positivos contra a razão dos falsos positivos. A área abaixo da curva permite inferir o poder explicativo do modelo, sendo maior o quão melhor for o modelo, variando entre 0,5 (50% de probabilidade de verdadeiro positivos e falsos positivos, um modelo puramente aleatório) e 1,0 (100% de verdadeiros positivos, um modelo "perfeito") (Mason & Grahan, 2002).

5 RESULTADOS

5.1 Análises estatísticas

A Tabela 1 lista os valores p obtidos nos testes U de Wilcoxon ou Binomial paras as análises estatísticas individuais dos parâmetros analisados (sendo "*m*e^{n}" equivalente a "*m* vezes 10 elevado a *n* potencia", como em 2,2e^{-16}), isto é, os menores valores de probabilidade de se rejeitar a hipótese nula (o parâmetro comporta-se igualmente para todas as variações) sendo esta verdadeira. Segundo o teste U de Wilcoxon, apresentaram significância estatística (para $\alpha = 0,05$) os parâmetros ΔΔGfmax (análise 1), DiSC, DiSClog (análise 2) e Lev (análise 3), e segundo o teste binomial exato apresentou significância estatística (para $\alpha = 0,05$) o parâmetro tipoV. Estes parâmetros apresentaram comportamento distintos entre o grupo das mutações e o grupo dos polimorfismos.

5.1.1 Parâmetros das Diferenças de energia livre de Gibbs de *folding* (ΔΔG*folding)*

A distribuição dos parâmetros, ΔΔGfmin, ΔΔGfmed, ΔΔGfmax pode ser visualizada nos gráficos da Figura 7. Para estas análises utilizaram-se os dados para uma janela deslizante de 28 nucleotídeos de comprimento, que se mostraram os mais significativos estatísticamente. Valores de janelas menores ou maiores mostraram pouca capacidade de distinção entre as formas *wild type* e variantes das mutações e polimorfismos, sendo portanto pouco informativas para a visualização de tendências que distinguissem os dois grupos. Para ambos os grupos os valores de ΔΔGfmed se situaram próximos a 0, indicando claramente pouca ou nenhuma diferença (*ΔGfolding var - ΔGfolding wt*) média na estabilidade termodinâmica entre as janelas equivalentes nos contextos genômicos das formas *wild type* e variante, dificultando a distinção do

comportamento dos parâmetros de ΔΔG*folding* entre as variantes mutações e as variantes polimorfismos.

Tabela 1 – Valores p dos parâmetros bioinformáticos obtidos para suas análises estatísticas e para seus coeficientes de regressão logística uni variável. * valor estatisticamente significativo para α = 0,05; ** estatisticamente significativo para α = 0,01; *** estatisticamente significativo para α = 0,001. "*m*e*n*" equivale a "*m* vezes 10 elevado a *n* potencia".

Análise	Parâmetro	Valor p (Wilcoxon)	Valor p (Binomial)	Valor p (Coeficiente de regressão logística univariável)
1	ΔΔGfmin	0,414	-	0,716
	ΔΔGfmed	0,3111	-	0,547
	ΔΔGfmax	0,009997 **	-	0,0385 *
	tipoV	-	0,04377 *	0,000807 ***
	Jaspar_alt	-	0,09842	0,191
	Jaspar_crd	-	0,9002	0,647
2	DiSC	$< 2,2e^{-16}$ ***	-	$2,74e^{-16}$ ***
	DiSClog	$< 2,2e^{-16}$ ***	-	$< 2,0e^{-16}$ ***
	DiTr	0,1511	-	0,188
3	Lev	$2,727e^{-8}$ ***	-	$4,45e^{-6}$ ***

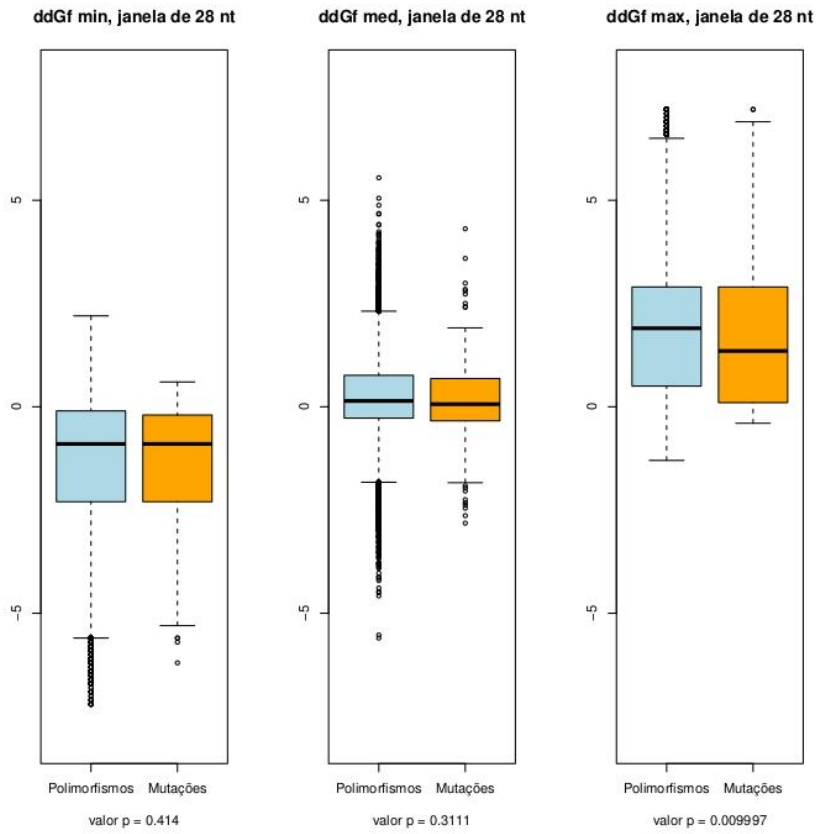

Figura 7 – Gráfico de caixa mostrando a distribuição empírica dos valores para os parâmetros ΔΔGfmin, ΔΔGfmed, ΔΔGfmax (n mutações =235, n polimorfismos =21.542). Valores p para teste U de Wilcoxon.

Metade dos valores de ΔΔGfmin, isto é, a menor diferença encontrada entre janelas correspondentes do contexto variante e *wild type,* apresentaram-se entre 0 e -2 kcal/mol, tanto no grupo das mutações como no grupo dos polimorfismos, indicando que a maioria das variações possuia janelas nas quais a sequência variante se apresentava como a mais estável (*ΔGfolding var < ΔGfolding wt,* ΔΔG*folding* < 0), porém com uma diferença de estabilidade pequena. Uma pequena parte das variações, 1/4, entretanto, nunca apresentaram janelas nas quais a variante se apresentava como mais estável (*ΔGfolding var > ΔGfolding wt*, ΔΔG*folding > 0*), sendo nestes casos a sequência *wild type* mais estável para todas as janelas. O contrário também foi observado nos valores de ΔΔGfmax em ambos os grupos, sendo metade dos valores das maiores diferenças encontradas, positivas (*ΔGfolding var > ΔGfolding wt*, ΔΔG*folding >*

0), entre 0 e +2 kcal/mol, sendo estas janelas mais estáveis para as sequências *wild type*. Foram encontradas em torno de 1/4 de variações nas quais a sequência variante sempre se apresentava como a mais estável (*ΔGfolding var* < *ΔGfolding wt,* ΔΔG*folding* < 0) em todas as janelas. Desta forma, de uma maneira geral, tanto em mutações quanto em polimorfismos, encontram-se algumas janelas nas quais a sequência variante é mais estável e algumas janelas nas quais a sequência *wild type* é mais estável. No total, observou-se em torno de 1/4 das variações puramente mais estável para todas as janelas variantes e aproximadamente 1/4 das variações puramente mais estável para todas as janelas *wild type.*

Nota-se, porém no parâmetro ΔΔGfmax uma leve tendência no grupo das mutações para valores menos positivos e uma maior proporção de variantes completamente mais estáveis (*ΔGfolding var* < *ΔGfolding wt,* ΔΔG*folding* < 0, para todas as janelas), 25 %, em comparação a 15% para os polimorfismos. Isto se reflete no valor p do teste de Wilcoxon (Tabela 1), significativo para α = 0,01 e pode ser visualizado no histograma da distribuição percentual dos valores de ΔΔGfmax (Figura 8). Desta forma há uma tendência, mesmo que pequena, de estruturas secundárias variantes completamente mais estáveis no grupo das mutações.

5.1.2 Tipo de Variação (tipoV)

As proporções de transversões e transições para os grupos das mutações e para os grupos dos polimorfismos estão demonstradas no gráfico de barras da Figura 9. Para os polimorfismos verificou-se que 33,618% eram transversões e 66,382% eram do tipo menos drástico, transições. Já no grupo das mutações verificou-se 44,068% de transversões e 55,932% de transições, mostrando uma quantidade significativamente maior (teste binomial, α = 0,05, Tabela 1). Portanto uma maior proporção de transversões, variações com maior diferença de caráter físico-químico entre a base original e a variante, foi encontrada dentre as mutações, causadoras de doenças.

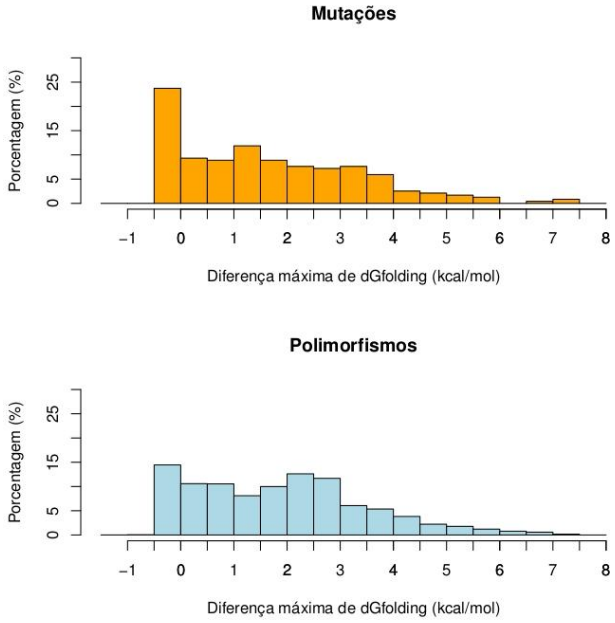

Figura 8 – Histogramas mostrando a distribuição percentual para o parâmetro de diferença máxima de *folding* (ΔΔGfmax), n mutações =235, n polimorfismos =21.542.

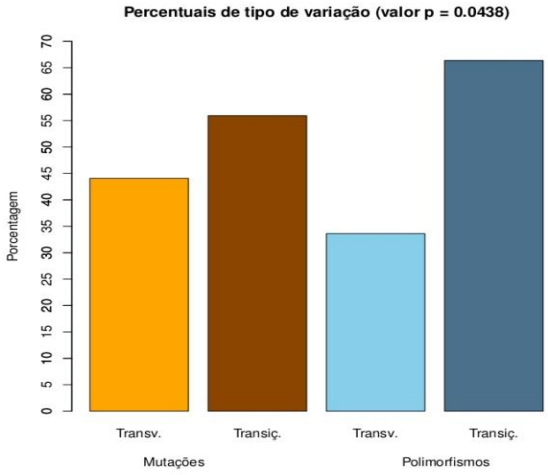

Figura 9 – Gráfico de barras mostrando os percentuais dos diferentes tipos de variação dentro dos grupos das mutações e dos polimorfismos (n mutações =235, n polimorfismos =21.542). Valor p referente ao teste binomial exato.

