INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS – *CAMPUS* SÃO JOÃO EVANGELISTA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Luan Patrik Silva Pinto

ANÁLISE COMPARATIVA DAS SEQUÊNCIAS DE REFERÊNCIA DOS GENES IGHV, IGHD e IGHJ DE HUMANOS SEQUENCIADOS COM A TECNOLOGIA *SHORT READS*

São João Evangelista
2023

LUAN PATRIK SILVA PINTO

ANÁLISE COMPARATIVA DAS SEQUÊNCIAS DE REFERÊNCIA DOS GENES IGHV, IGHD e IGHJ DE HUMANOS SEQUENCIADOS COM A TECNOLOGIA *SHORT READS*

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Sistemas de Informação do Instituto Federal de Minas Gerais – *Campu*s São João Evangelista para obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fábio Rodrigues Martins

São João Evangelista

2023

P659a Pinto, Luan Patrik Silva.

Análise comparativa das sequências de referência dos genes IGHV, IGHD e IGHJ de humanos sequenciados com a tecnologia short reads / Luan Patrik Silva Pinto – 2023.

43f.: il.

Orientador: Prof. Dr. Fábio Rodrigues Martins.

Trabalho de Conclusão de Curso (bacharelado em Sistemas de informação) – Instituto Federal Minas Gerais. *Campus* São João Evangelista, 2023.

1. Imunoglobulina. 2. Genoma de referência. 3. Genes. 4. Sequenciamento de nova geração. I. Pinto, Luan Patrik Silva. II. Instituto Federal de Minas Gerais *Campus* SJE. III. Título.

CDD 576.05

Catalogação: Esther Soares Cunha - CRB-6/MG-003372/P

LUAN PATRIK SILVA PINTO

ANÁLISE COMPARATIVA DAS SEQUÊNCIAS DE REFERÊNCIA DOS GENES IGHV, IGHD e IGHJ DE HUMANOS SEQUENCIADOS COM A TECNOLOGIA *SHORT READS*

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Sistemas de Informação do Instituto Federal de Minas Gerais – *Campu*s São João Evangelista para obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fábio Rodrigues Martins

Aprovado em 21/11/2023 pela banca examinadora:

gov.br Documento assinado digitalmente
FABIO RODRIGUES MARTINS
Data: 23/02/2024 10:36:21-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fábio Rodrigues Martins (Orientador) – IFMG - SJE

Dra. Carlena Tahina Navas de Reyes - UFMG

gov.br Documento assinado digitalmente
LUCIANA WERNECK ZUCCHERATO
Data: 22/02/2024 13:47:34-0300
Verifique em https://validar.iti.gov.br

Dra. Luciana Werneck Zuccherato – UFMG

Dedico este trabalho a todas as pessoas que sonham em alcançar grandes objetivos. À minha mãe Andréia, meus irmãos Luana, Carol e Theo aos quais eu prometo meu amor incondicional e eterno. Aos estudantes e professores do IFMG – Campus São João Evangelista por terem acreditado no meu potencial.

AGRADECIMENTOS

Agradeço à Deus pela dádiva da vida e por ter me abençoado em toda a minha vida e trajetória acadêmica.

Aos meus pais, Andréia e Antônio, por todo o amor, apoio emocional e financeiro. Meus irmãos Luana, Carol e Theo por serem o meu porto seguro. À toda a minha família por acreditarem em mim. Amo vocês mil milhões <3.

Ao meu orientador, Dr. Fábio, pelo auxílio na construção desse trabalho e por sempre acreditar que eu conseguiria entregar um ótimo resultado. Obrigado também por ter sido um amigo e saiba que você sempre será o meu Doutor favorito.

Aos professores e funcionários do IFMG - *Campus* São João Evangelista, em especial Eduardo Trindade e Ítalo Magno, com vocês eu aprendi muito além do que ser um profissional. Agradeço todos os ensinamentos e levarei todos em meu coração.

Ao meu amigo Mateus Pereira que considero como um irmão e sem você essa jornada teria sido muito mais árdua. Á Giovanna Kruk por ter acreditado no meu potencial até mesmo quando eu já não acreditava mais.

Aos meus colegas de turma, em especial Arthur, Daniel, José Guilherme e Karine e Mateus Henrique, pelo apoio em toda essa difícil jornada, pela ajuda nas disciplinas e serem o apoio que precisei.

À galera do Hotel Barbosa, Aguinaldo, Evandro, Luciene, Magno, Marcos Paulo, Thiago e Weslley por terem sido como uma grande família, nunca esquecerei todos os nossos momentos juntos, principalmente as reuniões de condomínio na cozinha.

Ao grupinho mais apocalíptico que eu já tive o prazer de conhecer formado pelos integrantes Armando, Dalisson, Enzo, Índia, Jainy, Junia, Mário, Mirelly, Samyra e Thata, agradeço pela amizade e por todos os momentos de felicidade nos bares, as danças, a cantoria, pelo apoio e acreditarem no meu potencial.

Ao bar Altas Horas, um beijo Diana e Anisio, por terem tocado Marília Mendonça alegrando as minhas noites, jamais esquecerei esses momentos. Obrigado pelo carinho e amizade.

À eterna rainha da sofrência Marília Mendonça por ter me acompanhado em toda essa jornada através de um fone de ouvido. Suas canções estavam comigo em momentos felizes e tristes. Por causa de você, as pessoas que me conhecem, assim que te ouvem, lembram de mim.

E por fim, agradeço a mim mesmo. A jornada até aqui não foi fácil, a vontade de desistir passou em minha mente muitas vezes, mas eu fui resiliente e persistente.

“Eu sei que tem pessoas que dizem que essas coisas não acontecem, e que isso serão apenas histórias um dia. Mas agora nós estamos vivos. E nesse momento, eu juro. **Nós somos infinitos**.”

Stephen Chbosky

RESUMO

Com o aumento na quantidade de dados genômicos resultantes do sequenciamento do DNA humano pelo Projeto Genoma Humano e das outras espécies sequenciadas até os dias atuais, surgiu o desafio de como armazenar, tratar e manipular esses dados em busca de extrair conhecimento. A fim de estudar as imunoglobulinas que são moléculas essências para a resposta imune, neste trabalho analisamos os genes de cadeia pesada V, D e J localizado no cromossomo 14. Sendo assim, este trabalho utiliza dados de genomas sequenciados com a tecnologia de sequenciamento de nova geração *short reads* para realizar uma análise comparativa das sequências de referência dos genes IGHV, IGHD e IGHJ de humanos, a fim de verificar se existem diferenças ou/e similaridades entre os genomas de referência versões CRCh37 e CRCh38. Após realizar as análises, foi possível concluir que existem diferenças na quantidade de nucleotídeos e presença de genes, além de similaridades entre os dois tipos de genomas de referência.

Palavras-chave: Imunoglobulina. Genoma de Referência. Genes. Sequenciamento de Nova Geração

ABSTRACT

With the increase in the amount of genomic data resulting from the sequencing of human DNA by the Human Genome Project and other species sequenced to date, the challenge of how to store, process, and manipulate this data in order to extract knowledge has emerged. In order to study immunoglobulins, which are essential molecules for the immune response, this work analyzes the V, D, and J heavy chain genes located on chromosome 14. Therefore, this work uses data from genomes sequenced with the short reads next-generation sequencing technology to perform a comparative analysis of the reference sequences of the IGHV, IGHD, and IGHJ genes of humans, in order to verify if there are differences or similarities between the reference genomes versions CRCh37 and CRCh38. After carrying out the analyses, it was possible to conclude that there are differences in the number of nucleotides and the presence of genes, as well as similarities between the two types of reference genomes.

