UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE BIOQUÍMICA E BIOLOGIA MOLECULAR
BACHARELADO EM BIOTECNOLOGIA

SARA FERREIRA PIRES

ANÁLISE *IN SILICO* DE POLIMORFISMOS DE UM ÚNICO NUCLEOTÍDEO (SNPs) EM DIABETES TIPO 2 EM HUMANOS

FORTALEZA
2018

SARA FERREIRA PIRES

ANÁLISE *IN SILICO* DE POLIMORFISMOS DE UM ÚNICO NUCLEOTÍDEO (SNPs) EM DIABETES TIPO 2 EM HUMANOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Biotecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Biotecnologia. Área de concentração: Genética de Populações.

Orientador: Prof. Dr. Murilo Siqueira Alves.

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

P747a Pires, Sara Ferreira.
Análise in silico de polimorfismos de um único nucleotídeo (SNPs) em diabetes tipo 2 em humanos / Sara Ferreira Pires. – 2018.
45 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Biotecnologia, Fortaleza, 2018.
Orientação: Prof. Dr. Murilo Siqueira Alves.

1. Envelhecimento. 2. Diabetes. 3. Variabilidade genética. 4. Associação genética. 5. SNP. I. Título.
CDD 661

SARA FERREIRA PIRES

ANÁLISE *IN SILICO* DE POLIMORFISMOS DE UM ÚNICO NUCLEOTÍDEO (SNPs) EM DIABETES TIPO 2 EM HUMANOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Biotecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Biotecnologia. Área de concentração: Genética de Populações.

Orientador: Prof. Dr. Murilo Siqueira Alves.

Aprovada em: 23/11/2018.

BANCA EXAMINADORA

_____
Prof. Dr. Murilo Siqueira Alves (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Nicholas Costa Barroso Lima
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Vicente Vieira Faria
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Ao Professor Dr. **Murilo Siqueira Alves**, por ter me permitido sonhar grande e ter aceitado o desafio de explorar novas possibilidades de pesquisa. Por não desacreditar de mim um único segundo e pelo estímulo constante a me fazer buscar novas ideias, novos caminhos. Pela inspiração a ser alguém melhor, pessoal e profissionalmente.

Ao Professor Dr. **José Tadeu Abreu de Oliveira**, pelo acolhimento no Laboratório de Proteínas Vegetais de Defesa e por ser exemplo de dedicação à pesquisa e busca incessante por conhecimento e por novas formas de repassá-lo. E aos colegas de pesquisa, que mesmo por pouco tempo me acompanharam e foram força, cumplicidade e carinho.

Aos membros da banca examinadora, pela disponibilidade de tempo e sugestões.

À Universidade e à Coordenação do Bacharelado em Biotecnologia, pela estrutura e capacitação. Pelos inúmeros caminhos que pude explorar em Ensino, Pesquisa e Extensão.

Ao Professor Dr. **Benildo Sousa Cavada** e à Professora Dra. **Kyria Santiago do Nascimento**, assim como aos colegas do Biomol-Lab, por me abrirem as portas da pesquisa e me mostrarem o mundo de possibilidades que a Ciência pode me proporcionar.

Aos meus familiares, em especial **minha mãe** (Jaqueline Pires), pelo exemplo de garra, força e determinação, e por sempre ter me estimulado a abrir minhas asas e explorar o mundo. Ao **meu pai** (Luis Gonzaga), *in memoriam*, por ter tentado me mostrar o poder transformador dos estudos. E ao **meu tio** (Daniel Barroso), por, além de ter sido uma figura paterna forte e acompanhado meu desenvolvimento, ter sido meu professor e despertado em mim a paixão pela Ciência.

Aos amigos, colegas de graduação e professores, que gostaria de citar um a um. Por todos os momentos de leveza e companheirismo ao longo de toda a jornada acadêmica (e da vida), sendo suporte, força, exemplo e amor.

“If you know you are on the right track, if you have this inner knowledge, then nobody can turn you off... no matter what they say.”

— Barbara McClintock

RESUMO

Com o aumento das ações visando à expansão da expectativa de vida da população, cresce também o risco de manifestação de doenças ligadas ao envelhecimento. Nesse cenário, a diabetes tipo 2 surge como uma das principais doenças crônicas associadas ao envelhecimento, caracterizada por hiperglicemia sanguínea persistente. Esta ocorre devido ao déficit na secreção de insulina pelas células beta-pancreáticas, o qual gera grande impacto negativo para a qualidade de vida do indivíduo afetado, devido à manifestação de diversas patologias associadas à alta glicemia sanguínea. A compreensão dos complexos mecanismos que levam à desregulação da homeostase da glicose sanguínea envolve o entendimento da carga genética associada à manifestação da diabetes. Dentro desse contexto, o presente trabalho teve como objetivo identificar variantes genéticas em regiões codificadoras de quatro genes (FKH1, PDX1, TP53 e TCF7L2) e estimar parâmetros de diversidade genética entre grupos de humanos saudáveis ou diagnosticados com diabetes tipo 2, a partir da análise de sequências consenso obtidas utilizando-se fragmentos de sequenciamento de nova geração (*reads*) depositados no banco de dados do NCBI. Os *reads* foram mapeados em sequências de DNA codificadoras (CDS) correspondentes aos genes de interesse através do *software* Geneious, as sequências consenso geradas foram alinhadas através do *software* MEGA X, e os principais descritores de diversidade intra e interpopulação foram estimados através do *software* DnaSP. Foi observada uma baixa diversidade nucleotídica entre as populações amostradas, embora tenham sido identificados sítios polimórficos e diversos haplótipos distintos. Polimorfismos de um único nucleotídeo (SNPs) foram mapeados e correlacionados com a probabilidade de manifestação da doença dentro de cada população, através da plataforma SNPStats. Os resultados sugerem ser possível rastrear variantes genéticas em regiões do exoma que possam ser associadas ao risco de manifestação de diabetes tipo 2 em populações de humanos, principalmente para os genes FOXO1 e TCF7L2.

Palavras-chave: envelhecimento, diabetes, variabilidade genética, associação genética, SNP.

ABSTRACT

Increase in actions aiming at expanding life expectancy also increase the risk of manifestation of age-related disorders. Type 2 diabetes appears as one of the main chronic diseases associated with aging, characterized by persistent hyperglycemia. This occurs due to impaired insulin secretion by beta-pancreatic cells, resulting in negative impact on life quality of the affected individual, caused by manifestation of several pathologies associated with high blood glucose levels. Elucidation of complex mechanisms which generates disorders of blood glucose homeostasis involves knowledge of genetic influence associated with diabetes manifestation. In this context, the present study aimed to identify genetic variants in coding regions of four genes (FKH1, PDX1, TP53 and TCF7L2) and to estimate genetic diversity parameters inside groups of healthy or type 2 diabetes previously diagnosed individuals, based on the analysis of consensus sequences obtained from next-generation sequencing fragments (reads) available at NCBI database. Reads were mapped to DNA coding sequences (CDS) corresponding to candidate genes using Geneious software. Consensus sequences obtained were aligned using MEGA X software, and main intra and interpopulation diversity descriptors were estimated using DnaSP software. Low nucleotide diversity was observed among populations sampled, although polymorphic sites and different haplotypes were identified. Single nucleotide polymorphisms (SNPs) were identified and correlated with risk of disease manifestation within each population, using SNPStats platform. Results suggest being possible to trace exome variants that can be associated with the risk of type 2 diabetes development in human populations, mainly for FOXO1 and TCF7L2.

Keywords: aging, diabetes, genetic variability, genetic association, SNP.

SUMÁRIO

1 INTRODUÇÃO 11
1.1 Envelhecimento 11
1.2 Diabetes tipo 2 12
1.2.1 Processos fisiológicos e moleculares associados à diabetes tipo 2 12
1.2.2 *Estudos genéticos e genômicos aplicados ao estudo de diabetes tipo 2* 13
2 OBJETIVOS 16
2.1 Objetivo geral 16
2.2 Objetivos específicos 16
3 METODOLOGIA 17
3.1 Escolha dos genes e informações adicionais 17
3.2 Obtenção das sequências nucleotídicas 18
3.3 Mapeamentos dos *reads* e análise das sequências consenso geradas 18
3.4 Alinhamentos múltiplos de sequências nucleotídicas 20
3.5 Análises de variabilidade genética 21
3.6 Identificação de polimorfismos de um nucleotídeo para estudos de associação 22
4 RESULTADOS 22
4.1 Escolha dos genes e informações adicionais 22
4.2 Mapeamentos dos *reads* e análise das sequências consenso geradas 24
4.3 Alinhamentos múltiplos de sequências nucleotídicas 29
4.4 Análises de variabilidade genética 32
4.4.1 Número de sítios polimórficos (S) e número total de mutações (η) 32
4.4.2 Número de haplótipos (h) e diversidade haplotípica (hd) 32
4.4.3 Diversidade nucleotídica (π) e número médio de diferenças nucleotídicas (k) 33
4.4.4 Coeficiente de Watterson (θ-W) 35
4.4.5 Parâmetros de recombinação 35
4.4.6 Teste de neutralidade (Fu & Li) 35
4.5 Identificação de polimorfismos de um nucleotídeo para estudos de associação 36

5 DISCUSSÃO 38
REFERÊNCIAS 41
APÊNDICE A – INFORMAÇÕES COMPLEMENTARES SOBRE O GENE ESCOLHIDO COMO CONTROLE NEGATIVO (P53) 44
ANEXO A – LOCALIZAÇÃO DOS GENES UTILIZADOS NO ESTUDO EM CROMOSSOMOS DE HUMANOS (*Homo sapiens*) 45
ANEXO B – REFERÊNCIA DE NOMENCLATURA DE BASES NUCLEOTÍDICAS 46
ANEXO C – FORMATOS DE ENTRADA E SAÍDA DAS PLATAFORMAS DE ANÁLISE UTILIZADAS PARA O ESTUDO 47

1 INTRODUÇÃO

1.1 Envelhecimento

A idade cronológica, isoladamente, fornece informações limitadas sobre os complexos processos biológicos que levam ao envelhecimento, considerando seu caráter multifatorial. A participação de moduladores genéticos, epigenéticos e fatores ambientais reforça a ideia de integração molecular e metabólica na determinação do envelhecimento biológico, que leva à redução progressiva do potencial replicativo e regenerativo celular (KHAN; SINGER; VAUGHAN, 2017).

As principais manifestações moleculares do envelhecimento envolvem perda de funcionalidade ou instabilidade genômica, alterações epigenéticas, desordem dos mecanismos de regulação do ciclo celular e reparo de DNA, acúmulo de estresses oxidativos e comprometimento das cascatas de sinalização celular (KHAN; SINGER; VAUGHAN, 2017; MARTINS; LITHGOW; LINK, 2015).

Clinicamente, o envelhecimento *per se* representa um fator de risco para a manifestação de doenças crônicas, tais como distúrbios cardiovasculares ou neurodegenerativos (KHAN; SINGER; VAUGHAN, 2017). O declínio progressivo de funções metabólicas e biológicas basais, como manutenção da homeostase, defesa, regeneração e reprodução celular, torna o organismo mais suscetível a estresses e doenças. A exposição prolongada a fatores estressores pode levar ao acúmulo drástico de erros ou mutações que resultam na progressão de patologias associadas ao envelhecimento (MOSKALEV *et al.*, 2014).

Apesar das manifestações do envelhecimento ocorrerem a velocidades e intensidades distintas em diferentes indivíduos, compreender a integração entre os eventos celulares e moleculares que determinam desordens idade-dependentes pode fornecer informações importantes para guiar o diagnóstico e possíveis intervenções terapêuticas direcionadas, individualizadas e mais eficazes (RÓNAI *et al.*, 2018).