5.1.3 Sequências Jaspar alteradas ou criadas (Jaspar_alt e Jaspar_crd)

Apesar da variedade de sequências possíveis de sítios de ligação de fatores de transcrição, 4342 sequências distintas, variando de 6 a 30 nucleotídeos de comprimento, abrangendo sequências encontradas em diversos vertebrados, apenas uma pequena porcentagem tanto das mutações quanto dos polimorfismos apresentou variações que alteravam ou que criavam tais sequências.

Sequências Jaspar alteradas (Jaspar_alt) na forma variante foram encontradas em apenas 5.578% dos polimorfismos e 8.439% das mutações, indicando uma leve tendência para uma maior proporção no grupo das mutações, porém não significativa estatisticamente (valor p = 0,09842, teste binomial, Tabela 1). Sequências Jaspar criadas pelas variantes ocorreram em 6.781% dos polimorfismos e 6.751% das mutações, não mostrando nenhuma tendência significativa (valor p = 0,9002, teste binomial, Tabela 1).

5.1.4 Parâmetros das Distâncias de Início da Sequências Codificantes (DiSC e DiSClog)

Para os valores de DiSC foram obtidos vários valores para cada variação, resultantes dos alinhamentos de um mesmo contexto para diferentes transcritos de região 5' UTR presentes no Biomart. A distribuição empírica das distâncias encontradas para as variantes está representada na Figura 10. O gráfico considerando todas os valores de DiSC encontrados para as variações (Figura 10, esquerda, n mutações = 180, n polimorfismos = 25.647) revela uma grande diferença para os valores entre as mutações e os polimorfismos. Mais de 75% dos polimorfismos se concentraram em regiões inferiores a 50 nucleotídeos de distância do início da sequência codificante. As mutações, por outro lado foram observadas mais bem distribuídas, com 50% distribuída na faixa de 50 a 150 nucleotídeos de distância do início da sequência codificante, quase

que de maneira exclusiva às distâncias ocupadas pelos polimorfismos, e com 25%, abrangendo a região de 150 a 350 nucleotídeos de distância da sequência codificante.

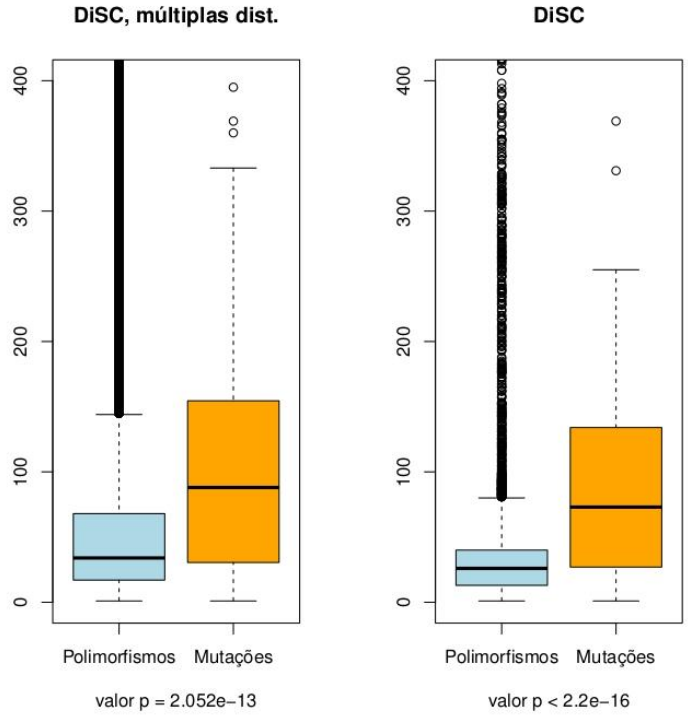

Figura 10 – Gráfico de caixa para os valores das Distâncias do início da Sequência Codificante (DiSC), O gráfico da esquerda apresenta os valores contendo as múltiplas distâncias encontradas em diversos transcritos para uma mesma variação (n mutações = 180, n polimorfismos = 25.647). O gráfico da direita apresenta apenas o valor da menor distância encontrada para cada variação (n mutações = 154, n polimorfismos = 19.597). Valores p referentes ao teste U de Wilcoxon.

A diferença entre os grupos se mostrou altamente significativa (valor p = $2{,}052e^{-13,}$ para $\alpha = 0{,}001$). A significância aumentou mais ainda quando se utilizou apenas dos menores valores de DiSC encontrados para as variações, únicos para cada variação (Figura 10, direita, n mutações = 154, n polimorfismos = 19.597), com valor p < $2{,}2e^{-16}$, também mostrado na Tabela 1, sendo que as mesmas observações se mantiveram de

maneira geral nesta análise, com exceção de uma esperada diminuição dos valores superiores para a faixa de 50 a 250 no grupo das mutações e uma concentração ainda maior dos valores faixa abaixo de 50 nucleotídeos de distância no grupo dos polimorfismos na. Para a geração dos modelos de regressão, utilizaram-se parâmetros gerados a partir dos menores valores de DiSC para cada variação, devido ao melhor desempenho estatístico. A Figura 11 permite a visualização ampla da distribuição percentual dos valores dos menores de DiSC (gráficos superiores) e de todos os valores de DiSC (gráficos inferiores).

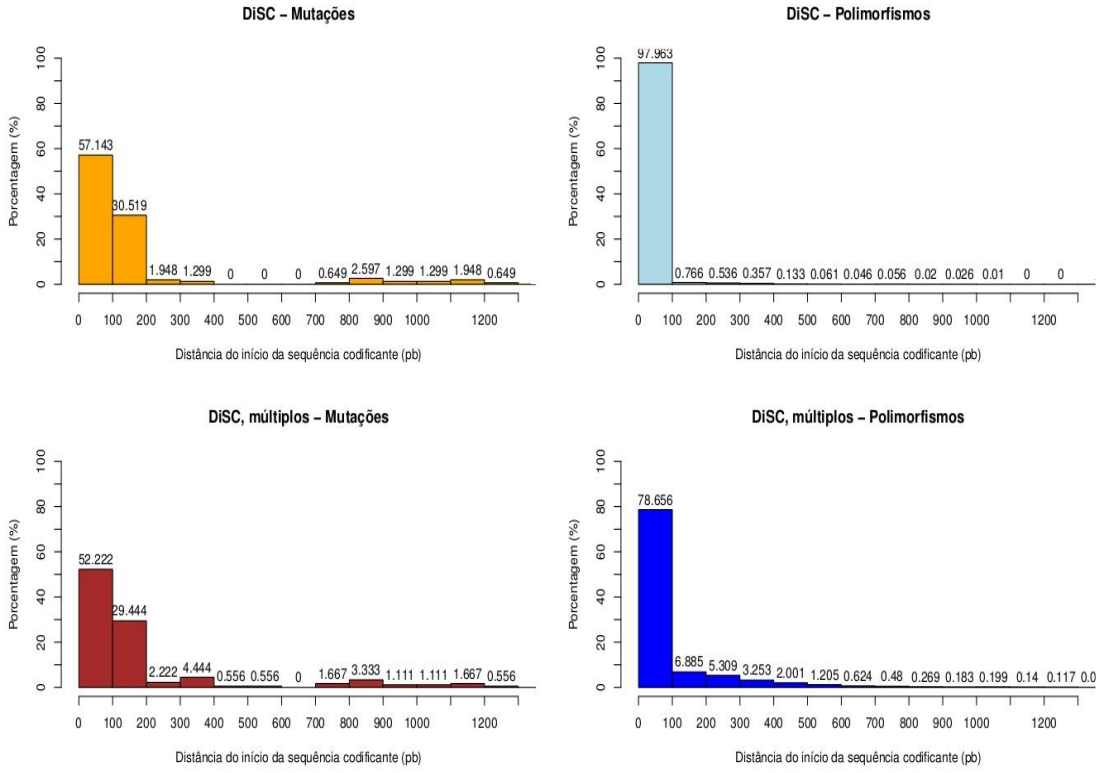

Figura 11 – Histogramas da distribuição dos valores de DiSC. Os gráficos superiores mostram a distribuição dos menores valores, únicos, para cada variação (n mutações = 154, n polimorfismos = 19.597), enquanto que os gráficos inferiores mostram a distribuição de todos os valores encontrados para cada variação pontual (n mutações = 180, n polimorfismos = 25.647). Os números sobre as barras indicam os valores percentuais.

Nota-se tanto nos gráficos de menores valores, quanto nos gráficos de valores múltiplos, a concentração dos valores de DiSC inferiores a 100 nucleotídeo no grupo

dos polimorfismos e a existência de uma ligeira concentração de mutações com valores de DiSC entre 700 e 1300 nucleotídeos, tanto para os menores valores de DiSC encontrados quanto para todos os valores encontrados, o que não se verifica em nenhuma situação para o grupo dos polimorfismos.

Os histogramas da Figura 12 permitem a visualização detalhada da distribuição para os menores valores de DiSC, únicos para cada variação. É possível notar uma proporção aumentada de mutações com valores de DiSC na faixa de 100 a 150 nucleotídeos (22,727%), além da predominância de valores inferiores a 50 nucleotídeos (41,558%), porém não tão grande quanto no grupo dos polimorfismos (90,683%).

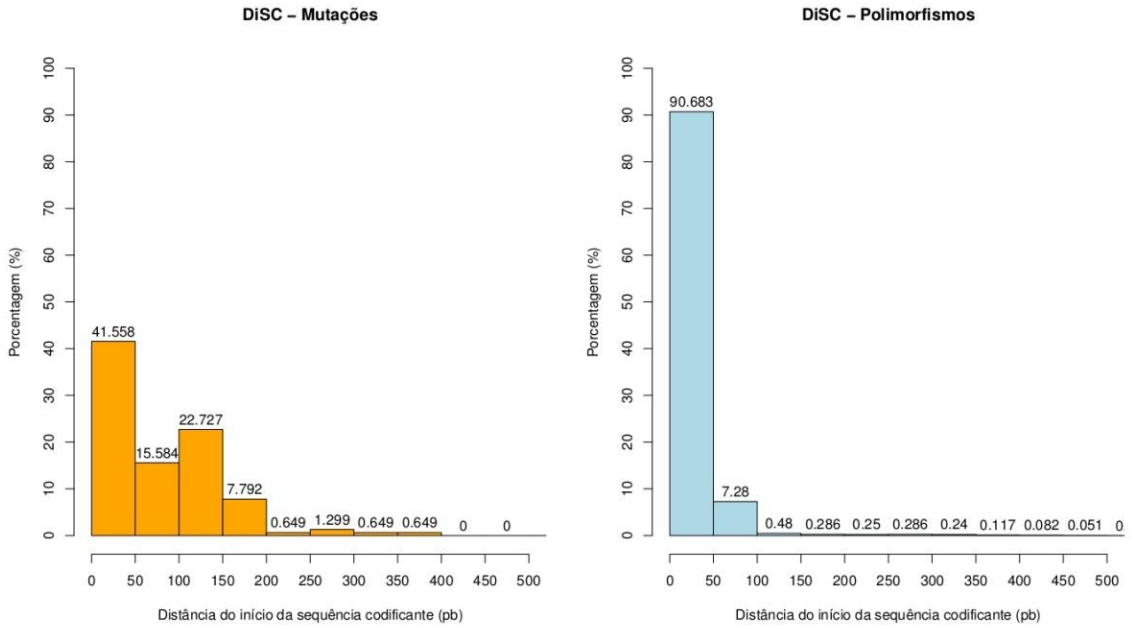

Figura 12 – Histogramas da distribuição dos menores valores de DiSC, únicos para cada variação (n mutações = 154, n polimorfismos = 19.597). Os números sobre as barras indicam os valores percentuais.