Keywords: Immunoglobulin. Reference genome. Genes. Next-Generation Sequencing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DDBJ - *DataBank of Japan*

DNA - Ácido Desoxirribonucleico

ENA - *European Nucleotide Archive*

GRCh37 - *Genome Reference Consortium Human Build 37*

GRCh38 - *Genome Reference Consortium Human Build 38*

Ig - Imunoglobulinas

IGH - *Immunoglobulin heavy locus*

IGHC - *Immunoglobulin heavy locus constant*

IGHD - *Immunoglobulin heavy locus diversity*

IGHJ - *Immunoglobulin heavy locus joining*

IGHV - *Immunoglobulin heavy locus variable*

IGK - *Immunoglobulin kappa locus*

IGL - *Immunoglobulin lambda locus*

IMGT - *International ImMunoGeneTics information system*

Kb - *Kilobase*

KDD - *Knowledge-Discovery In Databases*

NCBI - *National Center for Biotechnology Information*

NGS - Sequenciamento de Nova Geração

NHGRI - *National Human Genome Research Institute*

PGH - Projeto Genoma Humano

SUMÁRIO

1 INTRODUÇÃO

Em 1990, iniciava-se o Projeto Genoma Humano (PGH) como objetivo de sequenciar o DNA humano em busca de identificar e mapear os genes presentes nos 23 pares de cromossomos. Em abril de 2003, as atividades do Projeto foram encerradas com a conclusão da primeira versão do sequenciamento do genoma humano. Em setembro de 2011 já haviam disponíveis 37 espécies eucarióticas totalmente sequenciadas e estavam em andamento o sequenciamento de outras 1.178 espécies. Esses fatos impactaram diretamente no estudo de genomas, já que haviam poucos genomas sequenciados. (ZAHA, FERREIRA e PASSAGLIA, 2014).

Com o aumento na quantidade de dados genômicos sequenciados advindos dos avanços tecnológicos dos equipamentos em conjunto com as técnicas de sequenciamento de genoma, surge o desafio de como armazenar, tratar e manipular os dados para extrair conhecimento ou identificar padrões. Para esse fim, tornou-se indispensável o desenvolvimento de ferramentas na área de bioinformática (PEVSNER, 2015).

Segundo Lesk (2008), a bioinformática surge na década de 1980, com a finalidade de superar os limites impostos pelas ciências com o desenvolvimento de novas abordagens capazes de proporcionar a análise e a apresentação de dados biológicos, assim como de iniciar estudos de novos métodos para a resolução de questões complexas.

Em paralelo a área de bioinformática, existe a Imunoinformática que atua especificamente com dados de imunoglobulinas ou anticorpos. Estas moléculas desempenham papéis centrais durante a resposta imune que são: reconhecer antígenos (corpo estranho), desencadear o mecanismo de resposta imediata ou mecanismo de resposta adaptativa que seria a produção de anticorpos para combater o corpo estranho (LEFRANC; LEFRANC, 2001). A molécula de imunoglobulina é codificada por três loci gênicos independentes, ou seja, os genes IGK e IGL para a cadeia leve (*light*, L) localizados no cromossomo 2, cromossomo 22 e os genes IGH para a cadeia pesada (*heavy*, H), que estão no cromossomo 14 (MATSUDA, 1998).

O cromossomo 14 é um dos 23 pares de cromossomos presentes no genoma humano, sendo responsável pela codificação de diversos genes na cadeia pesada dos anticorpos que estão relacionados à saúde e ao desenvolvimento do organismo. O lócus da cadeia pesada da imunoglobulina (IGH, *Immunoglobulin heavy locus*) de humanos se encontra na banda 14q32.33 dentro da região telomérica do cromossomo 14, abrangendo 1.250 kilobases (kb) (LEFRANC; LEFRANC, 2001).

Com esse grande volume de dados a serem analisados, surge a necessidade da criação de novas ferramentas para trabalhar com dados advindos de sequenciamento genômico. Ainda neste contexto, vale ressaltar o Sequenciamento de Nova Geração, do inglês *New Generation Sequencing* (NGS), que descreve as abordagens que possuem a capacidade de sequenciar ácidos nucleicos aumentando a quantidade de sequência que podem ser obtidas de maneira rápida e econômica. O NGS segue três etapas básicas: preparação de bibliotecas, sequenciamento de ácidos nucleicos e análise de dados (KUMAR et al., 2019). É importante observar que essas técnicas além de agilizar o sequenciamento genômico proporcionaram uma economia considerável, e segundo o *National Human Genome Research Institute* (NHGRI), em 2022 seria possível sequenciar um genoma humano por 100 dólares.

Dada toda essa problemática apresentada acima, minerar esses dados é um desafio para a área da computação e ao mesmo tempo é necessário, para conhecermos melhor o genoma humano. Como afirma Griffiths (2001), tornou-se importante analisar e entender os assuntos que se referem ao DNA humano.

1.1 Justificativa

As imunoglobulinas são moléculas essenciais na resposta a um patógeno, neste contexto estudos e pesquisas que buscam melhorar entendimento do *lócus* que codifica estas moléculas é de extrema relevância, já que, pouco se sabe sobre o mesmo. É importante observar que para desvendarmos as sequências geradas pelo sequenciamento do DNA, é necessário que se tenha um genoma de referência de qualidade.

Sendo assim, este trabalho justifica-se pela necessidade em dar continuidade às pesquisas que estudam genomas de referência de humanos. Além disso, todo e qualquer estudo sobre este lócus codificante é de extrema relevância, uma vez que, os genes desse lócus compõem os anticorpos.

1.2 Objetivos

O objetivo geral deste trabalho consiste na criação de um *pipeline* computacional para filtrar e realizar a comparação das sequências dos genomas de referência versões GRCh37 e GRCh38 sequenciados com a tecnologia de *short reads.*

Para atingir o objetivo principal proposto, foram definidos os seguintes objetivos específicos:

a) Configurar o servidor e realizar o *download* dos genomas de referência GRCh37 e GRCh38;
b) Configurar e executar o programa disponibilizado por (MARTINS et al., 2021) no genoma de referência GRCh7 e GRCh38;
c) Desenvolver uma função para organização dos dados de saída;
d) Fazer uma análise dos genes de cada genoma de referência sequenciado com a tecnologia *short reads* e gerar uma visualização que permita identificar as diferenças nas sequencias de cada gene;
e) Fazer uma análise posicional dos genes indicando: alterações na quantidade de nucleotídeos (tamanho) e mudanças nas posições, disponibilizar uma visualização que permita identificar essas situações;
f) Fazer uma análise geral dos dados e criar diferentes visualizações.

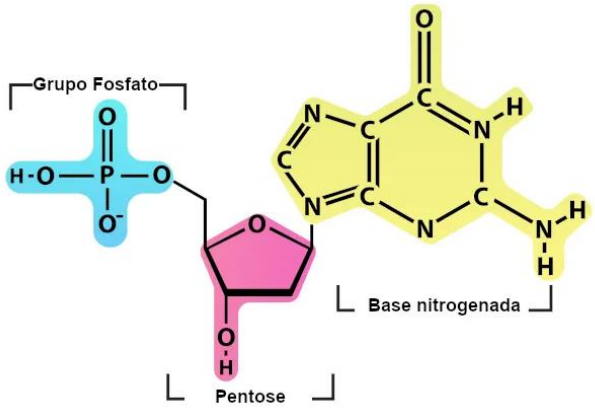
2 REFERENCIAL TEÓRICO

Nesta seção serão apresentados os conceitos e trabalhos correlacionais que embasam a construção deste trabalho.

2.1 Genoma

Segundo Góes e Oliveira (2014), o genoma é definido como uma sequência de ácido desoxirribonucleico (DNA) de um organismo, sendo o DNA formado pela ligação sequencial de moléculas chamadas nucleotídeos. Estes, representados na Figura 1, são formados por três partes: uma molécula de fosfato, a molécula de açúcar desoxirribose e a base nitrogenada. As bases nitrogenadas podem ser divididas em quatro tipos: adenina (A), timina (T), citosina (C) e guanina (G).