1.2 Diabetes tipo 2

1.2.1 Processos fisiológicos e moleculares associados à diabetes tipo 2

A diabetes tipo 2 apresenta-se como uma das principais doenças crônicas associadas ao envelhecimento e é caracterizada pela perda da função das células beta-pancreáticas e desenvolvimento progressivo de resistência à insulina. Nesta condição há disfunção da resposta aos estímulos do hormônio para o processamento adequado da glicose sanguínea, resultando em quadros de hiperglicemia (AL-QUOBAILI; MONTENARH, 2008; PRASAD; GROOP, 2015).

O risco de desenvolvimento e a prevalência da doença resultam de uma complexa interação entre fatores ambientais, genéticos e epigenéticos (PRASAD; GROOP, 2015). Suas implicações na desregulação do metabolismo da glicose a tornam um fator de risco para a ocorrência de casos clínicos mais graves, que incluem disfunção do fígado, doenças cardiovasculares, retinopatia, neuropatia e disfunções cerebrais, como demência associada à doença de Alzheimer (BROWN; WALKER, 2016; PRUZIN *et al.*, 2018).

Os sintomas de quadros mais acentuados de diabetes tipo 2 incluem poliúria (aumento na eliminação de urina), polidipsia (sede excessiva), polifagia (aumento do apetite e ingestão de alimentos), disfunções do sistema imunológico e comprometimento da visão. No entanto, alterações funcionais ocasionadas pela perturbação do metabolismo de carboidratos podem ocorrer sem a manifestação de sintomas clínicos, o que dificulta o diagnóstico precoce da doença (AMERICAN DIABETES ASSOCIATION, 2014).

Uma vez que a secreção de insulina pelas células beta-pancreáticas é cooperativamente coordenada por complexas cascatas de eventos moleculares que envolvem proteínas, fatores de transcrição e outros hormônios, a regulação da expressão de genes associados à modulação da síntese e secreção do hormônio em resposta a um aumento nos níveis glicêmicos é fundamental para a manutenção da homeostase da glicose (AL-QUOBAILI; MONTENARH, 2008).

1.2.2 Estudos genéticos e genômicos aplicados ao estudo da diabetes tipo 2

A arquitetura genética associada à diabetes tipo 2 ainda não se encontra completamente elucidada, visto que há uma complexa rede de interações entre fatores genéticos e ambientais relacionada ao desenvolvimento da patologia. Nesse contexto, o desenvolvimento de técnicas avançadas de sequenciamento e análise simultânea de grandes volumes de dados tornou possível a aplicação de abordagens em larga escala para rastrear variantes genéticas de risco e associá-las à manifestação da doença (PATNALA; CLEMENTS; BATRA, 2013*;* WU *et al.*, 2014; PRASAD; GROOP, 2015; BROWN; WALKER, 2016; RÓNAI *et al.*, 2018).

As plataformas de sequenciamento de nova geração, tais como a plataforma *Illumina* (empresa), permitem o sequenciamento do DNA com base na fragmentação e extensão de sequências específicas na presença de bases modificadas, que geram sinais detectáveis utilizados para a leitura dos fragmentos. Os dados obtidos, produzidos na forma de conjuntos de *reads* (ou subsequências curtas), podem ser utilizados, por exemplo, para mapeamento de genomas com base em referências de regiões conhecidas do DNA. A partir desses mapeamentos é possível rastrear sítios polimórficos que possam ser associados ao risco de manifestação da condição de interesse (ILLUMINA INC, 2010).

Uma vez que o genoma humano se mantém altamente conservado entre indivíduos distintos em populações, o estudo de marcadores polimórficos, como os polimorfismos de um único nucleotídeo (SNPs), permite analisar diretamente sítios de segregação que distinguem a população e compreender desta forma a extensão da influência genética na determinação de doenças. No caso da diabetes tipo 2, polimorfismos em genes relacionados principalmente à síntese e secreção de insulina ou ao funcionamento das células beta-pancreáticas foram previamente relatados como fatores de risco para o desenvolvimento da doença (BROWN; WALKER, 2016; RÓNAI *et al.*, 2018).

Muller *et al.* (2015) reportaram a associação de um SNP (rs2297627) presente na região intrônica do gene FOXO1 (*Forkhead box* O 1), com o risco de desenvolvimento de diabetes tipo 2 em uma população de 7.710 ameríndios (OR, *odds-ratio* = 1,19, valor de referência: OR > 1; p-valor, significância = 1×10^{-4}, valor de

referência: p-valor < 0,05). FOXO1 é um importante efetor da sinalização de insulina e IGF-1 (*Insulin Growth Factor* 1) envolvido na modulação de processos celulares cruciais de resposta a estresses, metabolismo, ciclo celular e apoptose (KATOH *et al.*, 2013; MARTINS; LITHGOW; LINK, 2015).

Steinthorsdottir *et al.* (2015) identificaram uma variante rara (c.651delT) que implica em alteração na janela de leitura (*frameshift mutation*) do gene PDX1 (*Pancreatic and duodenal homeobox fator* 1), associada a alto risco de manifestação de diabetes tipo 2 (OR = 1,45; p-valor = 8×10^{-3}) em uma população de 278.254 islandeses. PDX1, expresso em células beta-pancreáticas, atua tanto na regulação do transporte quanto na fosforilação da glicose. Em resposta a altos níveis de glicose, tem papel na ativação da transcrição de genes de sinalização e síntese de insulina, como o transportador celular GLUT-2 e a enzima hexoquinase (AL-QUOBAILI; MONTENARH, 2008; HARGREAVES *et al.*, 2017).

Barra *et al.* (2012) investigaram a associação de um SNP (rs7903146) presente na região intrônica do gene TCF7L2 (*Transcription factor* 7-*like* 2) com a ocorrência de diabetes tipo 2 em uma população de 252 brasileiros (OR = 1,5; p-valor = $3,2 \times 10^{-2}$). O SNP é reportado como tendo a associação mais consistente com a doença entre populações com diferentes ancestralidades. O fator de transcrição TCF7L2 atua na cascata de sinalização mediada pela família de proteínas Wnt, envolvida na regulação da proliferação, diferenciação e apoptose celular, e é um regulador chave da síntese e processamento de insulina, através da regulação de genes como PDX1. A ativação transcricional mediada por TCF7L2 regula positivamente a expressão de moléculas-sinal relacionadas às vias de sinalização do metabolismo, processamento e secreção da insulina e manutenção da homeostase da glicose (LYSSENKO *et al.*, 2007; SUDCHADA; SCARPACE, 2014; ZHOU *et al.*, 2014).

Embora não haja, até o momento, conhecimento sobre a totalidade da influência genética associada ao desenvolvimento de doenças complexas, estudos de associação fornecem uma base interessante para a identificação progressiva de novos candidatos biológicos e compreensão da sua contribuição funcional à patogênese da desordem investigada (PATNALA; CLEMENTS; BATRA, 2013; PRASAD; GROOP, 2015; BROWN; WALKER, 2016; RÓNAI *et al.*, 2018).

Considerando a diversidade da influência genética nos mecanismos de desenvolvimento de diabetes tipo 2, o presente estudo objetivou investigar o impacto de variantes genéticas em regiões codificadoras de quatro genes (FOXO1, PDX1, TP53 e TCF7L2) para a manifestação da patologia em uma população de humanos.

.

2 OBJETIVOS

2.1 Objetivo geral

Investigar o impacto de polimorfismos em regiões codificadoras de quatro genes (FKH1, PDX1, TP53 e TCF7L2) sobre a variabilidade genética de populações de humanos e na manifestação de doenças associadas ao envelhecimento em humanos, especificamente a diabetes tipo 2.

2.2 Objetivos específicos

Mapear regiões codificadoras de quatro genes (FOXO1, PDX1, TP53 e TCF7L2) em subpopulações de humanos sadios (controles) ou diagnosticados com diabetes tipo 2 (casos);

Validar a identidade das sequências obtidas com as regiões de interesse, correspondentes ao exoma de humanos;

Estimar parâmetros de diversidade genética intra e interpopulações;

Comparar as sequências de DNA e identificar sítios polimórficos nas populações de casos e controles;

Investigar a associação entre polimorfismos de um único nucleotídeo (SNPs) e a probabilidade de manifestação de diabetes tipo 2 em humanos.

3 METODOLOGIA

3.1 Escolha dos genes e informações adicionais

Para a execução das análises *in silico*, foram escolhidos 4 (quatro) genes: 2 (dois) candidatos, 1 (um) controle positivo e 1 (um) controle negativo, com base em referencial bibliográfico obtido dos bancos de publicações científicas PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) e Google Acadêmico (https://scholar.google.com.br/), e do banco de dados de polimorfismos de um único nucleotídeo, dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/).

O critério fundamental de seleção dos genes candidatos PDX1 (*Pancreatic and duodenal homeobox fator* 1) e FOXO1 (*Forkhead box* O 1) consistiu na correlação entre seus papéis funcionais e os possíveis mecanismos de desenvolvimento da patologia, com base em associação preliminar entre variantes genéticas presentes nas regiões com o risco de manifestação de diabetes tipo 2 em população de humanos (MULLER *et al.*, 2015; STEINTHORSDOTTIR *et al.*, 2015).

O gene TCF7L2 (*Transcription factor* 7-*like* 2) foi selecionado como controle positivo para este estudo por ter sido previamente relatado como gene cujas variantes genéticas apresentam, até o momento, a maior associação com a manifestação de diabetes tipo 2 em populações de humanos (LYSSENKO *et al.*, 2007; BROWN; WALKER, 2016). O gene P53 (*Tumor protein p*53) foi selecionado como controle negativo por não haver, até o momento, estudos mostrando a associação de variantes do gene com a ocorrência da doença (ver **Apêndice A**).

Os dados de localização de cada gene de interesse foram obtidos do banco de genomas Ensembl (https://www.ensembl.org/index.html), usando como referência a versão GRCh38 do genoma humano. Os dados de comprimento das sequências nucleotídicas de referência foram obtidos do banco de sequências NCBI *Nucleotide* (https://www.ncbi.nlm.nih.gov/nuccore).

3.2 Obtenção das sequências nucleotídicas

As sequências de CDS (*Coding DNA Sequences*) de referência de humanos utilizadas para mapeamento dos *reads* foram obtidas do banco NCBI *Nucleotide* em formato FASTA. Os dados de SRA (*Sequence Read Archives*) foram obtidos do estudo realizado por A Fakhro *et al.* (2016). O projeto, disponível na plataforma NCBI SRA (https://www.ncbi.nlm.nih.gov/sra), contém dados de sequenciamento de exoma de indivíduos adultos do Oriente Médio, saudáveis ou diagnosticados com diabetes tipo 2.

Todos os critérios de obtenção das sequências nucleotídicas foram padronizados durante a busca para que fossem obtidos apenas dados públicos de exoma (conjunto total de éxons do genoma) de humanos.

Os conjuntos de dados selecionados para as análises posteriores encontram-se relacionados na **Tabela 1**. Foram coletados dados de dois grupos: um grupo de 4 (quatro) indivíduos saudáveis (grupo controle), e um grupo de 12 (doze) indivíduos diagnosticados com diabetes tipo 2 (grupo caso). Ambos os grupos apresentam a mesma proporção entre indivíduos do sexo feminino e masculino (1:1).

Para cada indivíduo, os dados de *reads* dos cromossomos correspondentes às regiões genômicas contendo os genes de interesse (10, 13 e 17) foram coletados da plataforma SRA *Run Browser* (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser) em formato FASTQ.

3.3 Mapeamentos dos *reads* e análise das sequências consenso geradas

Os *reads* foram mapeados nas sequências nucleotídicas de referência utilizando o *software* privativo Geneious (https://www.geneious.com/), a fim de construir sequências consenso entre as referências e os fragmentos de sequenciamento (*reads*).

Tabela 1 – Caracterização dos conjuntos de *reads* utilizados para o estudo

Sexo	Característica	Arquivo SRA
Feminino	Diabético	SRR2125398, SRR2125405 SRR2125425, SRR2125435 SRR2130707, SRR2130893
	Saudável	SRR2130945; SRR2130912
Masculino	Diabético	SRR2130875; SRR2130935 SRR2125424; SRR2125426 SRR2125434; SRR2130878
	Saudável	SRR5264034; SRR2130908

Fonte: Elaborado pela autora. Dados obtidos da plataforma NCBI SRA.