Buscando-se a diminuição do impacto estatístico de valores discrepantes, calculou-se os valores dos logaritmos naturais dos valores de DiSC (DiSClog). Foram calculados e analisados os dados para todos os valores DiSC e para os menores valores DiSC, únicos para cada variação. Devido a número amostral reduzido de mutações, muitas variações pertenciam ao mesmo gene e portanto havia a possibilidade de

variabilidade restrita entre tais variações e de um desvio dos valores de DiSC para os intervalos ocupados por muitas variações de um mesmo gene. Para verificar esta possibilidade, analisou-se também o valor de DiSClog para variações únicas por gene, selecionadas aleatoriamente (n mutações = 82, n polimorfismos = 8.549). Os resultados para as diversas análises de DiSClog encontram-se na Figura 13.

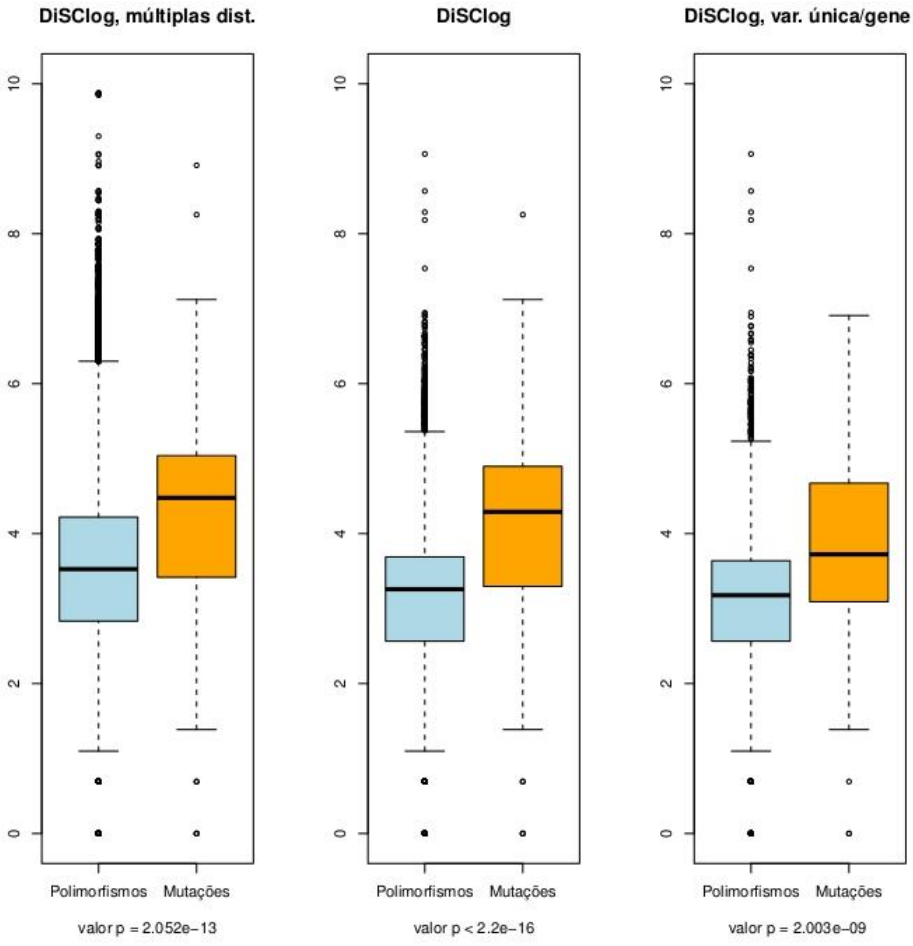

Figura 13 – Gráfico de Caixa para a distribuições dos valores calculados do logaritmo natural de DiSC (DiSClog). O gráfico da esquerda mostra a distribuição para todos os valores de DiSClog encontrados para as variações (n mutações = 180, n polimorfismos = 25.647). O gráfico central mostra a distribuição apenas para os menores valore de DiSClog, únicos por variação (n mutações = 154, n polimorfismos = 19.597). O gráfico da direita mostra os valores de DiSClog de apenas uma variação por gene, escolhidos aleatoriamente (n mutações = 82, n polimorfismos = 8.549). Os valores de p para teste U de Wilcoxon.

Novamente observa-se uma diferença altamente significativa ($\alpha = 0{,}001$) entre os grupos das mutações e polimorfismos (Figura 13, esquerda, n mutações = 180, n polimorfismos = 25.647). Tal tendência se tornou mais significativa para os valores menores de DiSClog (Figura 13, centro, n mutações = 154, n polimorfismos = 19.597), únicos por variação e teve redução de significância para a análise apenas de variações únicas por gene (Figura 13, direita, n mutações = 82, n polimorfismos = 8.549), porém não o suficiente para anular a diferença estatística encontrada entre os grupos. Para as demais análises de regressão de DiSClog foram utilizados os valores menores, únicos por variação, com diversas variações de um mesmo gene (n mutações = 154, n polimorfismos = 19.597), devido ao melhor desempenho estatístico.

5.1.5 Distância do início da Transcrição (DiTR)

Não se verificou diferença significativa entre os grupos das mutações e polimorfismos em relação à distância do início de transcrição (Tabela 1). Para as sequências estudadas verificou-se um valor médio ao redor de 50 nucleotídeos de distância da extremidade 5' dos transcritos, conforme a Figura 14.

5.1.6 Conservação filogenética por distância de Levenshtein (Lev)

Para as análises foram utilizados os dados para janela de 23 nucleotídeos de extensão, 11 de cada lado da base variável, devido ao seu melhor desempenho estatístico. Encontrou-se diferença altamente significativa ($\alpha = 0{,}001$, Tabela 1) entre os grupos, conforme ilustrado na Figura 15, com valores variando de 0 a 13 para as mutações e de 0 a 15 para os polimorfismos. Verificaram-se valores de distância de edição maiores dentre as mutações, 50% apresentando distâncias entre 6 e 10, sugerindo uma tendência destas substituições se localizarem em pontos menos conservados entre o ser humano e ratos e camundongos.

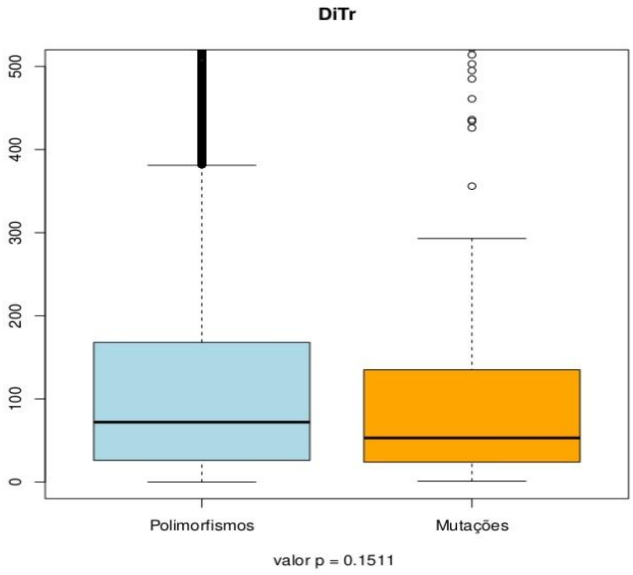

Figura 14 – Gráfico de caixa mostrando a distribuição empírica do parâmetro de Distância do início da Transcrição (DiTr, n mutações = 154, n polimorfismos = 19.597). Valor p para teste U de Wilcoxon.

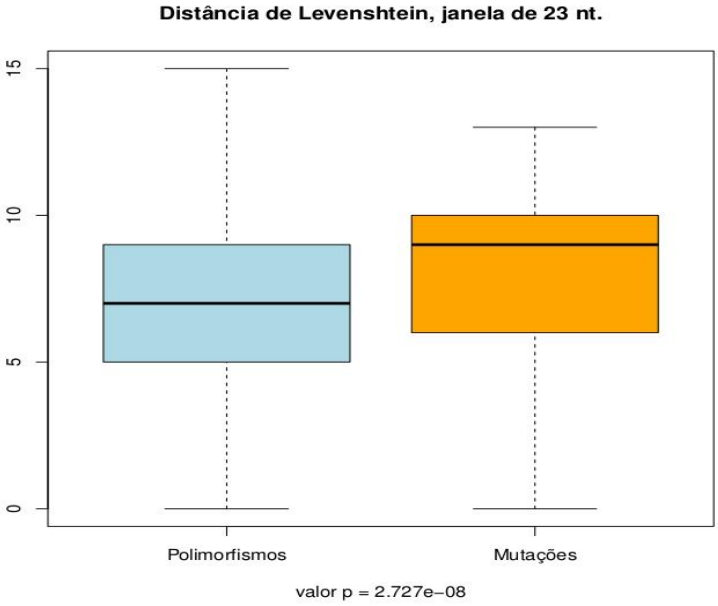

Figura 15 – Gráfico de caixa mostrando a distribuição empírica do parâmetro de Distância de edição de Levenshtein (Lev), para janelas de 23 nucleotídeos de extensão (n mutações = 171, n polimorfismos = 16.436). Valor p para teste U de Wilcoxon.

5.2 Modelos de regressões logísticas univariáveis e multivariáveis

Os valores p dos coeficientes de regressão logística univariáveis estão anotados na Tabela 1. Para todas as regressões logísticas univariáveis e multivariáveis realizada o valor p do intercepto foi $<2e^{-16}$, altamente significativos para $\alpha = 0{,}001$. Mostraram-se como parâmetros de coeficientes estatisticamente significativos o parâmetro ΔΔGfmax, para $\alpha = 0{,}05$, e altamente significativos tipoV, DiSC, DiSClog e Lev para $\alpha = 0{,}001$. Tais critérios foram também tidos como estatisticamente discriminantes para os grupos de mutações e polimorfismos nos testes de Wilcoxon e binomial (Tabela 1), e por isso foram selecionados para a geração de modelos de regressão logística multivariáveis e verificação do desempenho dos mesmos através da curva ROC. Desta forma seis modelos de regressão foram testados:

Modelo 1: $P_{mutação} = a_1 + b_1(\Delta\Delta Gfmax) + c_1(tipoV)$

Modelo 2: $P_{mutação} = a_2 + b_2(DiSClog)$

Modelo 3: $P_{mutação} = a_3 + b_3(Lev)$

Modelo 1+2: $P_{mutação} = a_{12} + b_{12}(\Delta\Delta Gfmax) + c_{12}(tipoV) + d_{12}(DiSClog)$

Modelo 1+3: $P_{mutação} = a_{13} + b_{13}(tipoV) + c_{13}(Lev)$

Modelo 1+2+3: $P_{mutação} = a_{123} + b_{123}(\Delta\Delta Gfmax) + c_{123}(tipoV) + d_{123}(DiSClog) + e_{123}(Lev)$

Onde $P_{mutação}$ é a probabilidade de uma variação ser uma mutação de acordo com as variáveis do modelo; a_n é intercepto da função logistica e b_n, c_n, d_n, e_n são os coeficientes para os parâmetros entre parênteses. Cada modelo foi gerado a partir de suas análises correspondentes. Através do modelo 1 verificou-se o desempenho dos parâmetros ΔΔGfmax e tipoV com o maior número amostral possível. Para o modelo 2 levou-se em consideração apena o parâmetro DiSClog, por este ser um equivalente a DiSC de desempenho estatístico melhor. Os modelos 2 e 3 foram gerados para a análise individualizada da importância de seus respectivos parâmetros (DiSClog e Lev), devido

à sua grande significância estatística nos testes (α = 0,001). Os modelos 1+2 e 1+3 foram gerados para se ver a interação de seus parâmetros enquanto se mantinha o maior número amostral possível. No caso do modelo 1+3, este originalmente incluía o parâmetro ΔΔGfmax, porém o mesmo obteve coeficiente não significativo para α = 0,05 e foi descartado. O modelo 1+2+3 mostrou pouco informativo, pois a sua grande redução amostral causou a perda de significância para α = 0,05 do parâmetro Lev, que foi descartado, tornando-se uma versão piorada do modelo 1+2+3 e portanto não mais analisado neste trabalho. Os valores p para os interceptos e coeficientes estão na Tabela 2.