Figura 1 - Composição de um Nucleotídeo



Fonte: (SANTOS, 2023)

A ordem com a qual os nucleotídeos estão organizados no DNA é que faz com que uma molécula se diferencie da outra. Esta diferença entre sequências pode ser determinada por meio do sequenciamento dos genomas. Como as moléculas de fosfato e açúcar são sempre as mesmas, a ordem da sequência é dada pelas bases nitrogenadas (Góes e Oliveira, 2014). Além disso, é importante ressaltar que o DNA genômico possui a maior parte da informação genética de um indivíduo, sendo um alvo no processo de sequenciamento.

2.1.1 Gene

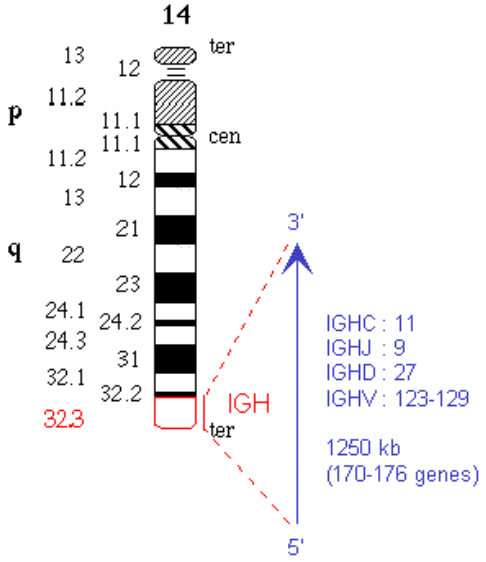
Um gene pode ser definido como uma sequência ordenada de nucleotídeos localizada em uma posição particular (*lócus*) em um cromossomo, que codifica um produto funcional específico (SANTOS, 2023). Em síntese, o gene pode ser descrito como um subgrupo extraído da sequência do genoma.

É importante observar que o gene possui duas partes principais, sendo o éxon e o íntron, estes dois segmentos se diferem pelo fato que na produção da proteína o íntron é eliminado pelo processo chamado de *splicing* alternativo, ou seja, no produto final teremos a presença dos éxons que são as partes que serão traduzidas em proteínas (AZEVEDO, 2021).

Ressalta-se, a relevância das informações provenientes dos genes, uma vez que podem ser utilizadas para o desenvolvimento de novos medicamentos, tratamentos para doenças adquiridas recentemente ou até mesmo para a prevenção daquelas que não se manifestaram.

Neste trabalho, serão investigados os segmentos de genes de imunoglobulinas que estão localizados no cromossomo 14 de humanos, na posição 14q32.33, onde 14 é a representação do cromossomo e 32.33 é a localização da banda cromossômica dos genes de IGH. A Figura 2 ilustra a localização do *loci* IGH que é alvo deste estudo. A seta azul indica a orientação de 5' até 3' do lócus e a ordem do grupo de genes no lócus, além de ser proporcional ao tamanho do lócus, indicado em kb. O número total de genes no lócus é mostrado entre parênteses.

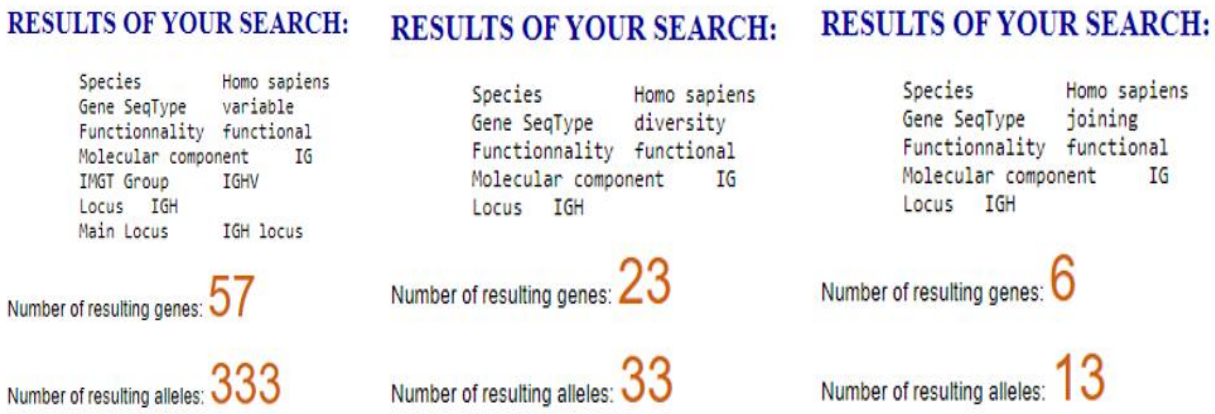
Figura 2 - Localização do loci IGH



Fonte: (GINESTOUX; LEFRANC, 2001)

No cromossomo 14 de humanos são catalogados de 123 a 129 IGHV dependendo dos haplótipos, 127 IGHD, 9 IGHJ, e frequentemente 11 IGHC. Em uma análise, 88 genes IGHV foram agrupados em apenas um subgrupo, já 41 pseudogenes apresentam divergências e não foram atribuídos a nenhum subgrupo (LEFRANC; LEFRANC, 2020). Porém, com o intuito de verificar a quantidade de genes IGHV funcionais, foi realizada em 09 de maio de 2023 uma busca no banco de dados do *International ImMunoGeneTics information system* (IMGT) (https://www.imgt.org/), onde foram filtrados por: espécie (*Homo Sapiens*), tipo de gene (*variable*, V) e funcionalidade. Como resultado dos genes catalogados no banco de dados IMGT/GENE-DB versão: 3.1.38 de 08 de novembro de 2022, foram retornados 57 genes IGVH funcionais e 333 alelos. Em suma, é de extrema importância, conhecer o número de genes IGHV já que esse será o objeto de estudo deste trabalho. Em 08 de novembro de 2023 foi realizada uma busca no site do IMGT versão 3.1.40 de 02 de outubro de 2023 para verificar a quantidade de genes IGHD e IGHJ, e foram identificados 23 genes D e 33 alelos, 6 genes J e 13 alelos. Esses resultados estão sendo ilustrados na Figura 3.

Figura 3 - Genes catalogados no banco de dados IMGT/GENE-DB

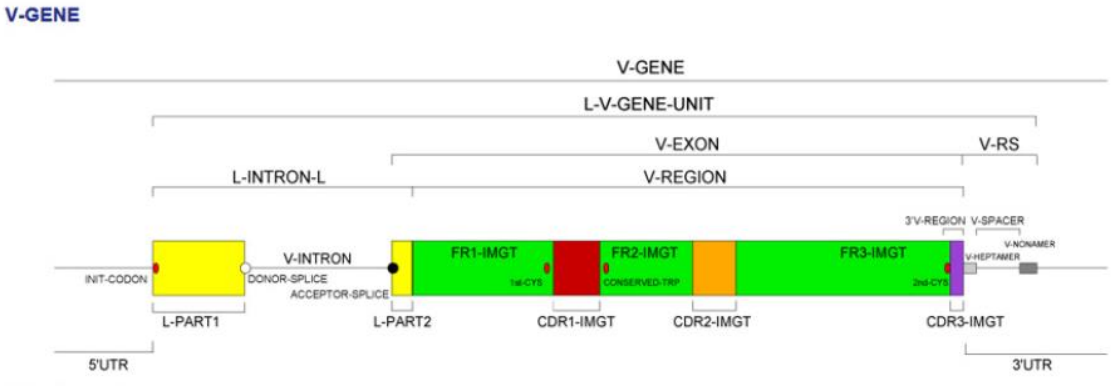


Fonte: (GIUDICELLI; CHAUME; LEFRANC, 2023)

As nomenclaturas utilizadas nos genes V, D e J seguem o padrão do IMGT. Como exemplo o geno IGHV é importante ressaltar que na Figura 4 o gene é apresentado com subdivisões, como exemplo L-V-GENE-UNIT, as partes V-EXONS, V-REGION e intrônicas. A estrutura do gene V é formada pela região líder (L-PART1), que possui também uma região v-íntron como mostrada na imagem abaixo pela Figura 4, abrangendo as regiões de íntron e posteriormente a região líder (L-PART2) como mostrada na Figura 4 em amarelo. A parte em verde, regiões de *frameworks*, são três: FR1-IMGT, FR2-IMGT e FR3-IMGT. Também é

possível identificar as regiões de determinação de complementariedade, do inglês *Complementarity-Determining Regions* (CDRs), tais regiões são apresentadas na Figura 4 sendo a parte vermelha o CDR1, em amarelo escuro o CDR2 e podemos identificar, em roxo, uma parte do CDR3.