Os parâmetros de mapeamento utilizados foram:

(1) Tecnologia de sequenciamento dos *reads*: Illumina

(2) Tipo de sequenciamento: Pair-end

(3) Tamanho do inserto: 101 pb

(4) Limiar de geração da sequência consenso: Highest quality

As sequências consenso obtidas foram armazenadas em formato FASTA e submetidas à análise para confirmação de sua identidade, através da plataforma NCBI, utilizando a ferramenta BLASTn (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch), como etapa de verificação da qualidade dos mapeamentos.

Os parâmetros de análise utilizados foram:

(1) Base de dados: Nucleotide colection

(2) Organismo: *Homo sapiens*

(3) Programa: Optimize for highly similar sequences (megablast)

Em seguida, a ferramenta ORF *finder* (https://www.ncbi.nlm.nih.gov/orffinder/) foi utilizada para identificar janelas abertas de leitura (ORF) para obtenção das sequências de aminoácidos correspondentes, como etapa adicional de verificação e identificação de regiões de interrupção de ORFs. As regiões de baixa cobertura de mapeamento, identificadas ao longo das sequências consenso pelo símbolo “?”, foram removidas antes de cada análise.

Para cada sequência, foi verificada a ORF com maior janela de leitura nos sentidos +1, e esta foi submetida à análise para determinação da identidade utilizando a ferramenta BLASTp (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins).

Os parâmetros de análise utilizados foram:

(1) Base de dados: Non-redundant protein sequences

(2) Organismo: *Homo sapiens*

(3) Programa: BLASTp (protein-protein BLAST)

3.4 Alinhamentos múltiplos de sequências nucleotídicas

As sequências consenso nucleotídicas obtidas para cada gene alvo foram alinhadas utilizando a ferramenta MUSCLE (*Multiple Sequence Comparison by Log Expectation*), interna ao software MEGA X (https://www.megasoftware.net/).

Os alinhamentos múltiplos gerados foram manualmente curados para aumentar a confiabilidade dos resultados gerados. Sequências de baixa qualidade ou regiões com excesso de bases redundantes foram removidas. A referência utilizada para nomenclatura das bases redundantes está relacionada no **Anexo B.**

Uma vez que a etapa de refinamento dos alinhamentos englobou a remoção de regiões internas à porção codificante (CDS), não foram feitas inferências sobre o efeito dos polimorfismos na sequência primária dos produtos proteicos.

3.5 Análises de variabilidade genética

A partir dos alinhamentos obtidos, o *software* DnaSP (DNA *Sequence Polymorphism* – http://www.ub.edu/dnasp/) foi utilizado para determinar os principais descritores gerais de variabilidade nucleotídica entre as sequências, conforme relacionados abaixo:

- **Número de sítios polimórficos (S):** quantidade de sítios que possuem quaisquer variações entre as sequências comparadas;

- **Diversidade nucleotídica (π):** número médio de diferenças nucleotídicas para cada comparação par a par possível entre as sequências;

- **Número total de mutações (η):** total de variações nucleotídicas observadas ao longo das sequências;

- **Número de haplótipos (h):** quantidade de sequências distintas obtidas a partir de diferenças nucleotídicas;

- **Diversidade haplotípica (hd):** probabilidade de observação de dois haplótipos diferentes dentro do espaço amostral;

- **Número médio de diferenças nucleotídicas (k):** média de diferenças nucleotídicas entre todas as sequências;

- **Coeficiente de Watterson (θ-W):** coeficiente de variabilidade genética, relação entre número efetivo da população e sua taxa de mutação, calculado sítio a sítio ou entre sequências.

Além de inferências de diversidade, foram determinados, ainda utilizando o *software* DnaSP, os coeficientes estatísticos D* e F* de Fu & Li (1993), para testar a hipótese de seleção neutra para as mutações observadas.

O parâmetro R (taxa de recombinação), que relaciona o tamanho da população com a média de diferenças nucleotídicas, também foi estimado.

Os possíveis sítios de recombinação (*hotspots*) intra e interpopulações foram identificados.

3.6 Identificação de polimorfismos de um nucleotídeo para estudos de associação

Para cada gene de interesse, as sequências consenso dos indivíduos caso e controle foram alinhadas com as sequências CDS de referência, seguindo método descrito no item 3.4.

Os blocos de SNPs foram extraídos dos alinhamentos múltiplos gerados utilizando a ferramenta SNP-sites (PAGE *et al.*, 2016). Para cada sítio polimórfico, os nucleotídeos correspondentes aos alelos divergentes à referência foram mapeados. Em seguida, a plataforma SNPStats (SOLÉ *et al.*, 2006) foi utilizada para a análise da associação entre os polimorfismos e os grupos-resposta (caso e controle).

Os seguintes parâmetros estatísticos foram estimados:

- ***Odds ratio* (OR), razão de probabilidade:** parâmetro de associação, determina a magnitude do impacto de determinada variável para a manifestação de uma característica-resposta, com intervalo de confiança de 95% (SZUMILAS, 2010);

- **p-valor:** significância estatística.

4 RESULTADOS

4.1 Escolha dos genes e informações adicionais

Os dados de localização, comprimento de sequência nucleotídica e associações utilizadas como critério de seleção dos genes de interesse aplicados para o estudo estão relacionados na **Tabela 2**.

Tabela 2 – Dados de localização, comprimento de sequência e associações utilizadas como critério de seleção dos genes de interesse aplicados para o estudo.

Gene	Localização	Sequência nucleotídica codificante (pb)	Associações identificadas	Aplicação
TCF7L2 (*transcription factor* 7 *like* 2)	10q25.2-q25.3	1791	Aterosclerose; Diabetes gestacional; Diabetes tipo 2; Obesidade; Retinopatia diabética	Controle positivo
PDX1 (*pancreatic duodenal homeobox factor* 1)	13q12.2	852	Câncer pancreático; Diabetes tipo 2	Gene teste
FOXO1 (*Forkhead box* O 1)	13q14.11	1968	Diabetes tipo 2; Nefropatia diabética; Obesidade	Gene teste
TP53 (*tumor protein* p53)	17p13.1	1065	Câncer de mama; Câncer de ovário; Câncer retal; Glioma; Leucemia; Osteosarcoma	Controle negativo

Fonte: Elaborado pela autora. A nomenclatura dos *loci* gênicos segue o padrão: número do cromossomo + braço do cromossomo onde se encontra o gene + posição no braço, em ordem crescente a partir do centrômero. Dados obtidos do banco de genomas Ensembl, do banco de sequências nucleotídicas NCBI *Nucleotide* e do banco de SNPs dbSNP.

4.2 Mapeamento dos *reads* e análise das sequências consenso geradas

Os mapeamentos dos *reads* contra as sequências de referência de cada gene estão exemplificados nas **Figuras 1-4**. Apesar da boa qualidade observada para todos os conjuntos de *reads*, representada pela coloração verde em cada fragmento das figuras, a menor disponibilidade de dados para PDX1 refletiu na redução da cobertura dos mapeamentos realizados para o gene. Enquanto para os demais genes cada conjunto de *reads* continha em torno de 800 a 2300 fragmentos, a maior quantidade observada para PDX1 foi de 513, porém com maioria entre 100 e 200. Tal observação pode ser justificada pela limitação da capacidade de ferramentas de sequenciamento em alcançar igualmente regiões diferentes do exoma.

Para os demais genes, os mapeamentos apresentaram cobertura satisfatória, sendo a sequência consenso majoritariamente coberta por pelo menos um *read* em cada ponto. Apesar do algoritmo do *software* otimizar a qualidade de resolução dos consensos, as regiões de baixa cobertura dos mapeamentos resultaram na geração de sítios com excesso de bases redundantes, ou não informativos (identificados pelo símbolo “?”).

Em todos os genes, o alinhamento dos *reads* extrapolou a região de referência em ambas as extremidades 5’ e 3’, gerando consensos de comprimento superior ao das sequências de referência. No entanto, ambas as etapas de verificação validaram a qualidade dos mapeamentos dos *reads*. A análise executada na plataforma BLASTn confirmou a identidade das sequências consenso com os genes de referência, sendo observado percentual de identidade inferior a 95% com as referências do NCBI apenas em algumas sequências dos genes P53 (92-94%) e TCF7L2 (92-93%).

No ORF finder, apenas as sequências com sítios de baixa confiabilidade produziram janelas de leitura de tamanho inferior ao comprimento das sequências de aminoácidos equivalentes às referências. Ainda assim, a análise na plataforma BLASTp das ORFs geradas permitiu confirmar a identidade de todas as sequências com os produtos proteicos de referência.

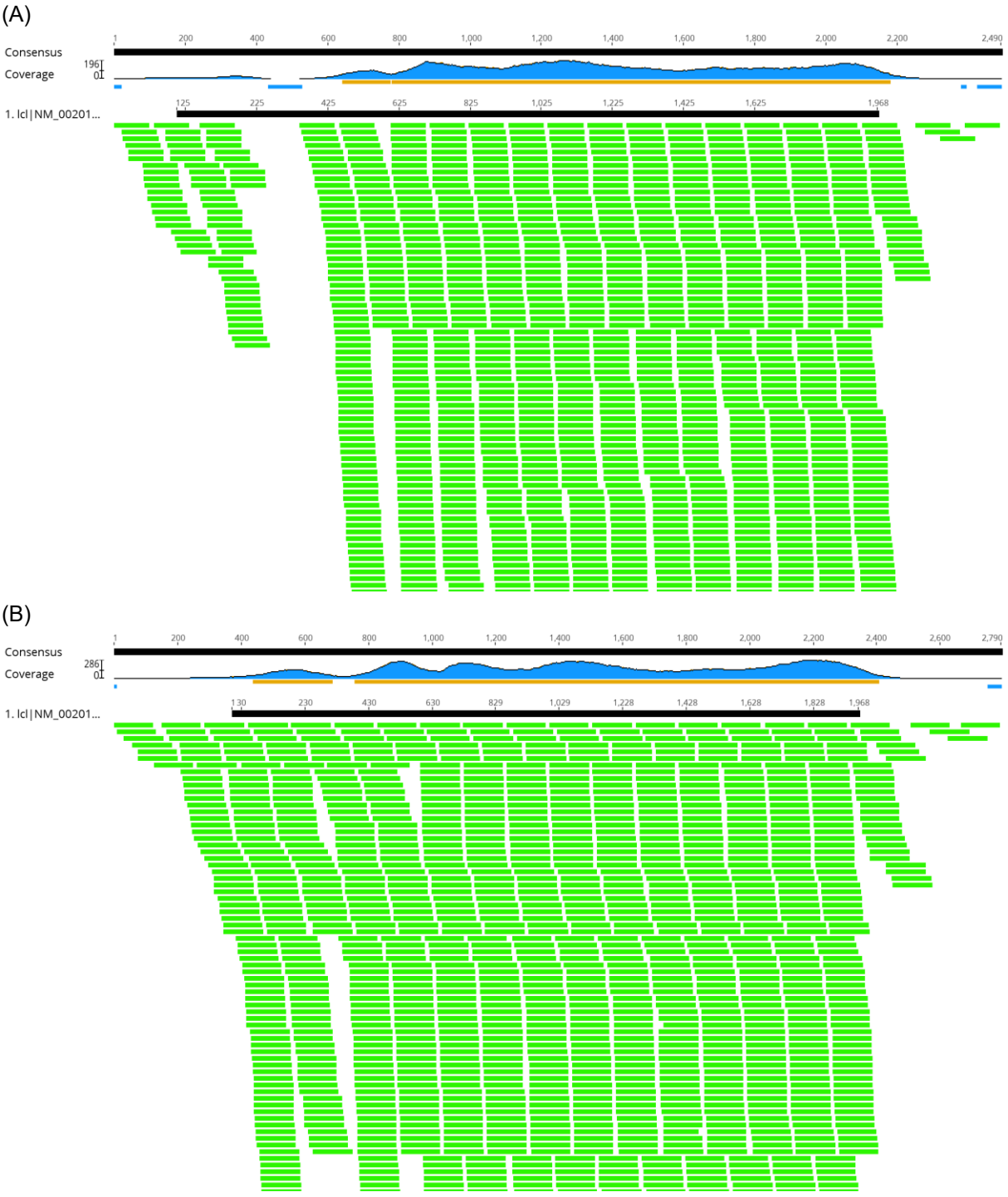
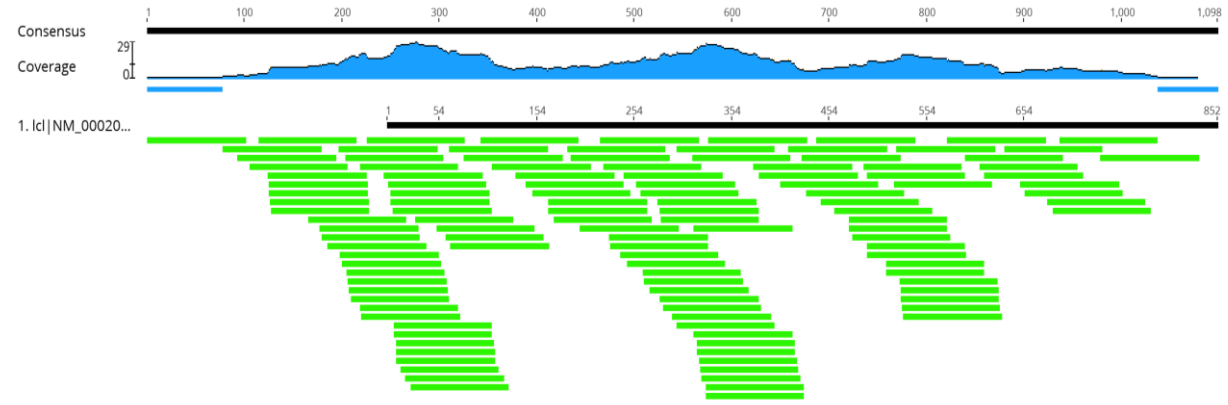