Tabela 2 – Valores p para o intercepto e os coeficientes dos modelos de regressão logística. * valor estatisticamente significativo para α = 0,05; ** estatisticamente significativo para α = 0,01; *** estatisticamente significativo para α = 0,001.

Modelos	Intercepto	ΔΔGfmax	tipoV	DiSClog	Lev
1		0,016532*	0.000343***	-	-
2		-	-	$<2e^{-16}$***	-
3	$<2e^{-16}$***	-	-	-	$4,45e^{-6}$***
1+2		0,002215**	0,000924***	$<2e^{-16}$***	-
1+3		-	0,00654**	-	$2,37e^{-6}$***

O desempenho dos modelos pode ser visualizado na Figura 16. De uma maneira geral, o desempenho dos modelos se mostrou proporcional ao valor p dos coeficientes dos parâmetros que os compunham. Parâmetros altamente significativos estatisticamente apresentaram maior contribuição explicativa, apresentado maiores áreas abaixo da curva. Os modelos 1, 3 e 1+3 apresentaram baixo poder explicativo com áreas debaixo da curva de 0,579, 0,622 e 0,632 respectivamente, indicando o baixo poder preditor dos parâmetros ΔΔGfmax (Max), tipoV e Lev para a determinação da probabilidade de uma variação ser uma mutação. O modelo apresentou área de curva de

0,752, mostrando um poder preditivo moderado para o parâmetro DiSClog. O resultado obtido pelo modelo 1+2 foi ligeiramente melhor, com área de 0,772, suferindo uma grande contribuição explicativa apenas do parâmetro DiSClog. Desta forma o modelo 1+2 possibilita a previsão do comportamento de uma variação pontual, ou seja, a probabilidade de que ela possua um impacto funcional comparável ao de uma mutação causadora de doença, com 77,2% de chance de retornar um valor verdadeiramente positivo.

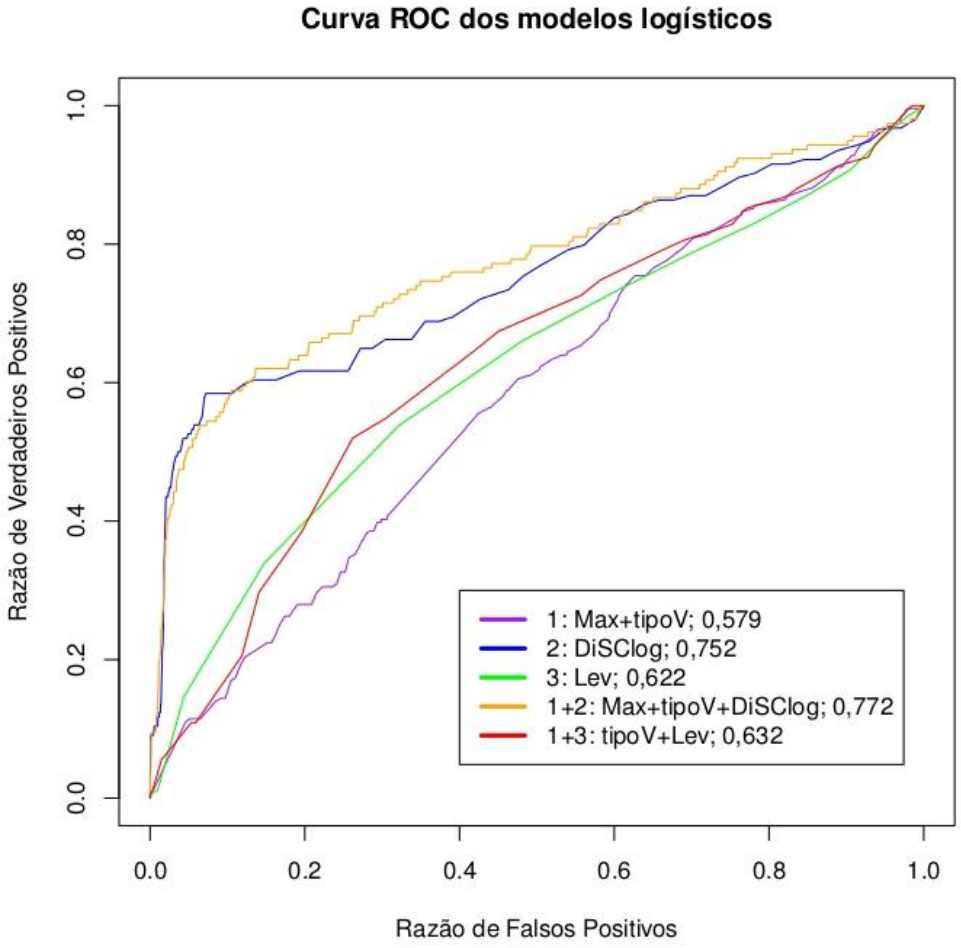

Figura 16 – Gráfico de Curvas de Características de Operação do Receptor, mostrando estimativas dos desempenhos dos modelos de regressão logística. A legenda mostra o número do modelo, seguido dos respectivos parâmetros preditores do modelo e da área debaixo da curva, indicando o desempenho do modelo.

6 DISCUSSÃO

Os resultados permitem a identificação de 4 parâmetros bioinformáticos distintivos entre os grupos de mutações patogênicas e polimorfismos de funcionalidade desconhecida, com diferentes poderes explicativos: ΔΔGfmax, tipoV, Lev e DiSClog. Devido a natureza puramente estatística do estudo é importante ressaltar a impossibilidade do estabelecimento de uma relação direta de causa e consequência entre os diferentes parâmetros e o efeito funcional conhecido da variação simples. Desta forma, pode se esperar dois tipos diferentes de parâmetros preditores: parâmetros que refletem uma contribuição diretamente causal para o efeito funcional da variação pontual, e parâmetros que refletem uma tendência mais ampla, de que determinados tipos funcionais de variações se apresentem concentrados em regiões do corpo gênico com determinadas propriedades estatísticas, por razões não causalmente relacionadas com a mesma e que ainda não foram esclarecidas.

A análise dos recentes estudos de mapeamento da variabilidade do genoma humano (Abecasis *et al*, 2012; MacArthur *et al*, 2012; Tennessen *et al*, 2012) mostra um comportamento interessante do genoma humano em relação a variantes funcionais de perda de função: O genoma humano se apresenta purificado de variações de perda função altamente deletérias, incluindo mutações com efeito patogênico. Tais variações são eliminadas ao longo da evolução, através da seleção natural, ao reduzirem drasticamente a viabilidade dos indivíduos portadores, sendo encontradas entre as variações de baixíssima frequência e de pouco tempo evolutivo na população humana. A mesma lógica evolutiva pode ser esperada no que se diz respeito ao efeito funcional de assinaturas genéticas (parâmetros estatísticos bioinformáticos): o genoma humano provavelmente se encontra purificado de características sequenciais que tenham por si só um efeito drástico e permitam a perda completa da função de um gene independentemente de outros fatores. Para que uma assinatura genética funcional deletéria seja observável no contexto de variações em quantidade amostral suficiente esta não deve interferir com os mecanismos regulatórios que normalmente regem a expressão gênica a ponto de tornar o funcionamento do organismo inviável e deve

vencer mecanismos naturais de reparo do DNA (Shih *et al*, 2012). Espera-se que falhas funcionais altamente deletérias no genoma sejam eliminadas ao longo da evolução. Da mesma forma, espera-se que assinaturas genéticas tenham um sucesso apenas parcial em prever o efeito funcional de uma variação pontual, atuando em conjunto com outras assinaturas ou sendo neutralizadas pelos mecanismos regulatórios normais ou alternativos da expressão gênica.

Apesar do desempenho de previsão limitado, o parâmetro de ΔΔGfmax parece ter um efeito funcional causal em pelo menos parte das mutações. Pelos resultados verificou-se pouca importância funcional de interações de auto complementariedade de curta distância (28 nucleotídeos) no contexto das variações, tanto nos polimorfismos quanto nas mutações. Tal resultado não é de todo inesperado, uma vez que já se sabe da capacidade da maquinaria de inicialização e da subunidade ribossomal 43S de varrer sobre estruturas secundárias com valores de ΔG inferiores e muito mais complexas do que as pequenas estruturas formadas pelas janelas curtas dos contextos das variações (Spruill & McDermott, 2009; Andreev *et al,* 2009; Vassilenko *et al*, 2011). Estas dificilmente ultrapassaram o valor de ΔG de -10 kcal/mol, sendo limitadas e pouco diferenciais mesmo entre os contextos *wild type* e variante. O percentual aumentado de contextos variantes mais estáveis que o *wild type* dentre o grupo das mutações indica, entretanto, que para um percentual considerável, de 25%, estas estruturas secundárias pequenas não só são distintas o bastante no contexto variante, mas também parecem ter um efeito funcional considerável ao aumentarem a estabilidade da estrutura secundária local, como mostram os resultados obtidos para o parâmetro de ΔΔGfmax e seu poder de explicar uma pequena parcela das mutações. Os efeitos deletérios destas pequenas estruturas possivelmente estão relacionados a outros elementos regulatórios próximos do contexto alterado pelas mesmas. Possíveis candidatos seriam sítios de ligação de RBPs ou de RNAs não codificantes ou outros elementos de ação localizada (Kumar *et al*, 2012, Abaza & Gebauer, 2008), sendo necessário um melhor conhecimento da interação de tais elementos com pequenas estruturas no mRNA.

Sob o ponto de vista funcional a caracterização dos tipos de variação se mostrou demasiadamente simples para o estabelecimento de qualquer hipótese de relação causal. Porém os resultados obtidos para parâmetro tipoV permitiram a distinção dos grupos.

Foram encontradas proporções de transversões idênticas às encontradas pelo trabalho de Shih *et al* (2012), sendo a proporção em torno de 33% de transversões para polimorfismos de caráter funcional neutro ou variado e de 44% dentre mutações causadoras de doenças. O mesmo trabalho mostra que, independentemente do tipo de variação, as variações patogênicas observadas são na maioria dos casos as que menos alteram as características eletrônicas de seu contexto próximo (transferência de carga do intervalo de 20 a 60 nucleotídeos ao redor da variação pontual), quando comparadas com as outras 2 possíveis variações de base no mesmo ponto. Tal estudo sugere que tanto variações patogênicas quanto variações neutras sejam pouco detectáveis pelos mecanismos de reparo de erros no DNA, de modo que alterações drásticas sejam detectadas e removidas.