Figura 4 - Estrutura do gene IGHV



Fonte: (LEFRANC; LEFRANC, 2020)

2.1.2 *Sequenciamento de Genoma*

De acordo com Fieto e Maciel (2015), o sequenciamento do genoma é uma técnica que obtém a ordem das bases nitrogenadas no DNA com alta confiabilidade. No final da década de 70 foram desenvolvidas as primeiras técnicas de sequenciamento que possuíam uma escala de sequenciamento de poucos kbs, na atualidade com a evolução tecnológica e a necessidade de realizar sequenciamentos de genomas inteiros e cada vez mais rápidos, surgiu o NGS.

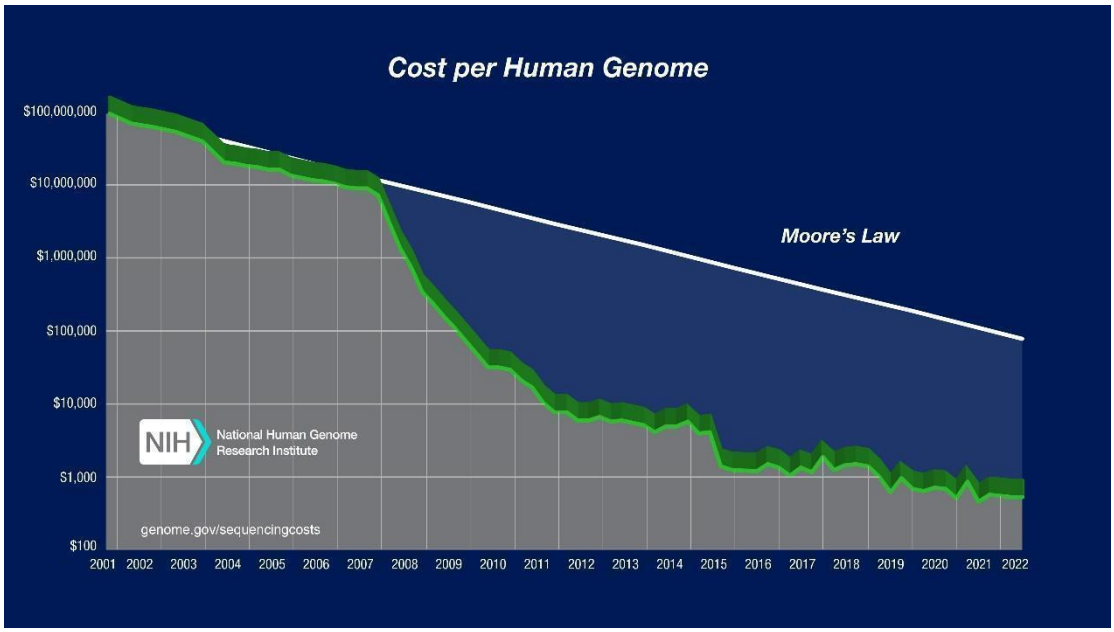
A plataforma NGS tem a característica de ser de alto rendimento e de sequenciar uma grande quantidade de moléculas diferentes de DNA (alta cobertura e alta profundidade). Capaz de realizar diversas interações em paralelo, essa nova abordagem proporcionou uma revolução no sequenciamento de DNA padronizado a ponto de um genoma humano inteiro agora pode ser sequenciado em 3 dias (KUMAR et al., 2019).

Neste trabalho, abordaremos apenas a tecnologia de sequenciamento *short reads*, capaz de ler sequências curtas. A empresa líder de mercado nesse segmento é a Illumina, com equipamentos e insumos de sequenciamento. Essa tecnologia foi bastante utilizada em estudos na área de imunologia (ROUET et al., 2018).

O Instituto Nacional de Pesquisa do Genoma Humano, do inglês *National Human Genome Research Institute*, (NHGRI) apresenta um panorama dos custos associados ao

sequenciamento de DNA com o intuito de verificar a oscilação destes valores ao longo do tempo. Esta informação pode ser observada na Figura 5, onde estão sendo apresentados os custos do ano de 2001 que corresponde a aproximadamente cem milhões de dólares para sequenciar um genoma humano até o ano de 2022, apresentando o custo de mil dólares por sequenciamento.

Figura 5 - Custo de Sequenciamento por genoma Humano



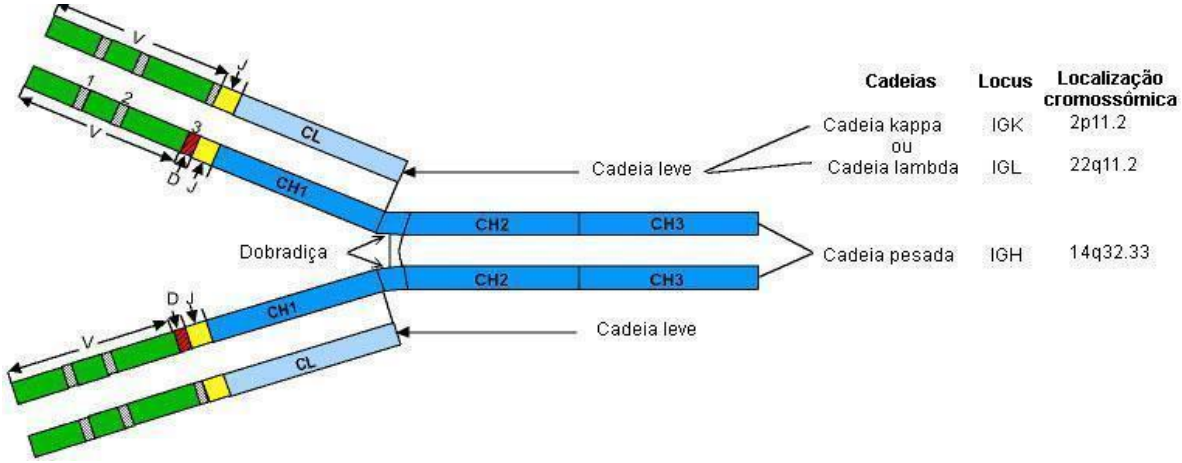
Fonte: (WETTERSTRAND, 2022)

2.2 Anticorpo

Os anticorpos ou imunoglobulinas (Ig) são proteínas responsáveis pela resposta imune humoral, que defende o corpo contra infecções e moléculas antigênicas. A molécula de anticorpo possui duas funções: 1) reconhecer e neutralizar patógenos ou a seus produtos que induziram a resposta imune; 2) recrutar outras células e moléculas, a fim de destruir o patógeno, quando o anticorpo estiver unido a ele (MURPHY, 2014).

O formato de uma molécula de anticorpo assemelha-se a um “Y” (Figura 5), e possui três segmentos de mesmo tamanho, ligados por uma porção flexível. O anticorpo humano é formado por quatro cadeias, sendo duas pesadas cujo lócus está presente no cromossomo humano 14, e duas cadeias leves, cujos loci estão presentes nos cromossomos 2 (lócus IGK kappa) e 22 (lócus IGL lambda) (LEFRANC; LEFRANC, 2020).