Figura 1. Mapeamento dos *reads* de indivíduos controle (A, amostra SRR2130945) e caso (B, amostra SRR2125398) contra sequência de referência do gene FOXO1. A barra preta superior representa a sequência consenso gerada; o gráfico em azul, a faixa de cobertura do mapeamento; as barras azul e amarela abaixo, as regiões de mínima e máxima cobertura, respectivamente; a barra preta inferior, a sequência de referência; e os retângulos verdes, o conjunto de *reads* utilizados para o mapeamento. Os *reads* foram coloridos tendo como base a qualidade do sequenciamento, variando de verde a vermelho para alta e baixa, respectivamente. Visualização do *software* Geneious.

(A)



(B)

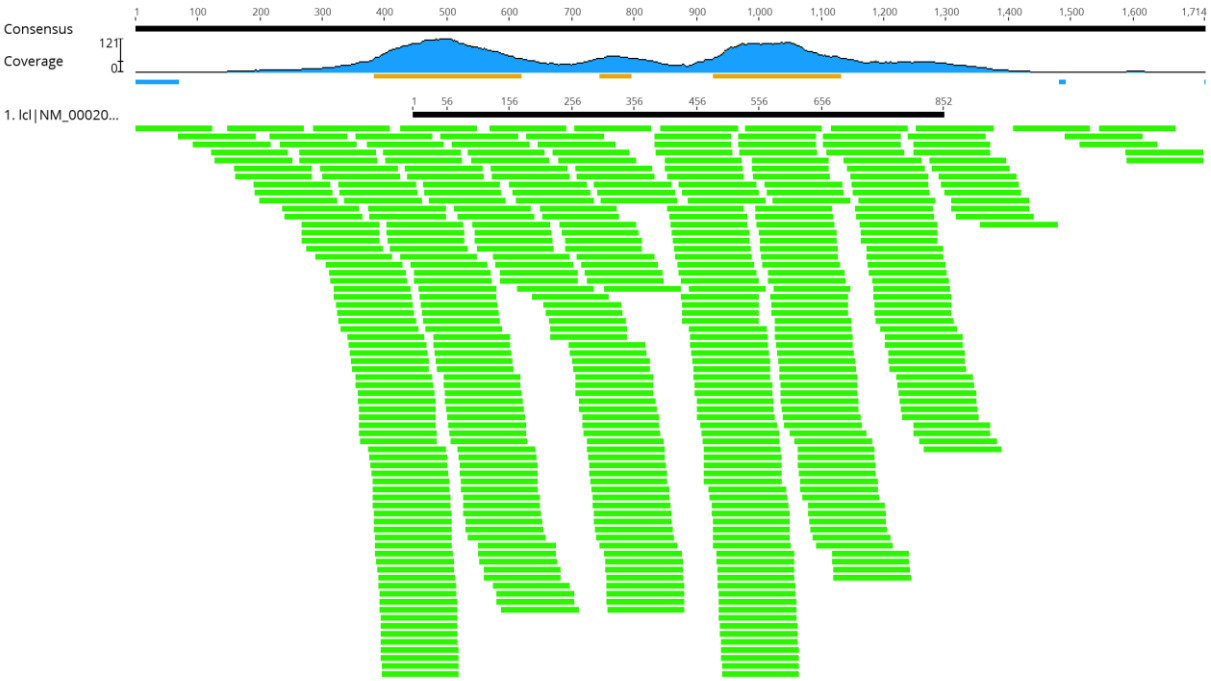
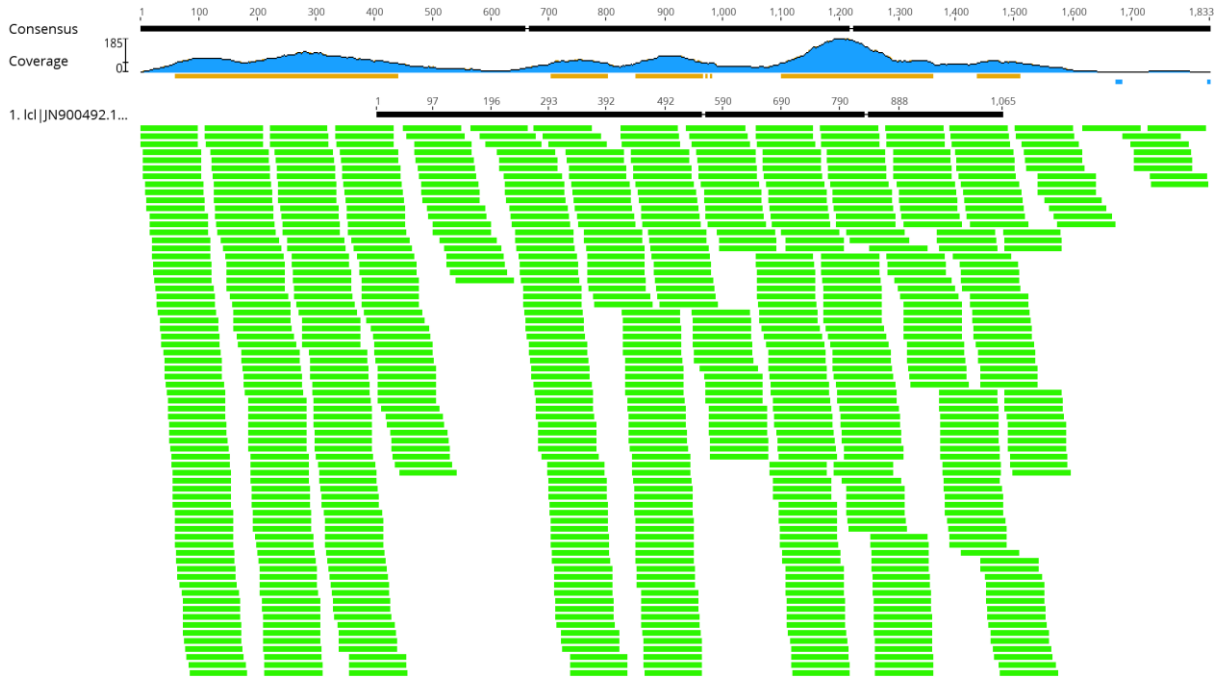


Figura 2. Mapeamento dos *reads* de indivíduos controle (A, amostra SRR5264034) e caso (B, amostra SRR2125398) contra sequência de referência do gene PDX1. A barra preta superior representa a sequência consenso gerada; o gráfico em azul, a faixa de cobertura do mapeamento; as barras azul e amarela abaixo, as regiões de mínima e máxima cobertura, respectivamente; a barra preta inferior, a sequência de referência; e os retângulos verdes, o conjunto de *reads* utilizados para o mapeamento. Os *reads* foram coloridos tendo como base a qualidade do sequenciamento, variando de verde a vermelho para alta e baixa, respectivamente. Visualização do *software* Geneious.

(A)



(B)

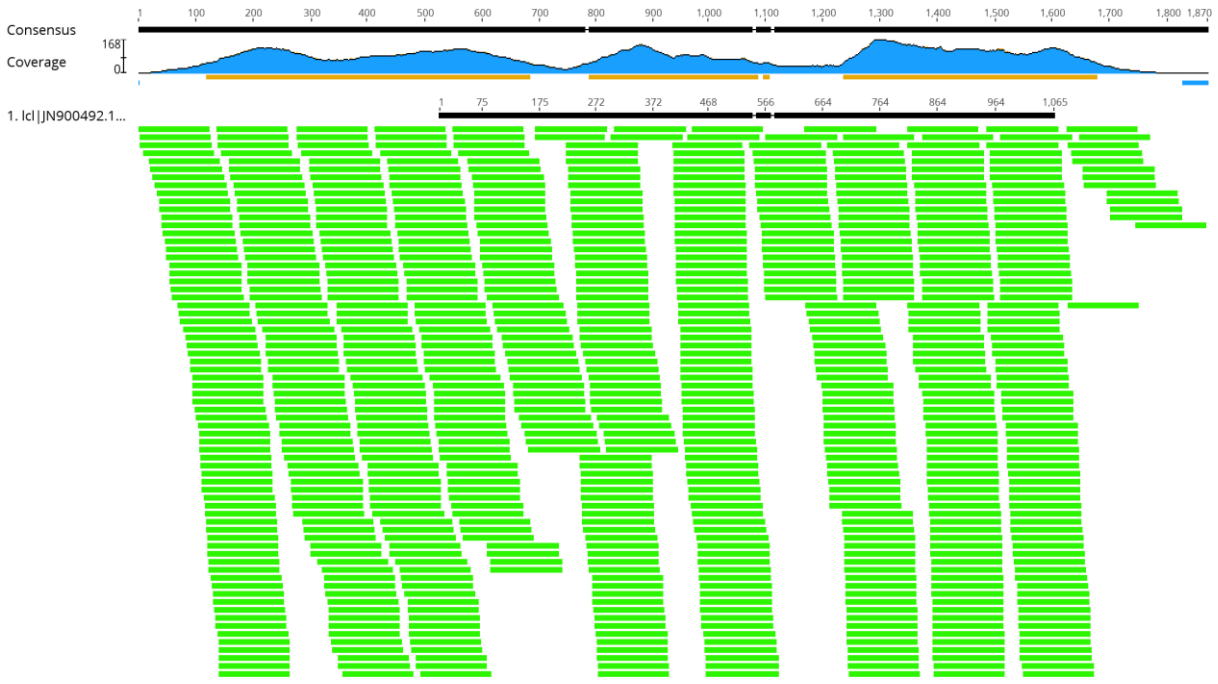


Figura 3. Mapeamento dos *reads* de indivíduos controle (A, amostra SRR2130945) e caso (B, amostra SRR2125398) contra sequência de referência do gene P53. A barra preta superior representa a sequência consenso gerada; o gráfico em azul, a faixa de cobertura do mapeamento; as barras azul e amarela abaixo, as regiões de mínima e máxima cobertura, respectivamente; a barra preta inferior, a sequência de referência; e os retângulos pretos, o conjunto de *reads* utilizados para o mapeamento. Os *reads* foram coloridos tendo como base a qualidade do sequenciamento, variando de verde a vermelho para alta e baixa, respectivamente. Visualização do *software* Geneious.