Existe a possibilidade da maior proporção de transversões dentre as variações pontuais patogênicas serem um indicativo não de seu efeito funcional, mas sim dos processos mutagênicos envolvidos na sua geração e disseminação. Diversos trabalhos mostram que a mutagênese, longe se ser um evento puramente aleatório, acontece de maneira direcionada em certas sequências ou arquiteturas cromossomais, criando os chamados *hotspots* de mutações (Walser & Furano, 2010; Cooper et al, 2011). É esperado que o aparecimento de mutações patogênicas seja maior em regiões de *hotspots*, uma vez que a maior quantidade absoluta de mutações e o possível comprometimento dos mecanismos reparatórios em tais regiões gerariam mais oportunidades para o aparecimento e propagação das mesmas. Uma lesão pré-mutagênica importante é 8-oxoguanina (8-oxoG), formada a partir da oxidação espontânea da base guanina, sendo uma das principais causas da transversão G>T. Camundongos deficientes para as enzimas MTH1 e MUTYH, reparadoras de DNA oxidado, mostram aumento de transversões e exibem maior tumorigênese espontânea (Nakabeppu *et al*, 2006). O trabalho de Ohno *et al* (2006) mostrou que áreas com alta densidade de SNPs se co-localizam com as regiões de alta densidade de 8-oxoG, sugerindo tal mecanismo com uma das principais causas da diversidade genômica humana.

A análise da conservação através da distância de conservação de Levenshtein (Lev) também sugere que tal assinatura seja um marcador relacionado com a maior

geração de mutações deletérias em regiões mais variáveis. Ao contrário do esperado, as mutações apresentaram uma distância de edição ligeiramente maior do que os polimorfismos, quando comparadas com os trechos de 23 nucleotídeos das respectivas sequências ortólogas em rato ou camundongo, indicando uma maior variabilidade da região de contexto destas mutações, mesmo em relação à variabilidade já aumentada da região 5' UTR (Resch *et al*, 2009; Fu & Lin, 2012). Tal condição de maior variabilidade é suportada também pelo enriquecimento de variantes raras com provável perda de função próximo às extremidades 5' dos genes (MacArthur *et al,* 2012).

Os parâmetros de ΔΔGfmax, tipoV e Lev apresentaram pequeno poder explicativo nos modelos de regressão logística, contribuindo de maneira similar quando adicionados a outros parâmetros. O melhor modelo formado apenas com estas variáveis (modelo 1+3) teve desempenho de 63% de explicação correta para a funcionalidade da variação, apenas 13% superior a um modelo puramente aleatório. Apesar da aparente pouca contribuição, este modelo se torna mais significativo quando se tem em mente o grande enriquecimento da região 5' UTR para elemento regulatórios com diversos mecanismos funcionais, conhecidos e desconhecidos (Chatterjee e Pal, 2009; Araujo *et al*, 2012). Variáveis altamente explicativas por si só seriam muito incomuns em tal sistema. Espera-se que melhores modelos sejam obtidos com a incorporação de tais elementos, como uAUGs e uORFs, e sítios de ligação de RBPs, e com a melhor elucidação dos processos de inicialização da tradução alternativos, como IRES e formas não canônicas de inicialização 5' dependentes, *cap* independentes.

Os parâmetros relacionados às sequências Jaspar (nJaspar_alt e nJaspar_crd) se mostraram pouco informativos. Uma provável explicação foi a escassez de sítios de ligação de fatores de transcrição encontrados na 5' UTR, o que não é totalmente inesperado. Entretanto a possibilidade da existência de uma correlação não pode ser completamente descartada sem a realização de novos estudos com um maio número de amostras de mutações causadoras de doenças.

O parâmetro relacionado a distância do início de transcrição (DiTr) também se mostrou pouco informativo também, sendo os valores obtidos o esperado para posições

dentro do comprimento da maioria da regiões 5' UTR humanas, em torno de 100 a 200 nucleotídeos.

Os parâmetros relacionados ao início da sequência codificante, ou seja, o início da tradução se apresentaram altamente significativos, sendo que DiSClog sozinho apresentou um desempenho moderado (área de curva de 0,752), que foi ligeiramente incrementado com a adição das variáveis ΔΔGfmax e tipoV ao modelo (área de curva de 0,772), de maneira que, considerando o contexto do estudo, este pode ser tido como um bom resultado. Ao se observar os valores de DiSC, entretanto, torna-se difícil o estabelecimento de alguma relação causal apenas com o processo de tradução propriamente dito. A verificação da distribuição dos valores de DiSC nas mutações e polimorfismos permite a identificação de três regiões distintas, em relação ao AUG da sequência codificante, que possivelmente contém a chave para um melhor entendimento dos resultados obtidos, seja por uma relação de causa e efeito, seja por uma variabilidade genética diferenciada.

A região de 0 a 50 nucleotídeos *upstream* concentrou 90,683% dos valores obtidos para os polimorfismos e 41,558% para as mutações. Ambos os grupos tiveram seu maior percentual de variações dentro desta faixa. A grande concentração de polimorfismo neste ponto parece ser um indicativo de menor pressão de seleção neste trecho, sendo possivelmente uma "região segura para a" ocorrência de polimorfismos. Sabe-se que tanto em procariotos quanto em eucariotos, a região imediatamente antes do AUG da sequência codificante apresenta-se pouco estruturada (Shabalina *et al*, 2012), o que é esperado levando em consideração a aparente variabilidade da região. Verifica-se uma queda abrupta do percentual de polimorfismos entre 50 e 100 nucleotídeos *upstream* (7,28%), sendo a proporção de polimorfismos a mais de 100 nucleotídeos *upstream* menor que 3%. As mutações por outro lado, apresentam-se de maneira mais distribuídas ao longo dos primeiros 200 nucleotídeos *upstream*, correspondendo ao comprimento esperado para a maioria das regiões 5' UTR humanas. No grupo das mutações é possível notar se um enriquecimento no intervalo de 100 a 150 nucleotídeos *upstream* ao AUG. Uma possível explicação para o enriquecimento de mutações neste trecho seria a presença de sequências complementares a trechos do 18S rRNA, que atuariam estabilizando o ribossomo próximo ao códon de inicialização. Trabalhos de

hibridização do 18S rRNA de camundongo mostraram a existência de regiões complementares a mRNAs separadas entre si por aproximadamente 100 a 300 nucleotídeos de distância (Shabalina *et al*, 2012). Finalmente, no grupo das mutações é possível observar-se a presença de um percentual considerável (em torno de 8%) de mutações na faixa de 700 a 1300 nucleotídeos *upstream* ao AUG, o que sugere a atuação de outros elementos regulatórios, como uORFs, estruturas secundárias ou junções de *splicing*. A maior abrangência de mutações em relação a DiSC indica a interação com mais de um tipo de elemento regulatório da região 5' UTR, sendo um melhor entendimento de tais elementos necessário para o estabelecimento de relações causais.

Entretanto existe também a possibilidade da grande diferença entre os grupos para os parâmetros DiSC e DiSClog se tratar de um artefato decorrente das análises, o que explicaria a significância maior em relação aos demais parâmetros encontrados. Tal artefato seria causado pelo fato dos polimorfismos terem sido obtidos do exoma ESP, cujo principal foco seria a sequência codificante, não abrangendo as partes mais distantes da região 5' UTR.

O melhor modelo de regressão logística, o modelo 1+2 (DiSClog+ΔΔGfmax+tipoV), obteve um desempenho, levando em consideração a área debaixo da curva, de 77,2% de explicações corretas, sendo quase que considerado como um bom modelo para este tipo de análise. Considerando-se a pouca quantidade de variáveis observadas, entretanto, tal modelo permite uma correlação considerável e até então não observada entre os parâmetros e o provável efeito funcional de variações na região 5' UTR. O maior obstáculo para a melhora do modelo é a obtenção de mais mutações patogênicas regiões 5' UTR assim como mais destas regiões curadas de humanos, camundongos e ratos, permitindo a análise de um maior número de variáveis e a obtenção de resultados mais confiáveis. Um melhor entendimento dos elementos regulatórios atuantes em tal região também contribuirá bastante ao desenvolvimento de modelos mais explanatórios. Finalmente, a análise da interação conjunta dos parâmetros encontrados entre si e com parâmetros adicionais, também levará a um melhor entendimento do comportamento de tais variações, uma vez que tais parâmetros parecem agir em conjunto para a determinação de efeitos deletérios das variações.

7 CONCLUSÃO

Os parâmetros bioinformáticos do valor máximo de diferença de energia livre de Gibbs de *folding (*ΔΔGfmax), tipo de variação (tipoV), Distância do início da Sequência Codificante (DiSClog), logaritmos natural da Distância do início da Sequência Codificante (DiSClog) e distância de edição de Levenshtein (Lev) permitiram a distinção entre o grupo dos polimorfismos, de caráter funcional diverso e desconhecido, e o grupo das mutações, conhecidamente causadoras de doença. Tais parâmetros permitiram a construção de modelos matemáticos de função logística razoavelmente explicativos, ainda que insuficientes para a explicação completa do comportamento das variações pontuais. O parâmetro ΔΔGfmax indicou uma relação entre as mutações e entre estruturas secundárias mais estáveis geradas pelas mesmas, provavelmente de caráter funcional significativo para parte das mesmas. Os parâmetros Lev e tipoV parecem indicar a possível origem das mutações como resultantes de regiões de maior variabilidade genômica. O parâmetro DiSC permitiu a identificação de regiões relativas ao AUG da sequência codificante dos transcritos que provavelmente possuem importância funcional. Apesar de não ter sido possível estabelecer relação causal entre os parâmetros e o impacto funcional das variações, ainda sim correlações aparentemente importantes entre os mesmos foram claramente estabelecidas, de forma que os presentes modelos se mostram promissores, podendo ser aperfeiçoados conforme o conheci mento científico dos elementos regulatórios da região 5' UTR for aprimorado.

REFERÊNCIAS*

Abaza I, Gebauer F. Trading translation with RNA-binding proteins. RNA. 2008;14(3):404-9. Epub 2008/01/24.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. Epub 2012/11/07.

Ait Ghezala H, Jolles B, Salhi S, Castrillo K, Carpentier W, Cagnard N, et al. Translation termination efficiency modulates ATF4 response by regulating ATF4 mRNA translation at 5' short ORFs. Nucleic acids research. 2012;40(19):9557-70. Epub 2012/08/21.

Andreev DE, Dmitriev SE, Terenin IM, Prassolov VS, Merrick WC, Shatsky IN. Differential contribution of the m7G-cap to the 5' end-dependent translation initiation of mammalian mRNAs. Nucleic acids research. 2009;37(18):6135-47. Epub 2009/08/22.

Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, et al. Before It Gets Started: Regulating Translation at the 5' UTR. Comparative and functional genomics. 2012;2012:475731. Epub 2012/06/14.

Arora A, Dutkiewicz M, Scaria V, Hariharan M, Maiti S, Kurreck J. Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. RNA. 2008;14(7):1290-6. Epub 2008/06/03.

* De acordo com a norma UNICAMP/FOP, baseadas na norma do International College Committee of Medical Journal Editor – Grupo de Vancouver. Abreviatura de periódicos em conformidade com o Medline.

Avery, OT, Macleod CM, Mccarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Journal of experimental medicine. 1944;79:137-58.