Figura 6 - Estrutura de um anticorpo



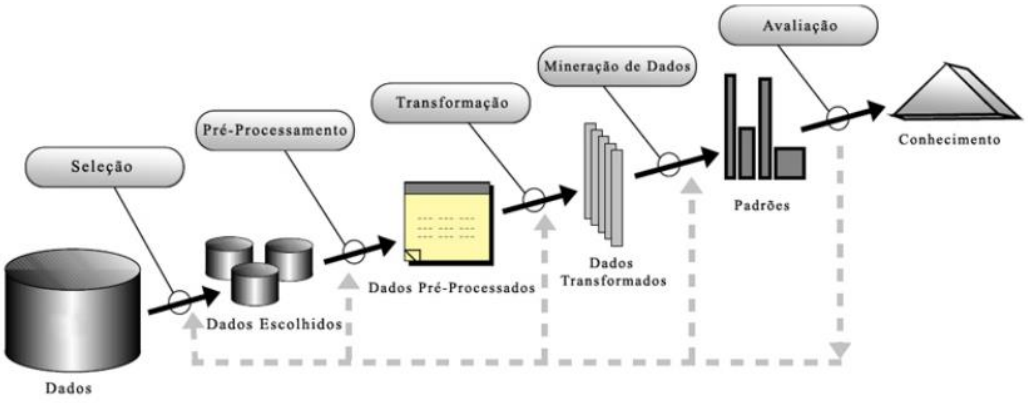
Fonte: (LEFRANC; LEFRANC, 2020)

2.3 Mineração de Dados

Segundo Hand et al. (2001), mineração de dados é definida como a análise de um grande volume de dados a fim de encontrar padrões, tendências e oportunidades que possam entregar algum tipo de valor agregado para o dono dos dados. Vale enfatizar que a mineração de dados pode ser entendida como a união de várias técnicas ou processos utilizando algoritmos preditivos ou descritivos que buscam extrair de um base de dados informações que geram algum tipo de conhecimento.

A descoberta de conhecimento nas Bases de Dados ou *Knowledge Discovery in Databases* (KDD) descreve todo o processo gerador de conhecimento, e nele a mineração de dados é descrita como uma das etapas (FAYYAD, 1996). Dado a complexidade de transformação de dados em conhecimento, tornou-se necessário a divisão em fases para se atentar aos detalhes e assim conseguir o melhor resultado possível. Na Figura 7, demonstramos as etapas do KDD.

Figura 7 - Etapas do Processo de KDD



Fonte: (FAYYAD, 1996)

Han e Kamber (2011) definem a Seleção como a primeira etapa do processo e nela os dados pertinentes são selecionados para serem usados na análise. A seguir, acontece o Processamento e a Transformação que visam manter a qualidade dos dados. Nessas duas etapas estão englobados a limpeza de dados inconsistentes e a utilização de técnicas que deixaram os dados em um formato adequado.

Em seguida, temos a Mineração de Dados parte na qual serão utilizados técnicas e algoritmos em busca de padrões significativos nos dados. Enfim, são feitos a Avaliação e Interpretação das informações geradas à procura de algum conhecimento que possa ser adquirido.

2.4 Python

De acordo com Borges (2014), Python é uma linguagem de programação com sintaxe clara e simplificada de alto nível, orientada a objetos, de tipagem dinâmica e forte. A linguagem possui estruturas (listas, dicionários, tuplas e outras), além de *frameworks* desenvolvidos por terceiros que podem ser adicionados às aplicações.

Diferente da sua utilização como linguagem de *scripting*[1], nos últimos dez anos Python tornou-se uma das maiores linguagens com foco em análise de dados (MCKINNEY, 2018). Esta nova realidade se deve às novas melhorias e facilidades entregues pelas bibliotecas de manipulação de dados como Pandas, *Seaborn*, *Matplotlib* e outras.

Segundo o *PopularitY of Programming Language* (PYPL), a linguagem de programação Python segue sendo a mais popular da atualidade, como evidenciado na Figura 8.

[1] Sequência de instruções executadas de maneira ordenada.

O índice de popularidade é calculado através da frequência com a qual tutoriais sobre a linguagem são feitos no Google, isto demonstra o interesse do público em aprender tal linguagem e em como ela tem crescido no mercado.

Figura 8 - Ranking das linguagens mais populares em 2023

Rank	Change	Language	Share	Trend
1		Python	27.27 %	-0.5 %
2		Java	16.35 %	-1.6 %
3		JavaScript	9.52 %	+0.2 %
4		C#	6.92 %	-0.3 %
5		C/C++	6.55 %	-0.4 %
6		PHP	5.1 %	-0.5 %
7		R	4.34 %	-0.2 %
8		TypeScript	2.88 %	+0.3 %
9	↑	Swift	2.3 %	+0.1 %
10	↓	Objective-C	2.13 %	-0.1 %
11	↑↑↑	Rust	2.08 %	+0.8 %
12	↑	Go	1.95 %	+0.4 %
13	↓	Kotlin	1.77 %	+0.1 %

Fonte: (POPULARITY OF PROGRAMMING LANGUAGE, 2023)

2.5 Linguagem R

A linguagem R é utilizada para a programação estatística e possui foco na manipulação, análise e visualização de dados (FARIA; PARGA, 2020). Embutido na própria linguagem já existem alguns pacotes, do inglês *package*, que consistem em um conjunto de ferramentas ou funções extras existentes dentro da mesma que podem ser adicionados através de linhas de comando, além disso, podem ser adicionados pacotes externos oriundos do desenvolvimento de terceiros e disponibilizados na plataforma oficial da linguagem ou no GitHub dos programadores (MAYER; ZEVIANI, s.d).

Além de ser uma linguagem de código aberto, ou seja, gratuita, outra vantagem é a sua ampla base de usuários, o que resulta em uma grande quantidade de informação sobre ela disponível na internet, para auxiliar no uso e no aprendizado da linguagem. Atualmente a linguagem R tem ocupado a 7ª posição no ranking de linguagens de programação mais populares, como ilustrado na Figura 8.

2.6 Shell Script

O Shell pode ser compreendido com o intérprete de comandos do Linux que une o usuário ao sistema operacional (SO), realizando as funções de ler e interpretar os comandos e retornar uma resposta ao usuário por meio das saídas do SO. Na programação *shell*, um *script* é definido como uma lista de comandos a serem executados em sequência. A sua utilização tornou-se de extrema importância dadas a agilidade ao escrever *scripts* em poucas linhas e a sua poderosa manipulação de arquivos que variam entre a criação, edição, localização e organizados dos mesmos (JARGAS, 2008; RIBEIRO, 2021).

2.7 Linguagem Java

Java é uma linguagem de programação orientada a objeto com uma variedade de aplicações dentre elas desenvolvimento web, *mobile* e análise de dados criada pela empresa Sun Microsystems (CLARO; SOBRAL, 2008). Na sua criação foi idealizada com a intenção de ser uma linguagem de programação multiplataforma, ou seja, que pudesse ser executada em qualquer plataforma, independentemente do sistema operacional.

Para a realização desta pesquisa, o Java foi utilizado na realização das operações de leitura e escrita de arquivos. Segundo Deitel et al. (2015), a leitura de arquivos é uma etapa fundamental em muitas análises de dados, com isto, a linguagem Java fornece uma variedade de classes e interfaces para leitura de arquivos de texto, binários e outros formatos.

2.8 Samtools

O *Samtools* é um pacote de software de código aberto que é usado para manipular e analisar dados de alinhamento de sequências de DNA (LI et al., 2009). Dentre as diversas funcionalidades ofertadas pelo software, utilizou-se a que permite indexar o arquivo FASTA

com isto possibilitando a procurar por regiões específicas que neste trabalho consiste no cromossomo 14.