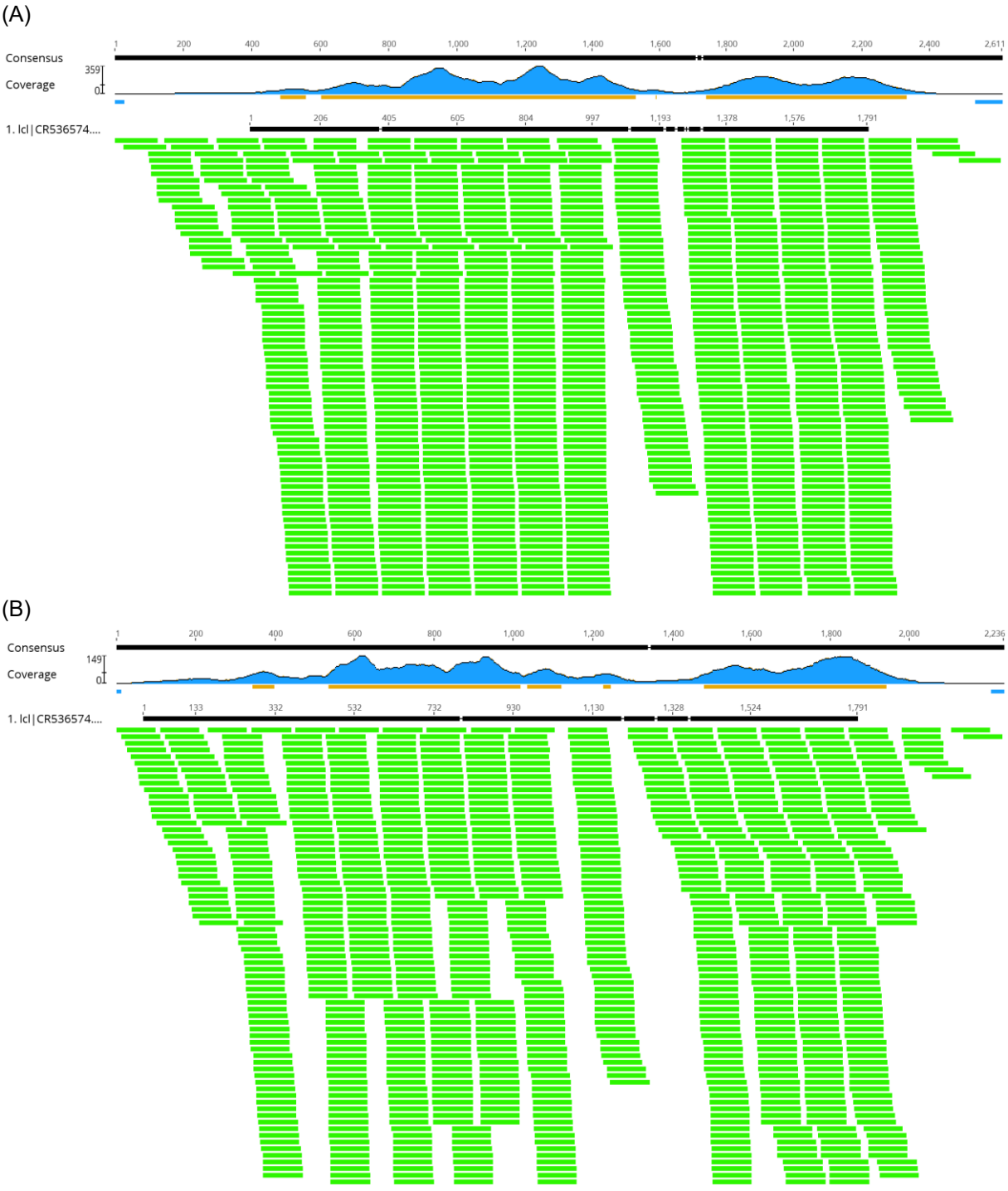


Figura 4. Mapeamento dos *reads* de indivíduos controle (A, amostra SRR2130912) e caso (B, amostra SRR2125398) contra sequência de referência do gene TCF7L2. A barra preta superior representa a sequência consenso gerada; o gráfico em azul, a faixa de cobertura do mapeamento; as barras azul e amarela abaixo, as regiões de mínima e máxima cobertura, respectivamente; a barra preta inferior, a sequência de referência; e os retângulos pretos, o conjunto de *reads* utilizados para o mapeamento. Os *reads* foram coloridos tendo como base a qualidade do sequenciamento, variando de verde a vermelho para alta e baixa, respectivamente. Visualização do *software* Geneious.

4.3 Alinhamentos múltiplos das sequências nucleotídicas

Os alinhamentos entre as sequências nucleotídicas consenso dos grupos (subpopulações) de casos e controles estão exemplificados nas **Figuras 5 e 6**.

A **Figura 6-A** ilustra uma região de completa homogeneidade, em que todas as posições nucleotídicas são idênticas entre os indivíduos (sítios monomórficos). Os mapeamentos apresentaram majoritariamente regiões com este perfil de alinhamento.

As posições onde houve divergência entre as sequências nucleotídicas e o consenso em pelo menos um indivíduo foram identificadas pela cor e nomenclatura correspondentes à base discordante. A nomenclatura segue padrão relacionado no **Anexo B**.

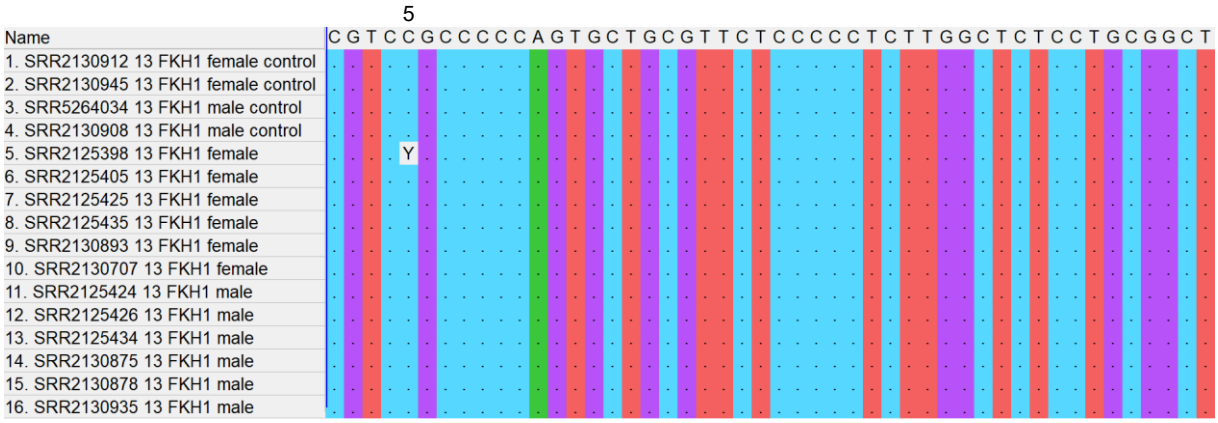
Três casos de heterogeneidade entre as sequências nucleotídicas foram observados (sítios polimórficos). No primeiro, exemplificado pelas **Figuras 5-A** (posição 5) e **6-C** (posição 48), apenas um indivíduo apresentou diferença nucleotídica em relação aos demais (*singleton*), o que caracteriza o sítio como pouco informativo.

No segundo, ilustrado pelas **Figuras 5-B** (posição 17) e **6-B** (posição 1), tal diferença pode ser observada para pelo menos dois indivíduos em uma mesma posição, o que caracteriza o sítio como parcimonioso, ou informativo.

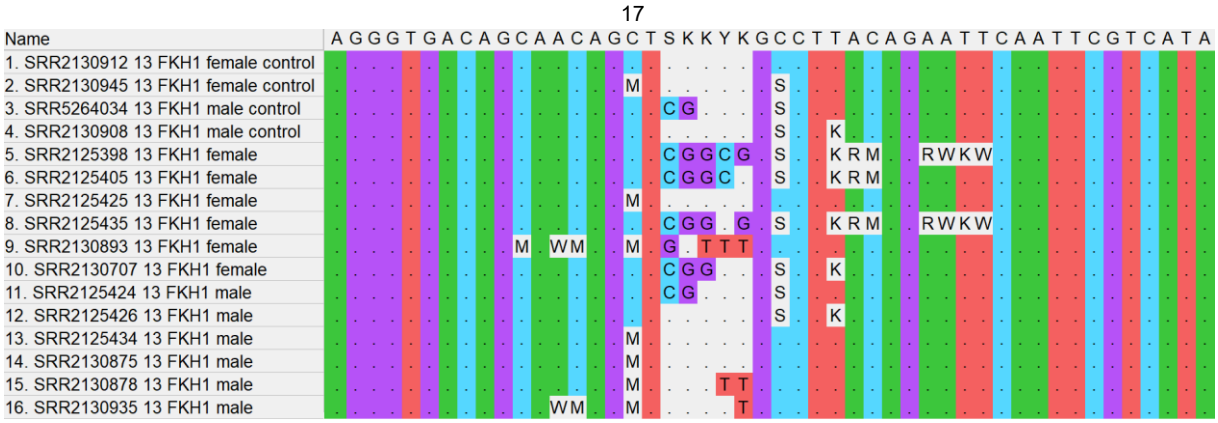
No último caso, ilustrado pela **Figura 7**, observa-se uma região marcada pelo excesso de bases redundantes ou grande divergência entre as posições nucleotídicas, regiões estas condizentes com o espectro de baixa cobertura dos mapeamentos realizados pelo Geneious.

O alinhamento múltiplo resultou em sequências nucleotídicas de comprimento similar ao das regiões de referência.

(A)



(B)



(C)