Babendure JR, Babendure JL, Ding JH, Tsien RY. Control of mammalian translation by mRNA structure near caps. RNA. 2006;12(5):851-61. Epub 2006/03/17.

Bompfunewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, et al. Variations on RNA folding and alignment: lessons from Benasque. Journal of mathematical biology. 2008;56(1-2):129-44. Epub 2007/07/06.

Bugaut A, Balasubramanian S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic acids research. 2012;40(11):4727-41. Epub 2012/02/22.

Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(18):7507-12. Epub 2009/04/18.

Cazzola M, Skoda RC. Translational pathophysiology: a novel molecular mechanism of human disease. Blood. 2000;95(11):3280-8. Epub 2000/05/29.

Chatterjee S, Pal JK. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. Biology of the cell / under the auspices of the European Cell Biology Organization. 2009;101(5):251-62. Epub 2009/03/12.

Cooper DN, Bacolla A, Ferec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM. On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Human mutation. 2011;32(10):1075-99. Epub 2011/08/20.

Crowe ML, Wang XQ, Rothnagel JA. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC genomics. 2006;7:16. Epub 2006/01/28.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic acids research. 2012;40(Database issue):D84-90. Epub 2011/11/17.

Fu GC, Lin WC. Identification of gene-oriented exon orthology between human and mouse. BMC genomics. 2012;13 Suppl 1:S10. Epub 2012/03/06.

Gaba A, Wang Z, Krishnamoorthy T, Hinnebusch AG, Sachs MS. Physical evidence for distinct mechanisms of translational control by upstream open reading frames. The EMBO journal. 2001;20(22):6453-63. Epub 2001/11/15.

Goss DJ, Theil EC. Iron responsive mRNAs: a family of Fe2+ sensitive riboregulators. Accounts of chemical research. 2011;44(12):1320-8. Epub 2011/10/27.

Gray NK, Hentze MW. Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs. The EMBO journal. 1994;13(16):3882-91. Epub 1994/08/15.

Griffth F. The significance of pneumococcal types. The Journal of Hygiene. 1928;27:113-59.

Hentze MW, Muckenthaler MU, Galy B, Camaschella C. Two to tango: regulation of Mammalian iron metabolism. Cell. 2010;142(1):24-38. Epub 2010/07/07.

Iacono M, Mignone F, Pesole G. uAUG and uORFs in human and rodent 5'untranslated mRNAs. Gene. 2005;349:97-105. Epub 2005/03/22.

Jackson RJ, Hellen CU, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nature reviews Molecular cell biology. 2010;11(2):113-27. Epub 2010/01/23.

Jenkins RH, Bennagi R, Martin J, Phillips AO, Redman JE, Fraser DJ. A conserved stem loop motif in the 5'untranslated region regulates transforming growth factor-beta(1) translation. PloS one. 2010;5(8):e12283. Epub 2010/09/25.

Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic acids research. 1987;15(20):8125-48. Epub 1987/10/26.

Kozak M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. The Journal of biological chemistry. 1991;266(30):19867-70. Epub 1991/10/25.

Kumar A, Wong AK, Tizard ML, Moore RJ, Lefevre C. miRNA_Targets: a database for miRNA target predictions in coding and non-coding regions of mRNAs. Genomics. 2012;100(6):352-6. Epub 2012/09/04.

Kumari S, Bugaut A, Huppert JL, Balasubramanian S. An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. Nature chemical biology. 2007;3(4):218-21. Epub 2007/02/27.

Lammich S, Kamp F, Wagner J, Nuscher B, Zilow S, Ludwig AK, et al. Translational repression of the disintegrin and metalloprotease ADAM10 by a stable G-quadruplex secondary structure in its 5'-untranslated region. The Journal of biological chemistry. 2011;286(52):45063-72. Epub 2011/11/09.

Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Doklady Akademii Nauk SSSR.1966;163(4):845-48.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823-8. Epub 2012/02/22.

Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly journal of the royal meteorological society. 2002;128:2145-66

Meijer HA, Thomas AA. Ribosomes stalling on uORF1 in the Xenopus Cx41 5' UTR inhibit downstream translation initiation. Nucleic acids research. 2003;31(12):3174-84. Epub 2003/06/12.

Nakabeppu Y, Sakumi K, Sakamoto K, Tsuchimoto D, Tsuzuki T, Nakatsu Y. Mutagenesis and carcinogenesis caused by the oxidation of nucleic acids. Biological chemistry. 2006;387(4):373-9. Epub 2006/04/12.

Ohno M, Miura T, Furuichi M, Tominaga Y, Tsuchimoto D, Sakumi K, et al. A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. Genome research. 2006;16(5):567-75. Epub 2006/05/03.

Palam LR, Baird TD, Wek RC. Phosphorylation of eIF2 facilitates ribosomal bypass of an inhibitory upstream ORF to enhance CHOP translation. The Journal of biological chemistry. 2011;286(13):10939-49. Epub 2011/02/03.

Parsons CJ, Stefanovic B, Seki E, Aoyama T, Latour AM, Marzluff WF, et al. Mutation of the 5'-untranslated region stem-loop structure inhibits alpha1(I) collagen expression in vivo. The Journal of biological chemistry. 2011;286(10):8609-19. Epub 2011/01/05.

Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001;276(1-2):73-81. Epub 2001/10/10.

Pickering BM, Willis AE. The implications of structured 5' untranslated regions on translation and disease. Seminars in cell & developmental biology. 2005;16(1):39-47. Epub 2005/01/22.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research. 2009;19(7):1316-23. Epub 2009/06/06.

Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. Evolution of alternative and constitutive regions of mammalian 5'UTRs. BMC genomics. 2009;10:162. Epub 2009/04/18.

Reynolds PR. In sickness and in health: the importance of translational regulation. Archives of disease in childhood. 2002;86(5):322-4. Epub 2002/04/24.

Scheper GC, van der Knaap MS, Proud CG. Translation matters: protein synthesis defects in inherited disease. Nature reviews Genetics. 2007;8(9):711-23. Epub 2007/08/08.

Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research. 2013;41(4):2073-94. Epub 2013/01/08.

Shahid R, Bugaut A, Balasubramanian S. The BCL-2 5' untranslated region contains an RNA G-quadruplex-forming motif that modulates protein expression. Biochemistry. 2010;49(38):8300-6. Epub 2010/08/24.

Shih CT, Wells SA, Hsu CL, Cheng YY, Romer RA. The interplay of mutations and electronic properties in disease-related genes. Scientific reports. 2012;2:272. Epub 2012/02/23.

Sobczak K, Krzyzosiak WJ. Structural determinants of BRCA1 translational regulation. The Journal of biological chemistry. 2002;277(19):17349-58. Epub 2002/03/06.

Spruill LS, McDermott PJ. Role of the 5'-untranslated region in regulating translational efficiency of specific mRNAs in adult cardiocytes. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2009;23(9):2879-87. Epub 2009/05/07.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome medicine. 2009;1(1):13. Epub 2009/04/08.

Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16-23. Epub 2000/05/17.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9. Epub 2012/05/19.

Terenin IM, Andreev DE, Dmitriev SE, Shatsky IN. A novel mechanism of eukaryotic translation initiation that is neither m7G-cap-, nor IRES-dependent. Nucleic acids research. 2013;41(3):1807-16. Epub 2012/12/27.

Vassilenko KS, Alekhina OM, Dmitriev SE, Shatsky IN, Spirin AS. Unidirectional constant rate motion of the ribosomal scanning particle during eukaryotic translation initiation. Nucleic acids research. 2011;39(13):5555-67. Epub 2011/03/19.

Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Molecular systems biology. 2010;6:400. Epub 2010/08/27.

Walser JC, Furano AV. The mutational spectrum of non-CpG DNA varies with CpG content. Genome research. 2010;20(7):875-82. Epub 2010/05/26.

Wang XQ, Rothnagel JA. 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. Nucleic acids research. 2004;32(4):1382-91. Epub 2004/03/03.

Watson JD, Crick FHC, Molecular structure of nucleic acid. A structure for deoxyribose nucleic acid. Nature. 1953;171:737-38

Wethmar K, Smink JJ, Leutz A. Upstream open reading frames: molecular switches in (patho)physiology. BioEssays : news and reviews in molecular, cellular and developmental biology. 2010;32(10):885-93. Epub 2010/08/21.

Wu C, Wei J, Lin PJ, Tu L, Deutsch C, Johnson AE, et al. Arginine changes the conformation of the arginine attenuator peptide relative to the ribosome tunnel. Journal of molecular biology. 2012;416(4):518-33. Epub 2012/01/17.

Xia X, Holcik M. Strong eukaryotic IRESs have weak secondary structure. PloS one. 2009;4(1):e4136. Epub 2009/01/07.

Glossário

Motif: intervalo de uma sequência de nucleotídeos pré-definida, podendo ser repetido ao longo do genoma, ser conservado entre espécies ou ter função conhecida, geralmente como sítio de ligação de proteínas ou de outras sequências.

Hotspots: Regiões genômicas de maior ocorrência de mutações.

Hairpin: Estrutura secundária formada pela auto complementariedade de uma sequência de ácidos nucleicos, em formato de "grampo de cabelo".

Stem loop: Estrutura secundária formada pela auto complementariedade de uma sequência de ácidos nucleicos, caracterizada pela presença de uma ou mais hastes (*stem*) complementares, com uma ou mais alças (*loop*), não complementares, no final ou entre as hastes.

ANEXO

Tabela 3 – Dados das mutações obtidas do banco de dados do HGMD.

Gene	Posição Genômica	Substituição	Fenótipo	Referência
AAAS	53715654	CT	Triple-A syndrome	Yassaee (2011) AMR 42, 163
ACTN4	39138352	CT	Glomerulosclerosis	Dai (2010) NDT 25, 824
ADSL	40742514	TC	Adenylosuccinate lyase deficiency	Race (2000) HMG 9, 2159
ALAS2	55057393	CT	Severe iron overload	Lee (2009) BCMD 42, 1
ALPL	21835920	CT	Hypophosphatasia	Taillandier (2000) HUM MUT 15, 293
ANK1	41655209	GA	Spherocytosis	Gallagher (1998) BCMD 24, 539
ANKH	14871567	CT	Chondrocalcinosis 2	Pendleton (2002) AJHG 71, 933
ANKRD26	27389368	AC	Thrombocytopaenia 2	Noris (2011) BLOOD 117, 6673
ANKRD26	27389373	CA	Thrombocytopaenia 2	Noris (2011) BLOOD 117, 6673
ANKRD26	27389374	CA	Thrombocytopaenia 2	Noris (2011) BLOOD 117, 6673
ANKRD26	27389376	AC	Thrombocytopaenia 2	Noris (2011) BLOOD 117, 6673
ANKRD26	27389381	TG	Thrombocytopaenia 2	Noris (2011) BLOOD 117, 6673
ANKRD26	27389382	AT	Thrombocytopaenia 2	Pippucci (2011) AJHG 88, 115
ANKRD26	27389382	AG	Thrombocytopaenia 2	Noris (2011) BLOOD 117, 6673
ANKRD26	27389383	GA	Thrombocytopaenia 2	Pippucci (2011) AJHG 88, 115
ANKRD26	27389389	GA	Thrombocytopaenia 2	Pippucci (2011) AJHG 88, 115
ASS1	133327612	CT	Citrullinaemia	Engel (2009) HUM MUT 30, 300
ATM	108094508	GA	Ataxia telangiectasia	Castellvi-Bel (1999) HUM MUT 14, 156
ATM	108093770	AG	Ataxia telangiectasia	Castellvi-Bel (1999) HUM MUT 14, 156
ATP7B	52585596	CA	Wilson disease	Nanji (1997) AJHG 60, 1423
ATP7B	52585606	AC	Wilson disease	Wan (2010) HEPATOLOGY 52, 1662
ATP7B	52585683	AT	Wilson disease	Wan (2010) HEPATOLOGY 52, 1662
ATP7B	52585551	AC	Wilson disease	Nanji (1997) AJHG 60, 1423
BMP7	55841179	GT	Eye / skeletal anomalies	Wyatt (2010) HUM MUT 31, 781
BMPR2	203241529	GA	Pulmonary arterial hypertension	Wang (2009) EJHG 17, 1063
BRCA2	32890572	GA	Breast and/or ovarian cancer	Capalbo (2006) ANN ONCO 17S7, vii34
CDKN1B	12870767	GC	Multiple endocrine neoplasia 1	Agarwal (2009) JCEM 94, 1826
CDKN1C	2906803	GA	Beckwith-Wiedemann syndrome	Lam (1999) JMG 36, 518
CDKN1C	2906804	GA	Beckwith-Wiedemann syndrome	Lam (1999) JMG 36, 518
CDKN2A	21974847	CT	Melanoma	Bisio (2010) HMG 19, 1479
CDKN2A	21974851	CT	Melanoma	Soufir (2004) BJC 90, 503
CDKN2A	21974860	GT	Melanoma	Liu (1999) NAT GENET 21, 128