2.9 Bancos de Dados Genômicos

Com o andamento do Projeto do Genoma Humano, a evolução tecnológica e o surgimento de novas técnicas de sequenciamento resultaram em uma grande quantidade de dados gerados pelos laboratórios, meios convencionais de armazenamento já não eram suficientes (CRITCHLOW et al., 2000). Para solucionar tal problema, surgem os bancos de dados biológicos que visam disponibilizar de maneira confiável os dados e ferramentas de análise, em vários casos, esses bancos são de domínio público (FÉLIX, 2002).

O Projeto GENCODE armazena e disponibiliza anotações genéticas de referência para os genomas humano e de camundongo. O consórcio GENCODE é responsável por produzir, manter e melhorar essas anotações (FRANKISH et al., 2019). Atualmente o banco de dados se encontra na versão 43, os dados podem ser obtidos no link <https://www.gencodegenes.org/human/>. São disponibilizados um arquivo no formato FASTA com as sequências e um outro arquivo no formato GFF3 contendo as anotações das sequências, ou seja, a posição dos genes no genoma.

O *National Center for Biotechnology Information* (NCBI) foi encarregado de criar sistemas automatizados para armazenar e analisar conhecimentos sobre biologia molecular, bioquímica e genética. O GenBank é o banco de dados de sequências genéticas oriundo do NCBI, o mesmo junto com o DNA DataBank of Japan (DDBJ), o European Nucleotide Archive (ENA) constituem o *International Nucleotide Sequence Database Collaboration* (NCBI, 2022). A cada dois meses ocorre o lançamento do GenBank, atualmente na versão 255.0.

No entanto, como o objetivo deste trabalho é analisar as sequências de referência dos genes IGHV, IGHD e IGHJ, utilizaremos o banco de dados de referência o IMGT, que armazena as sequências e posições dos genes de imunoglobulina de humanos, camundongo, cavalo e etc. É importante observar que IMGT utiliza a anotação do NCBI, RefSeq Reference Genome Annotation from build GRCh37 (CHURCH et al., 2011).

Com a geração de grandes quantidades de dados de sequenciamento das regiões de imunoglobulina por *Next Generation Sequencing* (NGS), a acurácia e completude dos bancos de dados de referência, como o IMGT pode ser verificada (SCHEIJEN et al., 2019; YU; CEREDIG; SEOIGHE, 2017).

2.10 Trabalhos Correlatos

Foram realizadas buscas em diversas fontes de dados com a finalidade de identificar trabalhos com características semelhantes a esse trabalho, ou até mesmo que utilizaram em parte as técnicas utilizadas neste trabalho, isto para realizar comparações com outros autores que estão na mesma área de atuação.

Khatri et al. (2021), em seu artigo, abordam a importância dos alelos germinativos correspondentes à população nos *loci* de imunoglobulinas (IG) em estudos relacionados a doenças infecciosas e vacinação em populações humanas. O artigo discute como os alelos germinativos das imunoglobulinas podem variar entre populações humanas devido a diferenças genéticas e evolutivas. Essas variações podem influenciar a resposta imunológica das pessoas a doenças infecciosas e vacinações, uma vez que os anticorpos produzidos pelo sistema imunológico são determinados pelos genes das imunoglobulinas.

No trabalho apresentado por Martins et al. (2021), os autores realizam uma mineração de dados de sequenciamento de genomas de humanos, o foco deste trabalho foi identificar novas variantes de imunoglobulinas, sendo que o trabalho identificou 10,909 variantes, dessas 10,828 são variantes não reportadas anteriormente. Porém, para esta identificação foi necessário fazer uma análise das sequências do *locus*, neste ponto o trabalho dos autores possui uma grande similaridade com a proposta deste trabalho, fato que, neste trabalho será utilizado um programa disponibilizado pelo autor.

Apesar dos trabalhos citados acima terem grande semelhança com o trabalho que está sendo desenvolvido, este trabalho se caracteriza pela análise mais detalhada dos genes de cadeia pesada de imunoglobulina, onde será explorada as mudanças posicionais que ocorreram entre as versões dos genomas de referência e até mesmo a geração de uma visualização de dados que permita identificar detalhes com mais facilidade.

3 METODOLOGIA

Este capítulo descreve os métodos de pesquisa que serão adotados, bem como a natureza da pesquisa, juntamente com seu caráter, os instrumentos, os materiais e procedimentos, a população e a amostra.

3.1 Natureza da Pesquisa

De acordo com Wazlawick (2014), a pesquisa científica pode ser definida como a busca pelo aumento do conhecimento humano sobre como o mundo funciona. Ela pode ser classificada de acordo com diferentes critérios, como a sua natureza, objetivos ou procedimentos técnicos, podendo um trabalho de pesquisa se encaixar em mais de um desses tipos.

Esta pesquisa possui, quanto a sua natureza, finalidade aplicada, pois visa gerar conhecimento científico para a solução dos mais variados problemas individuais ou coletivos. Os interesses desta pesquisa são locais, uma que servirá de auxílio para os pesquisadores da área.

Podemos considerar este trabalho em relação aos seus objetivos como uma pesquisa de caráter exploratória. Tal tipo de pesquisa busca examinar um conjunto de fenômenos com o objetivo de descobrir ou aperfeiçoar ideias, uma vez que ela não necessariamente possui uma hipótese ou objetivo definido (GIL, 2002).

Além dos três critérios apontados anteriormente, para Oliveira (2021) existe ainda o critério de abordagem da pesquisa. Quanto a este tópico, esta pesquisa encaixa-se como qualitativa, pois serão realizadas análises e a interpretação dos dados.

3.2 População e Amostra

Neste trabalho, a população e amostra designadas foram os dados genômicos de humanos disponibilizados no banco de dados genômico NCBI. Os dados são advindos de diversas fontes como projetos de sequenciamento genômicos.

3.3 Instrumentos utilizados

Para o desenvolvimento deste trabalho utilizou-se computadores com acesso à internet, além do servidor Linux disponibilizado pelo IFMG - Campus São João Evangelista. O servidor utilizado dispõe das seguintes configurações: Sistema Operacional Ubuntu Linux versão 20.04.6 LTS, processador AMD Phenon II, placa de vídeo AMD Rs880, 64 bits de OS.

3.4 Métodos e Procedimentos

Nesta seção será apresentada todas as etapas executadas durante a construção desta pesquisa. A metodologia proposta busca evidenciar todos os processos da coleta de dados até os resultados.

3.4.1 *Estrutura organizacional*

Foi criada uma estrutura de pasta para facilitar a organização e localização das informações. A pasta geral nomeada neste projeto como “analise-genomica” contém todas as demais. Inicialmente temos a pasta “bin” contendo os scripts nas linguagens java e python. Em seguida, “data” com os genomas de referências versões GRCh37 e GRCh38. A terceira pasta “docs” armazena arquivos informativos ou exemplos que visam auxiliar o entendimento sobre a utilização do projeto. Última pasta “results” armazena os arquivos finais gerados a partir dos scripts executados. Ainda dentro da pasta geral, existem os arquivos *shell script* responsáveis pela execução de todos os outros arquivos armazenados na pasta bin e outro pela criação das tabelas com os respectivos genomas de referência.

3.4.2 *Coleta dos dados*

Para obter os genomas de referência GRCh37 e GRCh38 foi realizada uma busca no site do NCBI disponível no link: https://ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml. Dentre os tipos de arquivos retornados pelo site, foram realizados o download dos arquivos com extensões FNA e GFF. Um arquivo com extensão FNA armazena informações das sequências de DNA, escritos em no formato de texto FASTA. Os arquivos do tipo GFF consistem em um arquivo de texto com as anotações genômicas.