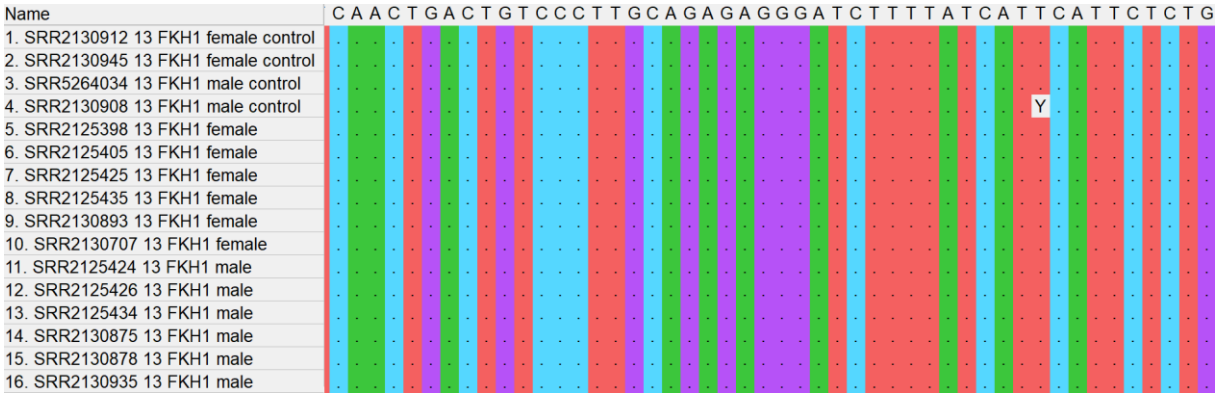
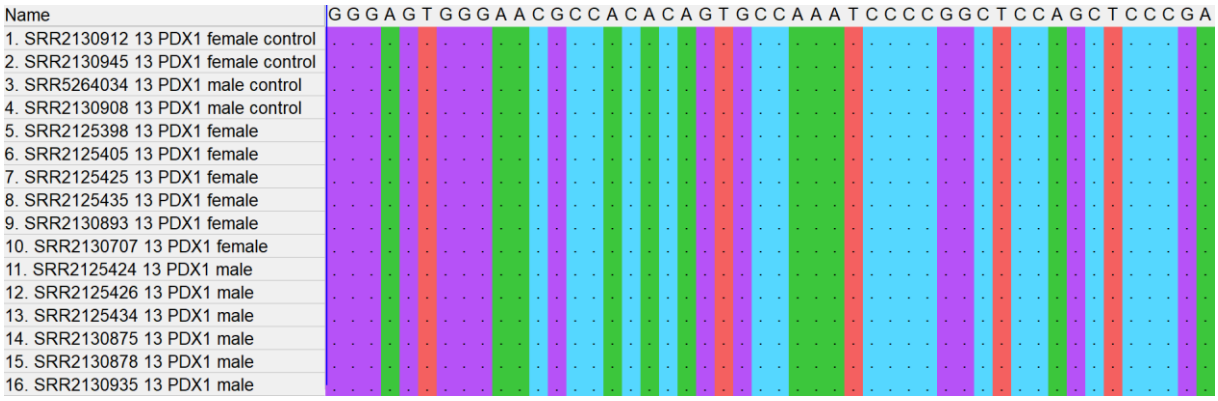
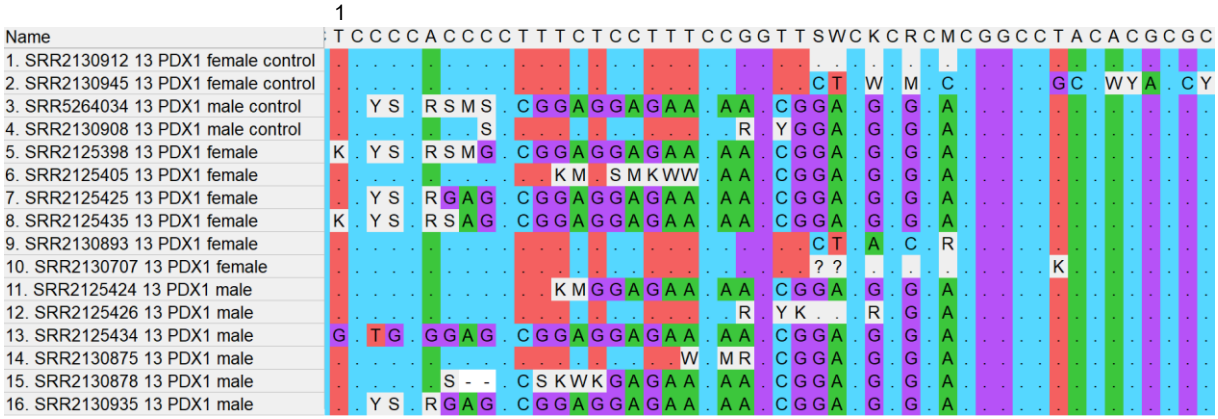


Figura 5. Alinhamento múltiplo de sequências para o gene FOXO1 após refinamento. Visualização parcial do início (A, sítios 1-48), regiões polimórficas (B, sítios 564-611), e término do alinhamento (C, sítios 2041-2088). A primeira linha, em cinza, representa o consenso entre as sequências nucleotídicas dos indivíduos. Os sítios identificados pelo símbolo "•" representam posições onde as sequências nucleotídicas dos indivíduos correspondem à mesma base do consenso gerado pelo alinhamento, sendo este a base de maior frequência para cada posição. Visualização do *software* MEGA X.

(A)



(B)



(C)

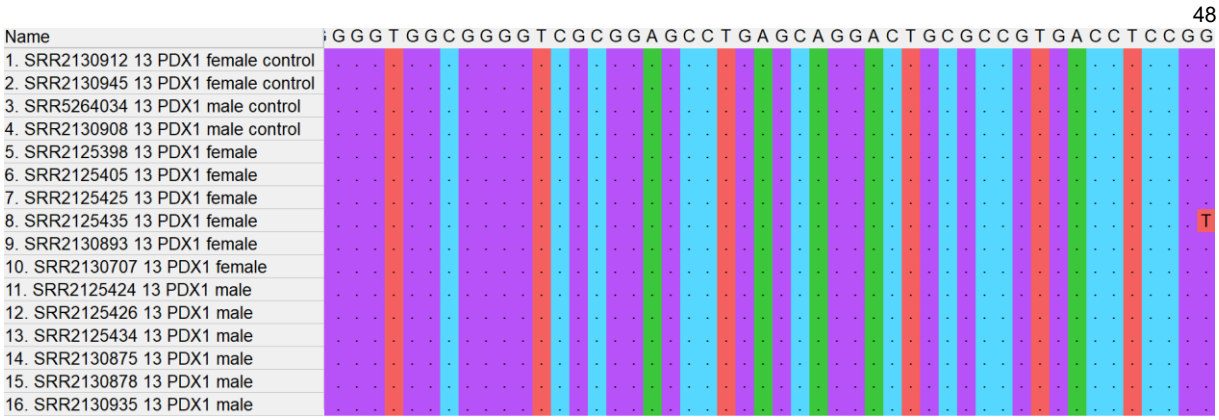


Figura 6. Alinhamento múltiplo de sequências para o gene PDX1 após refinamento. Visualização parcial do início (A, sítios 1-48), regiões polimórficas (B, sítios 516-563), e término do alinhamento (C, sítios 746-793). A primeira linha, em cinza, representa o consenso entre as sequências nucleotídicas dos indivíduos. Os sítios identificados pelo símbolo "•" representam posições onde as sequências nucleotídicas dos indivíduos correspondem à mesma base do consenso gerado pelo alinhamento, sendo este a base de maior frequência para cada posição. Visualização do *software* MEGA X.

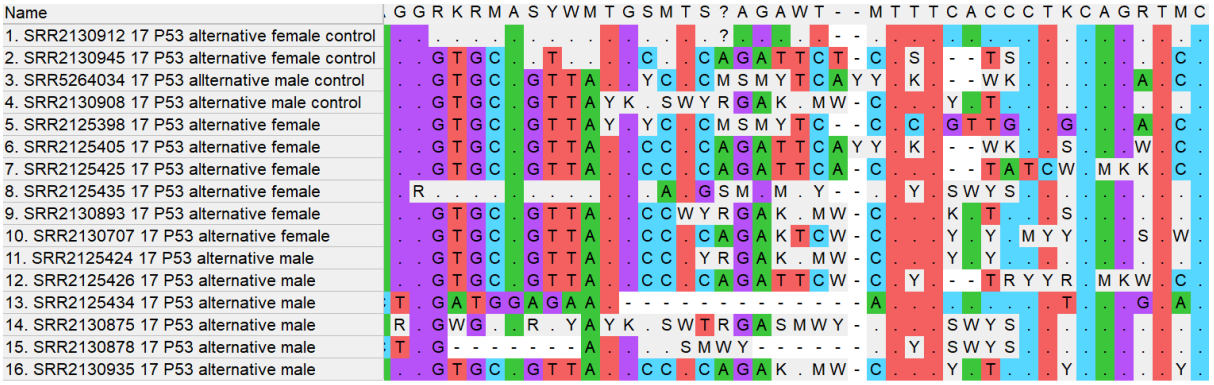


Figura 7. Região de baixa confiabilidade em alinhamento múltiplo de sequências para o gene P53. Visualização do *software* MEGA X.

4.4 Análises de variabilidade genética

Os principais descritores de variabilidade genética estimados para as subpopulações de casos e controles encontram-se relacionados na **Tabela 3**.

4.4.1 Número de sítios polimórficos (S) e número total de mutações (η)

O número máximo de sítios polimórficos (S) e mutações (η) entre as sequências nucleotídicas foi observado para os genes P53 e TCF7L2, utilizados respectivamente como controle negativo e positivo (**Tabela 3**). Para P53, a subpopulação de casos apresentou valores aproximadamente 2 (duas) vezes superiores à subpopulação de controles, para ambos os descritores. Para TCF7L2, os valores mantiveram-se próximos ao comparar as duas subpopulações.

As regiões correspondentes aos genes teste FOXO1 e PDX1 apresentaram menor abundância de sítios polimórficos quando em comparação às demais regiões genômicas.

Para o gene PDX1, os valores de ambos os descritores se mantiveram similares entre as duas populações. Em FOXO1, observou-se maior quantidade de sítios polimórficos e total de mutações para a subpopulação de casos.

4.4.2 Número de haplótipos (h) e diversidade haplotípica (hd)

Em todas as regiões genômicas, observou-se valor máximo de número de haplótipos (h) e diversidade haplotípica (hd) para as populações controle (valores de referência: $h_{máx}$ = número total de sequências; $hd_{máx}$ = 1,0), enquanto entre os casos, os valores foram máximos apenas para os genes P53 e TCF7L2 (**Tabela 3**).

Para os genes FOXO1 e PDX1, os valores de diversidade haplotípica (hd) entre as populações de casos foram mais baixos, mas próximos a 1 (um). O número de haplótipos também foi mais baixo, mas se aproximou do número máximo.

4.4.3 Diversidade nucleotídica (π) e número médio de diferenças nucleotídicas (k)

Para todas as regiões genômicas, a diversidade nucleotídica entre pares de sequência (π) apresentou valores aproximados entre as populações de casos e controles (**Tabela 3**). Todos os valores estimados para π se aproximaram de 0 (zero), sendo mais baixos para as populações de casos e controles de FOXO1.

O número médio de diferenças nucleotídicas entre todas as sequências (k) também apresentou valores mais baixos para as populações de casos e controles de FOXO1, sendo máximo para a população controle de TCF7L2.

4.4.4 Coeficiente de Watterson (θ-W)

O descritor de variabilidade genética $\theta\text{-}W_{sítio}$ apresentou valores próximos a 0 (zero) em todas as regiões genômicas, sendo menor para as populações de casos e controles de FOXO1, de forma similar ao observado para π (**Tabela 3**).

O descritor de variabilidade entre sequências $\theta\text{-}W_{sequência}$ também apresentou valores mais baixos para as populações de casos e controles de FOXO1, sendo máximo para a população controle de TCF7L2, comportamento semelhante ao de k.

Tabela 3 – Descritores de variabilidade genética para os genes FOXO1, PDX1, P53 e TCF7L2 em populações de casos e controles para diabetes tipo 2.

Região genômica	População	Número de sequências	Comprimento da região (pb)	S	η	h	hd	π	k	θ-W (sítio)	θ-W (sequência)
FOXO1	Caso	24	2088	19	19	14	0,942	0,0025	4,500	0,00244	5,088
	Controle	8	2088	10	10	8	1,000	0,00216	5,214	0,00185	3,857
	Total	32	2088	21	21	17	0,950	0,00238	4,97	0,0025	5,214
PDX1	Caso	24	793	28	33	20	0,982	0,01295	10,246	0,00948	7,498
	Controle	8	793	32	34	8	1,000	0,0168	13,321	0,01556	12,342
	Total	32	793	34	39	24	0,980	0,01441	11,401	0,01067	8,442
P53	Caso	24	1469	101	113	24	1,000	0,01432	21,025	0,01842	27,047
	Controle	8	1469	51	55	8	1,000	0,01321	19,393	0,0134	19,669
	Total	32	1469	115	127	32	1,000	0,01433	20,786	0,01945	28,555
TCF7L2	Caso	22	1954	84	90	22	1,000	0,00896	17,468	0,01182	23,043
	Controle	8	1954	82	85	8	1,000	0,01399	27,321	0,01619	31,625
	Total	30	1954	107	116	30	1,000	0,01018	19,841	0,01386	27,009

Legenda: Número de sítios polimórficos (S), Número total de mutações (η), Número de haplótipos (h), Diversidade haplotípica (hd), Diversidade nucleotídica (π), Número médio de diferenças nucleotídicas (k), Coeficiente de Watterson (θ-W). Fonte: Elaborado pela autora. Dados obtidos do *software* DnaSP.