Continua

Gene	Posição Genômica	Substituição	Fenótipo	Referência
CDKN2A	21974875	CA	Melanoma	Council (2009) EXP DERM 18, 485
CDKN2A	21974882	GT	Melanoma	Bisio (2010) HMG 19, 1479
CFC1	131356867	AG	Congenital heart defects	Roessler (2008) AJHG 83, 18
CFH	196620941	CT	Haemolytic uraemic syndrome	Westra (2010) NDT 25, 2195
CFTR	117120115	CT	Disseminated bronchiectasis	Lukowski (2011) HUM MUT 32, E2266
CHRNE	4806452	GA	Congenital myasthenic syndrome	Abicht (2000) ACTA MYO 19, 23
CHRNE	4806453	GA	Congenital myasthenic syndrome	Ohno (1999) NEU DIS 9, 131
CHRNE	4806454	CT	Congenital myasthenic syndrome	Nichols (1999) ANN NEUROL 45, 439
DIAPH3	60738072	GA	Auditory neuropathy	Schoen (2010) PNAS 107, 13396
DKC1	153991099	CG	Dyskeratosis congenita X-linked	Knight (2001) HUM GENET 108, 299
DMD	33229483	TA	Muscular dystrophy Duchenne	Flanigan (2009) HUM MUT 30, 1657
EDN3	57875743	GA	Hirschsprung disease	Sangkhathat (2006) JHG 51, 1126
EDN3	57875849	CA	Hirschsprung disease	Garcia-Barcelo (2003) CLIN CHEM 50, 93
EDNRB	78492734	GA	Hirschsprung disease	Amiel (1996) HMG 5, 355
ENG	130616761	CT	Haemorrhagic telangiectasia	Kim (2011) BMC MG 12,
F11	187186995	GA	Factor XI deficiency	Vasileiadis (2009) BCF 20, 309
F11	187187397	GA	Factor XI deficiency	Mitchell (2006) HUM MUT 27, 829
F7	113760155	CT	Factor VII deficiency	Kwon (2011) BCF 22, 102
F7	113760126	AC	Factor VII deficiency	Millar (2000) HUM GENET 107, 327
F7	113760124	AC	Factor VII deficiency	Kavlie (2003) TH 90, 194
F7	113760117	AG	Factor VII deficiency	Herrmann (2009) HAEMO 15, 267
F7	113760112	TC	Factor VII deficiency	Herrmann (2009) HAEMO 15, 267
F7	113760101	CT	Factor VII deficiency	Carew (2000) BLOOD 96, 4370
F7	113760101	CG	Factor VII deficiency	Herrmann (2009) HAEMO 15, 267
F7	113760097	TG	Factor VII deficiency	Kavlie (2003) TH 90, 194
F7	113760096	TC	Factor VII deficiency	McVey (2001) HUM MUT 17, 3
F7	113760096	TG	Factor VII deficiency	Herrmann (2009) HAEMO 15, 267
F7	113760095	TG	Factor VII deficiency	Arbini (1997) BLOOD 89, 176
F7	113760094	CT	Factor VII deficiency	Herrmann (2009) HAEMO 15, 267
F7	113760091	GC	Factor VII deficiency	Herrmann (2009) HAEMO 15, 267
F7	113760062	CG	Factor VII deficiency	Carew (1998) BLOOD 92, 1639
F7	113760060	CT	Factor VII deficiency	Nagaizumi (2002) BJH 119, 1052
F8	154250939	GA	Haemophilia A	Bogdanova (2007) HUM MUT 28, 54
F8	154251046	CT	Haemophilia A	Bogdanova (2007) HUM MUT 28, 54

Continuação

Gene	Posição Genômica	Substituição	Fenótipo	Referência
F8	154251084	TG	Haemophilia A	Riccardi (2009) JTH 7, 1234
F8	154251687	AG	Haemophilia A	Sanna (2008) HAEMO 14, 796
F9	138612102	GA	Haemophilia B	Tanimoto (1990) RIN BYORI 38, 1041
FASLG	172628081	TC	Sjoegren syndrome	Tsuzaka (2007) AUTOIMM 40, 497
FECH	55254103	GC	Protoporphyria erythropoietic	Di Pierro (2004) HUM GENET 114, 609
FMR1	146993615	GC	Fragile X retardation syndrome	Tarleton (2002) JMG 39, 196
FOXL2	138665815	GA	Blepharophimosis/ptosis	Li (2009) GTMB 13, 257
FOXP3	49114969	GT	IPEX syndrome	Myers (2006) ADC 91, 63
FSHR	49381679	AG	Decreased promoter activity	Wunsch (2005) FERT STER 84, 446
FSHR	49381694	AT	Increased promoter activity	Wunsch (2005) FERT STER 84, 446
FSHR	49381593	AG	Increased promoter activity	Wunsch (2005) FERT STER 84, 446
GALK1	73761239	TC	Increased GALK1 activity	Park (2009) BMC MG 10,
GCH1	55369403	CT	Dystonia dopa-responsive	Tassin (2000) BRAIN 123, 1112
GCH1	55369420	CT	Dystonia dopa-responsive	Bandmann (1998) ANN NEUROL 44, 649
GFI1	92949058	AC	Neutropaenia severe chronic	Hochberg (2008) PBC 50, 630
GHRHR	31003560	AC	GH deficiency (type 1B)	Salvatori (2002) MOL END 16, 450
GJB1	70443099	CT	Charcot-Marie-Tooth disease	Ionasescu (1996) NEUROLOGY 47, 541
GJB1	70443031	GC	Charcot-Marie-Tooth disease	Houlden (2004) ANN NEUROL 56, 730
GJB1	70443029	TG	Charcot-Marie-Tooth disease	Ionasescu (1996) NEUROLOGY 47, 541
GJB2	20767158	CT	Deafness autosomal recessive 1	Matos (2007) HUM GENET 121, 298
GJC2	228337561	AG	Pelizaeus-Merzbacher disease	Osaka (2010) ANN NEUROL 68, 250
GUCY2D	7906220	TC	Leber congenital amaurosis	Stone (2007) AJO 144, 791
HAMP	35773328	CT	Haemochromatosis HFE related	Island (2009) HAEMATOL 94, 720
HAMP	35773456	GA	Haemochromatosis juvenile	Matthes (2004) BLOOD 104, 2181
HAMP	35773453	GT	Haemochromatosis juvenile	Delbini (2008) HUM GENET 124, 313
HBA1	226707	GC	Thalassaemia alpha	Hadavi (2009) HEMOGLOBIN 33, 235
HBA2	222910	CT	Haemoglobin variant	Sarkar (2005) BJH 129, 282
HBA2	222891	CG	Haemoglobin variant	Lacerra (2004) HUM MUT 24, 338
HBB	5248343	CG	Thalassaemia beta	Ibn Ayub (2010) GTMB 14, 299
HBD	5255768	TC	Thalassaemia delta	Papadakis (1997) HUM MUT 9, 465
HBD	5255778	AG	Thalassaemia delta	Papadakis (1997) HUM MUT 9, 465
HBD	5255781	CT	Thalassaemia delta	Bouva (2006) HAEMATOL 91, 129
HBD	5255789	AT	Haemoglobin variant	De Angioletti (2002) HUM MUT 20, 358
HBD	5255790	TC	Thalassaemia delta	Matsuda (1992) BLOOD 80, 1347

Continuação

Gene	Posição Genômica	Substituição	Fenótipo	Referência
HBD	5255793	GA	Thalassaemia beta	Morgado (2007) EJH 79, 422
HBD	5255743	TC	Thalassaemia delta	Papadakis (2003) LSDB139 PC, 2238
HBD	5255744	AG	Thalassaemia delta	Frischknecht (2005) HEMOGLOBIN 29, 151
HBD	5255749	CA	Thalassaemia delta	Papadakis (1997) HUM MUT 9, 465
HBM	209709	TC	Thalassaemia alpha	De Gobbi (2006) SCIENCE 312, 1215
HFE	26087649	GA	Del-Castillo-Rueda (2010) EJH	
HNF1A	121416453	GA	Diabetes MODY3	Godart (2000) HUM MUT 15, 173
HNF1A	121416448	GC	Diabetes mellitus type 1	Yoshiuchi (1999) DIABETOL 42, 621
HNF1A	121416385	CT	Diabetes MODY3	Godart (2000) HUM MUT 15, 173
HNF1A	121416354	TC	Diabetes MODY3	Godart (2000) HUM MUT 15, 173
HNF1A	121416289	AC	Diabetes MODY3	Gragnoli (1997) DIABETES 46, 1648
HNF1A	121416110	GA	Diabetes mellitus type 2	Cox (1999) DIABETOL 42, 120
HNF1A	121416034	GC	Diabetes MODY	Radha (2009) JCEM 94, 1959
HNF1A	121416510	CG	Diabetes MODY3	Godart (2000) HUM MUT 15, 173
HNF1A	121416475	TG	Diabetes MODY3	Godart (2000) HUM MUT 15, 173
HNF4A	42984309	AG	Diabetes	Wirsing (2010) DIAB MED 27, 631
HNF4A	42984276	CT	Diabetes	Wirsing (2010) DIAB MED 27, 631
HNF4A	42984264	GA	Diabetes MODY1	Hansen (2002) JCI 110, 827
HNF4A	42984253	CG	Diabetes MODY1	Ek (2006) DIABETES 55, 1869
HSPB1	75931813	TC	Amyotrophic lateral sclerosis	Dierick (2007) HUM MUT 28, 830
IDUA	980871	CG	Mucopolysaccharidosis I	Wang (2011) CLIN GENET epub, epub
INS	2182419	AC	Diabetes neonatal	Garin (2010) PNAS 107, 3105
INS	2182532	CG	Diabetes neonatal	Garin (2010) PNAS 107, 3105
INS	2182532	CA	Diabetes neonatal	Garin (2010) PNAS 107, 3105
INS	2182533	CG	Diabetes neonatal	Garin (2010) PNAS 107, 3105
IRF6	209979367	GA	Van der Woude syndrome	de Lima (2009) GENET MED 11, 241
IRF6	209979367	GA	Van der Woude syndrome	de Lima (2009) GENET MED 11, 241
IRF6	209979435	CT	Van der Woude syndrome	de Lima (2009) GENET MED 11, 241
IRF6	209975332	CA	Van der Woude syndrome	de Lima (2009) GENET MED 11, 241
IRF6	209975361	AT	Van der Woude syndrome	Kondo (2002) NAT GENET 32, 285
ITGA2B	42470923	GA	Glanzmann thrombasthenia	Kannan (2009) JTH 7, 1878
KCNJ11	17409772	GT	Hyperinsulinism	Tornovsky (2004) JCEM 89, 6224
KCNJ11	17409692	CT	Hyperinsulinism	Huopio (2002) JCEM 87, 4502
KLF1	12998078	TC	Blood group variant In(Lu)	Singleton (2008) BLOOD 112, 2081
LDLR	11200212	AG	Hypercholesterolaemia	Medeiros (2010) ATHEROS 212, 553
LDLR	11200091	CT	Hypercholesterolaemia	Miyake (2009) ATHEROS 203, 153
LDLR	11200089	CG	Hypercholesterolaemia	De Castro-OrÃƒÂ³s (2011) HUM MUT 32, 868
LDLR	11200086	CA	Hypercholesterolaemia	Fouchier (2005) HUM MUT 26, 550