3.4.3 *Obtendo dados do cromossomo 14*

Para gerar os arquivos necessários para a análise dos dados, foi executado o arquivo principal intitulado *runMainVerifyReference.sh* com os parâmetros exigidos. O *script* inicia com a declaração das variáveis que recebem os respectivos valores passados na ordem dos parâmetros, elas possibilitam a utilização do mesmo arquivo com os diferentes genomas de referência. As variáveis foram declaradas como *path*, *filegff_gz*, *filegff*, *version*, *filefna*, *queryfasta*, *filefna_gz*, *filefna_gzt* representando respectivamente o nome da pasta com os arquivos GFF e FNA, nome do arquivo GFF compactado, nome do arquivo GFF dentro do arquivo compactado, número da versão do genoma de referência, nome do arquivo FNA compactado com as sequencias de DNA, código identificador do cromossomo, nome do arquivo FNA compactado e o nome temporário para o arquivo FNA compactado.

Com a finalidade de trabalharmos com arquivos menores, dentre as funções presentes no *pipeline*[2] desenvolvido em *shell Linux* e disponibilizado por MARTINS et al 2021, foram utilizados no arquivo GFF os comandos (*gunzip data/reference/$path/$filegff_gz*) para descompactar o genoma de referência e o programa Java (*java -jar bin/filterCromossomoFileGff.jar NC_000014 data/reference/$path/$filegff > data/reference/$path/$filegff-gff-filtrado.txt*) recebe como parâmetros de entrada o código identificador do cromossomo e o arquivo GFF descompactado, logo depois gera um arquivo de texto com os dados filtrados.

Em seguida, para obter os dados do arquivo FNA foram realizadas a descompactação via comando (*gunzip data/reference/$path/$filefna_gz*). Foi criado um programa Java (*java -jar bin/extractchr14fna.jar data/reference/$path/$filefna_gzt $queryfasta > data/reference/$path/$filefna.tmp*), que recebe como parâmetros o nome do arquivo FNA compactado temporário e o código identificador do cromossomo, depois gera um arquivo de texto temporário com os dados filtrados. O arquivo temporário retorna as informações em múltiplas linhas, a fim de solucionar tal eventualidade o programa (*java -cp bin/formatFileFasta.jar formatfilefasta.FormatFileFastaOneLine data/reference/$path/$filefna.tmp data/reference/$path/$filefna*). Após a finalização dessas operações, foram excluídos o arquivo temporário e a compactação do arquivo filtrado.

3.4.4 *Identificar a ocorrência dos genes*

[2] É uma série de etapas que são executadas para processar um dado ou uma tarefa.

Em busca de identificar possíveis ocorrências dos genes V, J e D nos arquivos GFF são necessários especificar dois parâmetros, o arquivo de texto de cada gene, além do arquivo filtrado e posteriormente foi gerado um arquivo texto com os resultados. Dando prosseguimento, foi utilizado um script *Samtools* buscando obter a sequência de nucleotídeos de referência apenas para conferência, após é realizada uma formatação através da função de indexação que possibilidade realizar buscas em regiões especificas.

3.4.5 *Obtendo a sequência de nucleotídeos de referência*

Para obter a sequência de nucleotídeos de referência IMGT para uma conferência, inicialmente é feita uma formatação das sequências de genes utilizando o *Samtools* através do comando (*java -jar bin/formatResultSeqGenesSamtools.jar*). Cada arquivo FASTA contendo a sequência dos genes V, D e J retornados anteriormente foram transformados em uma linha pelo código a seguir especificando qual a função de formatação desejada (*java -cp bin/formatFileFasta.jar formatfilefasta.FormatFileFastaOneLine*).

3.4.6 *Gerando e transformando os dados*

Para obtermos os dados referentes ao genoma de referência GRCh7 utilizou-se o seguinte comando (*./runMainVerifyReference.sh ncbi-imgt37 GRCh37_latest_genomic.gff GRCh37_latest_genomic.gff 37 GRCh37-chromossomo14.fna NC_000014.8 GRCh37_latest_genomic.fna.gz GRCh37_latest_genomic.fna*) e no GRCh38 (*./runMainVerifyReference.sh ncbi-imgt38 GRCh38_latest_genomic.gff GRCh38_latest_genomic.gff 38 GRCh38-chromossomo14.fna NC_000014.9 GRCh38_latest_genomic.fna.gz GRCh38_latest_genomic.fna*).

A fim de gerar as tabelas de comparação para uma melhor visualização dos dados, foi desenvolvido para esse projeto um programa utilizando a linguagem Python que pode ser encontrado no APÊNDICE A. O programa em questão foi desenvolvido de maneira simples e objetiva realizando operações *split* (separação), *strip* (remoção), *replace* (substituição) para que sobrem apenas as informações de extrema relevância para a análise comparativa.

Os métodos e procedimentos foram propostos com base nos objetivos específicos do trabalho, conforme descrito na Tabela 1.

Tabela 1 - Métodos e Procedimentos

Objetivos	Tarefas
Configuração do ambiente de desenvolvimento e realizar primeira explorada nos dados	Configurar o servidor e fazer o *download* dos genomas de referência GRCh37 e GRCh38.
Execução e organização de tarefas	Compilar o programa disponibilizado por (MARTINS et al., 2021) no genoma de referência GRCh37 e GRCh38. Desenvolver uma função para organização dos dados de saída.
Realizar as análises e gerar suas respectivas visualizações	Fazer uma análise dos genes de cada genoma de referência sequenciado com a tecnologia *short reads* e gerar uma visualização que permita identificar as diferenças nas sequencias de cada gene. Fazer uma análise posicional dos genes indicando: alterações na quantidade de nucleotídeos (tamanho) e mudanças nas posições, disponibilizar uma visualização que permita identificar as situações. Fazer uma análise geral dos dados e criar diferentes visualizações.

Fonte: Elaborado pelo autor, 2023

4 RESULTADOS

Com o objetivo de comparar os genes IGHV, IGHD e IGHJ dos genomas de referências versões GRCh37 e GRCh38, foi realizada uma análise das posições iniciais e finais de cada um. Podem ser identificadas nas tabelas, além das posições iniciais e finais, uma coluna "Gene" com a identificação dos genes e "T" representando a quantidade de nucleotídeos do gene (tamanho total).

4.1 Comparação das posições dos segmentos gênicos IGHV

Totalizando 45 segmentos gênicos IGHV descobertos no genoma de referência GRCh38 apenas 40 genes foram achados no genoma de referência GRCh37. Portanto, os segmentos gênicos IGHV7-4-1, IGHV3-63D, IGHV5-10-1, IGHV2-70D, IGHV1-69-2, IGHV1-69-D foram identificados na versão 38 e não identificados na versão 37. O gene IGHV3-62 foi o único presente na GRCh37 e não apareceu no genoma de referência GRCh38.

Excluindo os genes que diferem entre as versões, é possível observar que a quantidade de nucleotídeos dos genes em sua maioria permaneceu igual. Entretanto apresentaram tamanhos diferentes da versão 37 para a 38 os segmentos gênicos IGHV4-4 com 3 nucleotídeos e IGHV3-20 com 1 posição (Tabela 2).

Ao considerar os genes que estão presentes nas duas versões, em relação a posição inicial, o gene IGHV3-74 teve o menor deslocamento, sendo 107218674 na versão 37 e 106810440 na versão 38. Já no gene IGHV4-4 teve o maior deslocamento entre as versões, iniciando com 106478108 no GRCh37 e 106011922 no GRCh38.

Outro ponto a ser observado na Tabela 2 são as posições iniciais dos genomas de referências que na versão GRCh37 iniciam em 106405609 e finalizam em 107218674. Já na versão GRCh38 os genes iniciam na posição 105939754 e finalizam em 106810440. Podemos notar que ouve algum tipo de perca entre os dois genomas de referência.