4.4.5 Parâmetros de recombinação

A estimativa dos parâmetros de recombinação está relacionada na **Tabela 4.**

Em todas as regiões genômicas, observou-se valores mais altos de recombinação total (R-gene) nas populações sadias, sendo o valor mais alto observado para TCF7L2 (R-gene = 303).

Os valores de recombinação média por sítio (R-sítio) se aproximaram de 0 (zero) em todos os genes, sendo o maior valor observado para a subpopulação sadia de TCF7L2 (R-sítio = 0,1552).

A variância dos dados apresentou valores mais altos para os genes P53 e TCF7L2 (var = 144,717 e 117,398, respectivamente), sendo máxima na população de casos no gene P53 (var = 144,717). Entre casos e controles, os valores foram maiores para casos para as sequências de FOXO1 (var = 15,767) e P53 (var = 144,717) e para controles, em PDX1 (76,976) e TCF7L2 (var = 117,398).

A estimativa da quantidade mínima de eventos de recombinação foi mais alta para a população caso de P53 (Rm = 9), seguida das populações caso e controle de TCF7L2 (Rm = 7, para ambas). Os valores mais baixos foram observados para as populações caso e controle de FOXO1, e para a população controle de PDX1 (Rm = 1, para todas).

4.4.6 Teste de neutralidade (estimativa de *Fu & Li*)

Em nenhuma das subpopulações os parâmetros de neutralidade estimados apresentaram p-valor correspondente à faixa de significância estatística necessária para corroborar com a hipótese de seleção neutra (valor de referência: p-valor < 0,05).

Tabela 4 – Estimativa de parâmetros de recombinação para os genes FOXO1, PDX1, P53 e TCF7L2 em populações de casos e controles para diabetes tipo 2.

Região	População	R-gene	R-sítio	var	Rm
FOXO1	Caso	4,5	0,0022	15,767	1
	Controle	20,9	0,0100	7,996	1
	Total	5,5	0,0026	13,624	1
PDX1	Caso	5,5	0,0069	50,752	2
	Controle	8,4	0,0106	76,976	1
	Total	5,3	0,0670	62,207	2
P53	Caso	13,5	0,0092	144,717	9
	Controle	42,7	0,0291	93,718	3
	Total	15,7	0,0107	130,231	10
TCF7L2	Caso	61,2	0,0313	55,195	7
	Controle	303	0,1552	117,398	7
	Total	70,1	0,0359	62,499	10

Legenda: Taxa de recombinação total (R-gene), Taxa de recombinação por sítio (R-sítio), Variância (var), Número mínimo de eventos de recombinação (Rm). Fonte: Elaborado pela autora. Dados obtidos do *software* DnaSP.

4.5 Identificação de polimorfismos de um nucleotídeo para estudos de associação

Os principais descritores da associação entre os SNPs identificados com a probabilidade de manifestação da doença estão exemplificados na **Tabela 5**.

A maior quantidade de SNPs foi observada para os genes P53 e TCF7L2 (112 e 145, respectivamente). Os demais genes, FOXO1 e PDX1, apresentaram, respectivamente, 19 e 41 SNPs em comparação às sequências CDS de referência.

Tabela 5 – Estimativa de p-valor, OR e IC de SNPs identificados na região FOXO1.

Região genômica	SNP*	p-valor	OR (IC 95%)
FOXO1	1	0,011	0,00 (0,00-NA**)
	2	0,440	NA (0,00-NA)
	3	0,440	NA (0,00-NA)
	4	0,260	NA (0,00-NA)
	5	0,260	NA (0,00-NA)
	6	0,370	3 (0,24-37,67)
	7	0,540	0,47 (0,04-5,90)
	8	0,540	0,47 (0,04-5,90)
	9	0,099	0,00 (0,00-NA)
	10	0,260	0,00 (0,00-NA)
	11	0,260	0,00 (0,00-NA)
	16	0,440	NA (0,00-NA)
	17	0,440	NA (0,00-NA)
	18	0,260	NA (0,00-NA)
	19	0,260	NA (0,00-NA)

Legenda: p-valor (significância), *odds-ratio* (OR), Intervalo de 95% de confiança (IC 95%). * Os SNPs 12, 13, 14, e 15 foram desconsiderados por apresentarem caráter monomórfico. ** NA (*not-available*).
Fonte: Elaborado pela autora. Dados obtidos da ferramenta SNPStats.

Destacam-se duas situações observadas a partir das análises de associação. Na primeira, exemplificada pelo SNP 1 (T/A) de FOXO1 (**Tabela 5**), apesar do p-valor estimado se inserir na faixa de significância estatística, não foi possível determinar o intervalo de confiança para a validação dos dados de associação. Na segunda, ilustrada pelo SNP 6 (C/A) de FOXO1 (**Tabela 5**), apesar do valor de OR (*odds-ratio*, razão de probabilidade) ser elevado (OR = 3; valor de referência de associação: OR > 1), o intervalo de confiança se encontra em faixa muito

ampla (IC 95% = 0,24 – 37,67). Além disso, o p-valor estimado (0,370; valor de referência: p-valor < 0,05) encontra-se fora da faixa de significância estatística.

A mesma discordância entre os valores de OR e p-valor foi observada para os demais genes. O SNP 53 (G/C) de TCF7L2, por exemplo, apresentou valor de OR de 1.2 (valor de referência de associação: OR > 1), porém tanto o p-valor observado (0,88; valor de referência: p-valor < 0,05) quanto o intervalo de confiança (0,12 – 11,87) estão fora da faixa de significância estatística.

5 DISCUSSÃO

Todos os mapeamentos realizados no *software* Geneious (**Figuras 1-4**) apresentaram, em geral, espectro de cobertura satisfatório. Foram observadas regiões de baixa cobertura mais expressivas principalmente nos mapeamentos realizados para o gene PDX1, para o qual houve maior restrição na quantidade de dados de *reads* disponíveis para as montagens de sequência.

Estudos com base em sequenciamento de exoma estão sujeitos à influência de variáveis operacionais que podem induzir erros de predição ou interpretação de variantes associadas a tais doenças, sendo as principais limitações os algoritmos de análise e processamento de dados e o potencial de replicação e validação das análises executadas. O espectro de cobertura de ferramentas de sequenciamento de exoma pode sofrer influência de características biológicas da região-alvo (conteúdo de G-C, por exemplo) ou limitações das técnicas utilizadas para montagem das sequências e identificação de variantes, o que explica a heterogeneidade observada para algumas regiões dos mapeamentos (ADAMS; ENG, 2018).

A observação de regiões de baixa confiabilidade ao longo das sequências geradas indica a necessidade de etapas adicionais de controle de qualidade dos *reads*, fazendo uso de *softwares* especializados, a fim de refinar os dados e aumentar a homogeneidade dos consensos. A otimização da qualidade dos *reads* utilizados para as análises poderá também minimizar os artefatos operacionais de mapeamento e

aumentar a confiabilidade dos polimorfismos identificados posteriormente ao longo das sequências.

Ainda assim, como observado nos alinhamentos múltiplos realizados com o *software* MEGA X (**Figuras 5 e 6**), as sequências consenso apresentaram perfil majoritariamente homogêneo entre os indivíduos. Os descritores de diversidade nucleotídica π e θ-$W_{sítio}$ (**Tabela 3**) reforçam a homogeneidade das sequências para todas as regiões de estudo, sendo FOXO1 o gene com o perfil mais uniforme observado para ambas as análises. No entanto, a baixa diversidade nucleotídica não permite explicar a alta quantidade de haplótipos distintos observada entre as subpopulações por efeito de mutações. O mesmo se aplica aos parâmetros de Fu & Li, que não permitem inferir o efeito de seleção natural atuando sobre a população.

Já os altos valores de R estimados (**Tabela 4**), especialmente para a região correspondente ao gene TCF7L2 (R-gene = 303), indicam que a alta diversidade haplotípica (hd), verificada para cada gene e entre as subpopulações, pode ser resultado de eventos frequentes de recombinação. A hipótese é sustentada pela identificação prévia de um *hotspot* de recombinação localizado no éxon 14 de TCF7L2, associado às altas taxas de recombinação observadas para o gene em uma população ameríndia (ACOSTA *et al.*, 2016). Análises posteriores utilizando softwares especializados na detecção de *hotspots* de recombinação devem ser realizadas para a confirmação desta hipótese.

No presente estudo, foram estimados parâmetros de associação (*odds-ratio* e p-valor, exemplificados na **Tabela 5**) entre os SNPs e os grupos-resposta diferentes dos previamente reportados na literatura. O estudo realizado por Müssig *et al.* (2009), por exemplo, observou a associação entre variantes genéticas e a ocorrência de diabetes tipo 2 em uma população de 5957 finlandeses. Um SNP presente no gene FOXO1 (rs2721068) foi reportado como variante de risco tanto para a doença (OR, *odds ratio* = 1,532; p-valor, significância = 2×10^{-3}) quanto para a manifestação de fenótipos pré-diabéticos, como deficiência na secreção de insulina pelas células beta-pancreáticas. Embora tenha-se estimado um alto valor de associação entre um dos SNPs de FOXO1 (SNP6) e a manifestação da doença (OR = 3, valor de referência de associação: OR > 1), a baixa confiabilidade estatística constatada para as análises (p-valor observado > 0,05) reforça a necessidade de

maior capacidade de processamento para refinamento e validação adequada dos dados (KIEZUN *et al.*, 2012).

Todos os genes escolhidos para o estudo (com exceção de P53) codificam fatores de transcrição relacionados ao metabolismo da glicose, função das células beta-pancreáticas ou síntese e secreção de insulina. Os SNPs previamente reportados para esses genes foram identificados predominantemente em regiões intrônicas (não codificantes) do genoma, o que não explica a totalidade dos mecanismos moleculares associados à patogênese da doença. Compreender o impacto de alterações funcionais em regiões do exoma, por outro lado, pode complementar a elucidação de tais mecanismos, uma vez que efeitos funcionais em regiões codificantes podem ser mais facilmente identificados por alterações diretas na função dos produtos proteicos correspondentes (KIEZUN *et al.*, 2012).

As análises realizadas sugerem ser possível rastrear variantes genéticas em regiões do exoma que possam ser associadas ao risco de manifestação de diabetes tipo 2 em populações de humanos, especialmente para os genes FOXO1 e TCF7L2. Novas análises, utilizando dados refinados de *reads* de um número maior de indivíduos nas subpopulações, devem ser realizadas a fim de possibilitar a avaliação do impacto funcional de tais variantes para a ocorrência da doença.

.

REFERÊNCIAS

A FAKHRO, Khalid *et al*. The Qatar genome: a population-specific tool for precision medicine in the Middle East. **Human Genome Variation**, v. 3, n. 1, p.1-7, jun. 2016.

ACOSTA, Jose Luis *et al*. Rare intronic variants of TCF7L2 arising by selective sweeps in an indigenous population from Mexico. **Bmc Genetics**, v. 17, n. 1, p.1-15, maio 2016.

ADAMS, David R.; ENG, Christine M.. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. **New England Journal Of Medicine**, v. 379, n. 14, p.1353-1362, out. 2018.

AL-QUOBAILI, Faizeh; MONTENARH, Mathias. Review: Pancreatic duodenal homeobox factor-1 and diabetes mellitus type 2. **International Journal Of Molecular Medicine**, p.399-404, abr. 2008.

AMERICAN DIABETES ASSOCIATION. Diagnosis and Classification of Diabetes Mellitus. **Diabetes Care**, v. 37, Supplement 1, p.81-90, jan. 2014.

BARRA, Gustavo Barcelos *et al*. Association of the rs7903146 single nucleotide polymorphism at the Transcription Factor 7-like 2 (TCF7L2) locus with type 2 diabetes in Brazilian subjects. **Arquivos Brasileiros de Endocrinologia & Metabologia**, v. 56, n. 8, p.479-484, nov. 2012.