Continuação

Gene	Posição Genômica	Substituição	Fenótipo	Referência
LDLR	11200086	CG	Hypercholesterolaemia	Smith (2007) EJHG 15, 1186
LDLR	11200211	CA	Hypercholesterolaemia	Day (1997) HUM MUT 10, 116
LDLR	11200085	CG	Hypercholesterolaemia	Alonso (2009) CLIN BIO 42, 899
LDLR	11200085	CT	Hypercholesterolaemia	Marduel (2010) HUM MUT 31, E1811
LDLR	11200069	CT	Hypercholesterolaemia	Dedoussis (2004) HUM MUT 23, 285
LDLR	11200037	CT	Hypercholesterolaemia	Fouchier (2005) HUM MUT 26, 550
LDLR	11200019	CT	Hypercholesterolaemia	Lind (2002) ATHEROS 163, 399
LDLR	11200202	AC	Hypercholesterolaemia	Mozas (2004) HUM MUT 24, 187
LDLR	11199958	AG	Hypercholesterolaemia	Marduel (2010) HUM MUT 31, E1811
LDLR	11200220	CT	Hypercholesterolaemia	Fouchier (2005) HUM MUT 26, 550
LIF	30642690	CA	Female infertility	Giess (1999) MHR 5, 581
MEFV	3306599	CG	Mediterranean fever familial	Haverkamp (2005) LSDB167 PC, .
MEFV	3306917	GA	Mediterranean fever familial	Notarnicola (2007) LSDB167 AB, .
MEFV	3306969	CG	Mediterranean fever familial	Oosta (2007) LSDB167 AB, .
MEFV	3307201	CG	Mediterranean fever familial	Notarnicola (2007) LSDB167 AB, .
MEN1	64577603	CA	Hyperparathyroidism	Jager (2006) MCE 249, 123
MLH1	37034932	CG	Colorectal cancer non-polyposis	Zhong (2007) BG 45, 671
MLH1	37035012	CA	Colorectal cancer non-polyposis	Hitchins (2011) CANC CELL 20, 200
MSH2	47630150	GA	Colorectal / endometrial cancer	Shin (2002) CANCER RES 62, 38
MSH2	47630106	GC	Colorectal cancer non-polyposis	Shin (2002) CANCER RES 62, 38
NAGS	42078968	CA	Hyperammonaemia	Heibel (2011) HUM MUT 32, 1153
NKX2-5	172662542	CT	Reduced activity	Hermanns (2011) JCEM 96, E977
OTC	38211584	AG	Ornithine transcarbamylase deficiency	Luksan (2010) HUM MUT 31, E1294
PARK2	163148721	GT	Parkinson disease recessive	Hedrich (2002) NEUROLOGY 58, 1239
PARK7	8021919	CG	Parkinson disease recessive	Tarantino (2009) PRD 15, 324
PAX9	37130036	GA	Hypodontia	Mendoza-Fandino (2011) CLIN GENET 80, 265
PCSK9	55505180	CA	Hypercholesterolaemia dominant	Blesa (2008) JCEM 93, 3577
PDE6H	15130918	GC	Cone dystrophy	Piri (2005) OPHTHALMOL 112, 159
PEX7	137143759	CT	Rhizomelic chondrodysplasia punctata	Braverman (2002) HUM MUT 20, 284
PIGM	160001799	CG	Glycosylphosphatidylinositol	Almeida (2006) NAT MED 12, 846
PKLR	155271258	AG	Pyruvate kinase deficiency	Manco (1999) HUM GENET 105, 188
PKLR	155271259	GC	Pyruvate kinase deficiency	Marcello (2008) BCMD 41, 261
PKLR	155271269	GC	Pyruvate kinase deficiency	van Wijk (2003) BLOOD 101, 1596
PLEKHG4	67313932	CT	Cerebellar ataxia dominant	Ishikawa (2005) AJHG 77, 280
PLP1	103031893	CT	Pelizaeus-Merzbacher disease	Kawanishi (1996) HUM MUT 7, 355

Continuação

Gene	Posição Genômica	Substituição	Fenótipo	Referência
POMC	25387652	CA	Obesity adrenal insufficiency	Krude (1998) NAT GENET 19, 155
PRM1	11375202	GC	Oligozoospermia ?	Ravel (2007) MHR 13, 461
PROC	128175994	TG	Protein C deficiency type I	Labrouche (2003) BCF 14, 531
PROK2	71834207	CA	Kallmann syndrome	DodÃƒÂ© (2006) PG 2, e175
PROS1	93692761	CT	Protein S deficiency	Sanda (2007) BJH 138, 663
PROS1	93692597	CG	Protein S deficiency	Espinosa-Parrilla (2000) HUM MUT 15, 463
PTEN	89623226	TC	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623142	CT	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623116	AG	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623056	CT	Cowden disease	Tan (2011) AJHG 88, 42
PTEN	89623049	CT	Cowden disease	Tan (2011) AJHG 88, 42
PTEN	89622988	AG	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623462	AG	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623428	GC	Cowden disease	Teresi (2007) AJHG 81, 756
PTEN	89623392	CT	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623373	CG	Cowden disease	Teresi (2007) AJHG 81, 756
PTEN	89623365	GT	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623331	TC	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623306	GT	Cowden disease	Zhou (2003) AJHG 73, 404
PTEN	89623296	GA	Cowden disease	Zhou (2003) AJHG 73, 404
RB1	48877900	GT	Retinoblastoma	MacÃƒÂas (2008) CB 4, 93
RB1	48877899	GC	Retinoblastoma	Cowell (1996) ONCOGENE 12, 431
RB1	48877860	GT	Retinoblastoma	Sakai (1991) NATURE 353, 83
RB1	48877856	TA	Retinoblastoma	Taylor (2007) HUM MUT 28, 284
RB1	48877856	TG	Retinoblastoma	Taylor (2007) HUM MUT 28, 284
RB1	48877851	GA	Retinoblastoma	Sakai (1991) NATURE 353, 83
RB1	48877837	GA	Retinoblastoma	Richter (2003) AJHG 72, 253
RB1	48878045	CG	Retinoblastoma	Barbosa (2009) BMC MG 10, 75
RET	43572680	CG	Hirschsprung disease	Angrist (1995) HMG 4, 821
RET	43572670	GC	Hirschsprung disease	Garcia-Barcelo (2003) CLIN CHEM 50, 93
SEPT9	75316275	GC	Neuritis	Kuhlenbaumer (2005) NAT GENET 37, 1044
SERPING1	57365118	CG	Angioneurotic oedema	Verpy (1996) AJHG 59, 308
SERPING1	57365057	AG	Angioneurotic oedema	Uyguner (2008) HUM GENET 124, 309
SERPING1	57365055	CT	Angioneurotic oedema	Verpy (1996) AJHG 59, 308
SHH	156061506	GA	Holoprosencephaly	Jeong (2008) NAT GENET 40, 1348
SHOX	591568	CA	Leri-Weill dyschondrosteosis	Grigelioniene (2001) HUM GENET 109, 551
SLC26A4	107301244	AG	Enlarged vestibular aqueduct	Choi (2009) HUM MUT 30, 599

Continuação

Gene	Posição Genômica	Substituição	Fenótipo	Referência
SLC26A4	107301201	TC	Pendred syndrome	Yang (2007) AJHG 80, 1055
SLC4A1	42340296	GA	Spherocytosis	Alloisio (1996) BLOOD 88, 1062
SLC5A5	17983075	CT	Iodide transport defect	Nicola (2011) JCEM 96, E1100
SMN1	70220892	AG	Spinal muscular atrophy	Wang (2010) ELECTROPH 31, 2396
SNURF	25200131	GA	Phenotype modifier ?	Maina (2007) JHG 52, 297
SOX9	68676303	TC	Pierre Robin sequence	Benko (2009) NAT GENET 41, 359
SPINK1	147211355	GA	Pancreatitis chronic	Kaneko (2001) JHG 46, 293
SPINK1	147211193	CT	Pancreatitis chronic	Witt (2000) NAT GENET 25, 213
SPR	73114549	GA	Dystonia dopa-responsive	Steinberger (2004) NEUROGENET 5, 187
SRY	2655774	GC	XY sex reversal	Ravel (2009) HUM GENET 126, 333
SRY	2655719	GA	Gonadal dysgenesis	Poulat (1997) HUM MUT Sup1, S192
SYNPO	150027814	GA	Glomerulosclerosis	Dai (2010) NDT 25, 824
SYNPO	150019898	CT	Glomerulosclerosis	Dai (2010) NDT 25, 824
TBX1	19747128	CT	Cardiovascular defects	Gong (2001) JMG 38, e45
TBX22	79277760	CG	Cleft palate and ankyloglossia	Marcano (2004) JMG 41, 68
TGFB3	76447266	GA	Arrhythmogenic ventricular dysplasia	Beffagna (2005) CARDIO RES 65, 366
TH	2193085	TA	Tyrosine hydroxylase deficiency	Verbeek (2007) ANN NEUROL 62, 422
TH	2193086	GA	Tyrosine hydroxylase deficiency	Verbeek (2007) ANN NEUROL 62, 422
TH	2193087	CT	Tyrosine hydroxylase deficiency	Ribases (2007) MGM 92, 274
THPO	184094078	GT	Thrombocythaemia essential	Ghilardi (1999) BJH 107, 310
TMEM127	96931137	CT	Phaeochromocytoma	Yao (2010) JAMA 304, 2611
TRPM1	31369151	CT	Stationary night blindness congenital	Audo (2009) AJHG 85, 720
TTPA	63998581	CT	Ataxia isolated vitamin E deficiency	Usuki (2000) JNNP 69, 254
ZIC2	100634295	CT	Holoprosencephaly	Roessler (2009) HUM MUT 30, E541
serca2b	110719585	CG	Colon cancer	Korosec (2006) CGC 171, 105
serca2b	110718567	GT	Colon cancer	Korosec (2006) CGC 171, 105
serca2b	110718411	GT	Colon cancer	Korosec (2006) CGC 171, 105

Conclusão