BROWN, Audrey E.; WALKER, Mark. Genetics of Insulin Resistance and the Metabolic Syndrome. **Current Cardiology Reports**, v. 18, n. 8, p.1-8, jun. 2016.

HARGREAVES, Adam D. *et al*. Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. **Proceedings Of The National Academy Of Sciences**, v. 114, n. 29, p.7677-7682, jul. 2017.

ILLUMINA INC. **Illumina Sequencing Technology:** Highest data accuracy, simple workflow, and a broad range of applications.. 2010. Disponível em: <https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf>. Acesso em: 22 out. 2018.

KANETO, Hideaki *et al*. Role of Reactive Oxygen Species in the Progression of Type 2 Diabetes and Atherosclerosis. **Mediators Of Inflammation**, v. 2010, p.1-11, 2010.

KATOH, Masuko *et al*. Cancer genetics and genomics of human FOX family genes. Cancer Letters, v. 328, n. 2, p.198-206, jan. 2013.

KHAN, Sadiya S.; SINGER, Benjamin D.; VAUGHAN, Douglas E.. Molecular and physiological manifestations and measurement of aging in humans. **Aging Cell**, v. 16, n. 4, p.624-633, maio 2017.

KIEZUN, Adam *et al*. Exome sequencing and the genetic basis of complex traits. **Nature Genetics**, v. 44, n. 6, p.623-630, jun. 2012.

KITAMURA, Tadahiro. The role of FOXO1 in β-cell failure and type 2 diabetes mellitus. **Nature Reviews Endocrinology**, v. 9, n. 10, p.615-623, ago. 2013.

KUNG, Che-pei; MURPHY, Maureen. The role of the p53 tumor suppressor in metabolism and diabetes. **Journal Of Endocrinology**, v. 231, n. 2, p.61-75, set. 2016.

LYSSENKO, Valeriya *et al*. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. **Journal Of Clinical Investigation**, v. 117, n. 8, p.2155-2163, ago. 2007.

MARTINS, Rute; LITHGOW, Gordon J.; LINK, Wolfgang. Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. **Aging Cell**, v. 15, n. 2, p.196-207, dez. 2015.

MULLER, Yunhua L. *et al*. Assessing FOXO1A as a potential susceptibility locus for type 2 diabetes and obesity in American Indians. **Obesity**, v. 23, n. 10, p.1960-1965, set. 2015.

MÜSSIG, Karsten *et al*. Association of Common Genetic Variation in the FOXO1 Gene with β-Cell Dysfunction, Impaired Glucose Tolerance, and Type 2 Diabetes. **The Journal Of Clinical Endocrinology & Metabolism**, v. 94, n. 4, p.1353-1360, abr. 2009.

PAGE, Andrew J. *et al*. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. **Microbial Genomics**, v. 2, n. 4, p.1-5, abr. 2016.

PATNALA, Radhika; CLEMENTS, Judith; BATRA, Jyotsna. Candidate gene association studies: a comprehensive guide to useful in silico tools. **Bmc Genetics**, v. 14, n. 1, p.1-11, 2013.

PEVZNER, Pavel; SHAMIR, Ron. **Bioinformatics for Biologists**. Nova York. Cambridge University Press, 2011. 394 p.

PONUGOTI, Bhaskar; DONG, Guangyu; GRAVES, Dana T.. Role of Forkhead Transcription Factors in Diabetes-Induced Oxidative Stress. **Experimental Diabetes Research**, v. 2012, p.1-7, set. 2012.

PRASAD, Rashmi; GROOP, Leif. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. **Genes**, v. 6, n. 1, p.87-123, mar. 2015.

PRUZIN, J. J. *et al*. Review: Relationship of type 2 diabetes to human brain pathology. **Neuropathology And Applied Neurobiology**, v. 44, n. 4, p.347-362, mar. 2018.

RÓNAI, Zsolt *et a*l. Investigation of the genetic background of complex diseases. **Orvosi Hetilap**, v. 159, n. 31, p. 1254-1261, 2018.

SLIWINSKA, Agnieszka *et al*. Tumour protein 53 is linked with type 2 diabetes mellitus. **Indian Journal Of Medical Research**, v. 146, n. 2, p.237-243, ago. 2017.

SOLÉ, Xavier *et al*. SNPStats: a web tool for the analysis of association studies. **Bioinformatics**, v. 22, n. 15, p.1928-1929, maio 2006.

STEINTHORSDOTTIR, Valgerdur *et al*. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. **Nature Genetics**, v. 46, n. 3, p.294-298, jan. 2014.

SUDCHADA, P.; SCARPACE, K.. Transcription factor 7-like 2 polymorphisms and diabetic retinopathy: a systematic review. **Genetics And Molecular Research**, v. 13, n. 3, p.5865-5872, ago. 2014.

SZUMILAS, Magdalena. Explaining Odds Ratios. **Journal of the Canadian Academy of Child and Adolescent Psychiatry**, v. 19, n. 3, p.227-229, ago. 2010.

WU, Yanling *et al*. Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. International Journal Of Medical Sciences, v. 11, n. 11, p.1185-1200, set. 2014.

ZHOU, Yuedan *et al*. TCF7L2 is a master regulator of insulin production and processing. **Human Molecular Genetics**, v. 23, n. 24, p.6419-6431, jul. 2014.

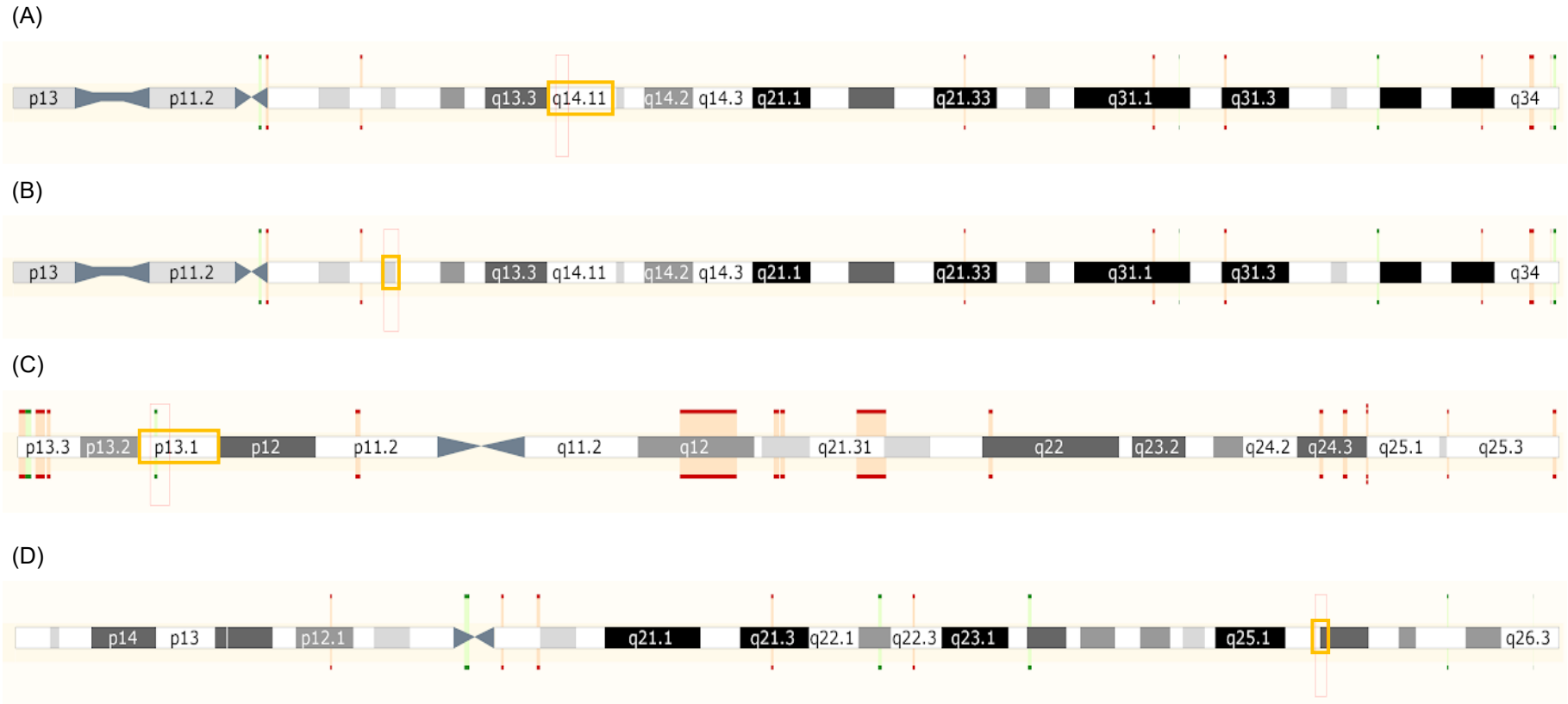
APÊNDICE A – INFORMAÇÕES COMPLEMENTARES SOBRE O GENE ESCOLHIDO COMO CONTROLE NEGATIVO (P53)

P53 (*tumor protein* p53)

A proteína p53 (*tumor protein* p53) é um fator de transcrição de resposta a condições de estresse que regula a expressão de genes-alvo envolvidos no controle do ciclo celular, metabolismo, reparo do DNA e apoptose (SLIWINSKA *et al.*, 2017). O papel da proteína na indução da expressão de tais genes pode ser modulado por múltiplos fatores estressores, tais como danos ao DNA, hipóxia, restrição nutricional e erros de replicação (KUNG; MURPHY, 2016; SLIWINSKA *et al.*, 2017).

Sua atividade afeta funções fisiológicas como desenvolvimento, replicação e senescência celular, o que implica que mutações no gene de p53 ou outros genes que interfiram na atividade da proteína podem estar associadas à ocorrência de doenças degenerativas, principalmente processos carcinogênicos (KUNG; MURPHY, 2016).

ANEXO A – LOCALIZAÇÃO DOS GENES UTILIZADOS NO ESTUDO EM CROMOSSOMOS DE HUMANOS (*Homo sapiens*)



Figuras A-D. Localização dos genes (controle e teste) utilizados para o estudo em cromossomos de humanos (*Homo sapiens*). A – FOXO1, 13q14.11; B – PDX1, 13q12.2; C – P53, 17p13.1; D – TCF7L2, 10q25.2-25.3. As regiões que contêm os genes estão destacadas em quadros amarelos. A nomenclatura dos *loci* gênicos segue o padrão: número do cromossomo + braço do cromossomo onde se encontra o gene + posição no braço, em ordem crescente a partir do centrômero. Fonte: Ensembl.

ANEXO B – REFERÊNCIA DE NOMENCLATURA DE BASES NUCLEOTÍDICAS

Anexo B – Referência de nomenclatura de bases nucleotídicas

Código	Base
A	Adenina
C	Citosina
G	Guanina
T (ou U)	Timina (ou Uracila)
R	A ou G
Y	C ou T
S	G ou C
W	A ou T
K	G ou T
M	A ou C
B	C ou G ou T
D	A ou G ou T
H	A ou C ou T
V	A ou C ou G
N	Qualquer base
-	Gap (espaçamento)

Fonte: Kyoto Encyclopedia of Genes and Genomes (KEGG).

ANEXO C – FORMATOS DE ENTRADA E SAÍDA DAS PLATAFORMAS DE ANÁLISE UTILIZADAS PARA O ESTUDO

Anexo C – Formatos de entrada e saída das plataformas de análise utilizadas para o estudo.

Ferramenta	Formato de entrada	Formato de saída
Illumina	mRNA	FASTQ
Geneious	FASTQ	FASTA
ORF finder	FASTA	Protein FASTA
MEGA X	FASTA	MEGA ou FASTA
DnaSP	FASTA	Descritores
SNP-sites	FASTA	FASTA
SNPStats	Tabela de variantes	p-valor, *odds ratio*, IC

Fonte: Elaborado pela autora.