Tabela 2 - Comparação das posições dos genes IGHV

	NCBI - GRCh37			NCBI - GRCh38		
Gene	P. Inicio	P. Fim	T	P. Inicio	P. Fim	T
IGHV6-1	106405609	106406056	447	105939754	105940201	447
IGHV1-2	106452669	106453106	437	105986582	105987019	437
IGHV1-3	106471244	106471681	437	106005095	106005532	437
IGHV4-4	106478108	106478539	431	106011922	106012356	434
IGHV7-4-1	0	0	0	106025145	106025581	436
IGHV2-5	106494134	106494577	443	106037902	106038345	443
IGHV3-7	106518398	106518853	455	106062149	106062604	455
IGHV3-64D	0	0	0	106088122	106088573	451
IGHV5-10-1	0	0	0	106107972	106108513	541
IGHV3-11	106573231	106573680	449	106116635	106117084	449
IGHV3-13	106586135	106586587	452	106129540	106129992	452
IGHV3-15	106610311	106610772	461	106153622	106154083	461
IGHV1-18	106641561	106641997	436	106184899	106185335	436
IGHV3-20	106667579	106668033	454	106210936	106211391	455
IGHV3-21	106691671	106692124	453	106235062	106235515	453
IGHV3-23	106725199	106725654	455	106268606	106269061	455
IGHV1-24	106733142	106733579	437	106276546	106276983	437
IGHV2-26	106757649	106758092	443	106301395	106301838	443
IGHV4-28	106780511	106780945	434	106324254	106324688	434
IGHV3-30	106791003	106791456	453	106335080	106335533	453
IGHV4-30-2	106805207	106805644	437	106349283	106349720	437
IGHV3-33	106815720	106816173	453	106359791	106360244	453
IGHV4-34	106829592	106830024	432	106373661	106374093	432
IGHV3-35	106845321	106845774	453	106389390	106389843	453
IGHV4-39	106877617	106878055	438	106421709	106422147	438
IGHV3-43	106926187	106926644	457	106470263	106470720	457
IGHV1-45	106962929	106963366	437	106506996	106507433	437
IGHV1-46	106967047	106967484	437	106511115	106511552	437
IGHV3-48	106993812	106994267	455	106537810	106538265	455
IGHV3-49	107012936	107013397	461	106556936	106557397	461
IGHV5-51	107034727	107035162	435	106578742	106579177	435
IGHV3-53	107048670	107049120	450	106592676	106593126	450
IGHV1-58	107078371	107078808	437	106622357	106622794	437
IGHV4-59	107083254	107083685	431	106627249	106627680	431
IGHV4-61	107095124	107095561	437	106639119	106639556	437
IGHV3-62	107099133	107099588	455	0	0	0
IGHV3-64	107113739	107114194	455	106657723	106658178	455
IGHV3-66	107131031	107131481	450	106675015	106675465	450
IGHV1-69	107169929	107170367	438	106714682	106715120	438
IGHV2-70D	0	0	0	106723574	106724017	443
IGHV1-69-2	0	0	0	106737110	106737547	437
IGHV1-69D	0	0	0	106762092	106762530	438
IGHV2-70	107178819	107179262	443	106770577	106771020	443
IGHV3-72	107198930	107199391	461	106790692	106791153	461
IGHV3-73	107210930	107211391	461	106802692	106803153	461
IGHV3-74	107218674	107219129	455	106810440	106810895	455

Fonte: Elaborado pelo autor, 2023

4.2 Comparação das posições dos segmentos gênicos IGHD

Foram encontrados 23 segmentos gênicos IGHD presentes nas duas versões 37 e 38, conforme ilustrada na Tabela 3. Podemos observar que a quantidade de nucleotídeos dos genes são iguais, entretanto as posições iniciais e finais se diferem entre os dois genomas de referência.

Apesar dos genes encontrados apresentarem a mesma quantidade nucleotídeos, ao considerar a posição inicial, o gene IGHD7-27 teve o menor deslocamento, sendo 106331761 na versão 37 e 105865551 na versão 38. Já no gene IGHD3-16 teve o maior deslocamento entre as versões, iniciando com 106361492 no GRCh37 e 105895634 no GRCh38.

Outro ponto a ser observado na Tabela 3 são as posições iniciais dos genomas de referências que na versão GRCh37 iniciam em 106331761 e finalizam em 106385361. Já na versão GRCh38 os genes iniciam na posição 105865551 e finalizam em 105919502, com isto, podemos notar um deslocamento de nucleotídeos entre os dois genomas de referência.

Tabela 3 - Comparação das posições dos genes IGHD

	NCBI - GRCh37			NCBI - GRCh38		
Gene	P. Inicio	P. Fim	T	P. Inicio	P. Fim	T
IGHD7-27	106331761	106331771	10	105865551	105865561	10
IGHD1-26	106346892	106346911	19	105881034	105881053	19
IGHD6-25	106347397	106347414	17	105881539	105881556	17
IGHD3-22	106351889	106351919	30	105886031	105886061	30
IGHD2-21	106354409	106354436	27	105888551	105888578	27
IGHD1-20	106357049	106357065	16	105891191	105891207	16
IGHD6-19	106357557	106357577	20	105891699	105891719	20
IGHD5-18	106359400	106359419	19	105893542	105893561	19
IGHD4-17	106360366	106360381	15	105894508	105894523	15
IGHD3-16	106361492	106361528	36	105895634	105895670	36
IGHD2-15	106363815	106363845	30	105897957	105897987	30
IGHD6-13	106367000	106367020	20	105901142	105901162	20
IGHD5-12	106368507	106368529	22	105902649	105902671	22
IGHD3-10	106370355	106370385	30	105904497	105904527	30
IGHD3-9	106370539	106370569	30	105904681	105904711	30
IGHD2-8	106373069	106373099	30	105907211	105907241	30
IGHD1-7	106375766	106375782	16	105909907	105909923	16
IGHD6-6	106376269	106376286	17	105910410	105910427	17
IGHD5-5	106378116	106378135	19	105912257	105912276	19
IGHD4-4	106379081	106379096	15	105913222	105913237	15
IGHD3-3	106380218	106380248	30	105914359	105914389	30
IGHD2-2	106382685	106382715	30	105916826	105916856	30
IGHD1-1	106385361	106385377	16	105919502	105919518	16

Fonte: Elaborado pelo autor, 2023

4.3 Comparação das posições dos segmentos gênicos IGHJ

Conforme mostrado na Tabela 4, foram retornados um total de 6 segmentos gênicos IGHJ presentes nas duas versões. Podemos identificar que a quantidade de nucleotídeos dos genes permaneceram iguais, entretanto as posições iniciais e finais se diferem entre os dois genomas de referência.

Apesar dos genes encontrados apresentarem a mesma quantidade nucleotídeos, ao considerar a posição inicial, o gene IGHJ4 teve o menor deslocamento, sendo 106330423 na versão 37 e 105863196 na versão 38. Já no gene IGHJ2 teve o maior deslocamento entre as versões, iniciando com 106331407 no GRCh37 e 105865197 no GRCh38.

Outro ponto a ser observado na Tabela 4 são as posições iniciais dos genomas de referências que na versão GRCh37 iniciam em 106329406 e finalizam em 106331668. Já na versão GRCh38 os genes iniciam na posição 105863196 e finalizam em 105865458, com isto, podemos notar que ouve algum tipo de perca entre os dois genomas de referência.

Tabela 4 - Comparação das posições dos genes IGHJ

	NCBI - GRCh37			NCBI - GRCh38		
Gene	P. Inicio	P. Fim	T	P. Inicio	P. Fim	T
IGHJ6	106329406	106329470	64	105863196	105863260	64
IGHJ5	106330022	106330074	52	105863812	105863864	52
IGHJ4	106330423	106330472	49	105864213	105864262	49
IGHJ3	106330795	106330846	51	105864585	105864636	51
IGHJ2	106331407	106331461	54	105865197	105865251	54
IGHJ1	106331615	106331668	53	105865405	105865458	53

Fonte: Elaborado pelo autor, 2023

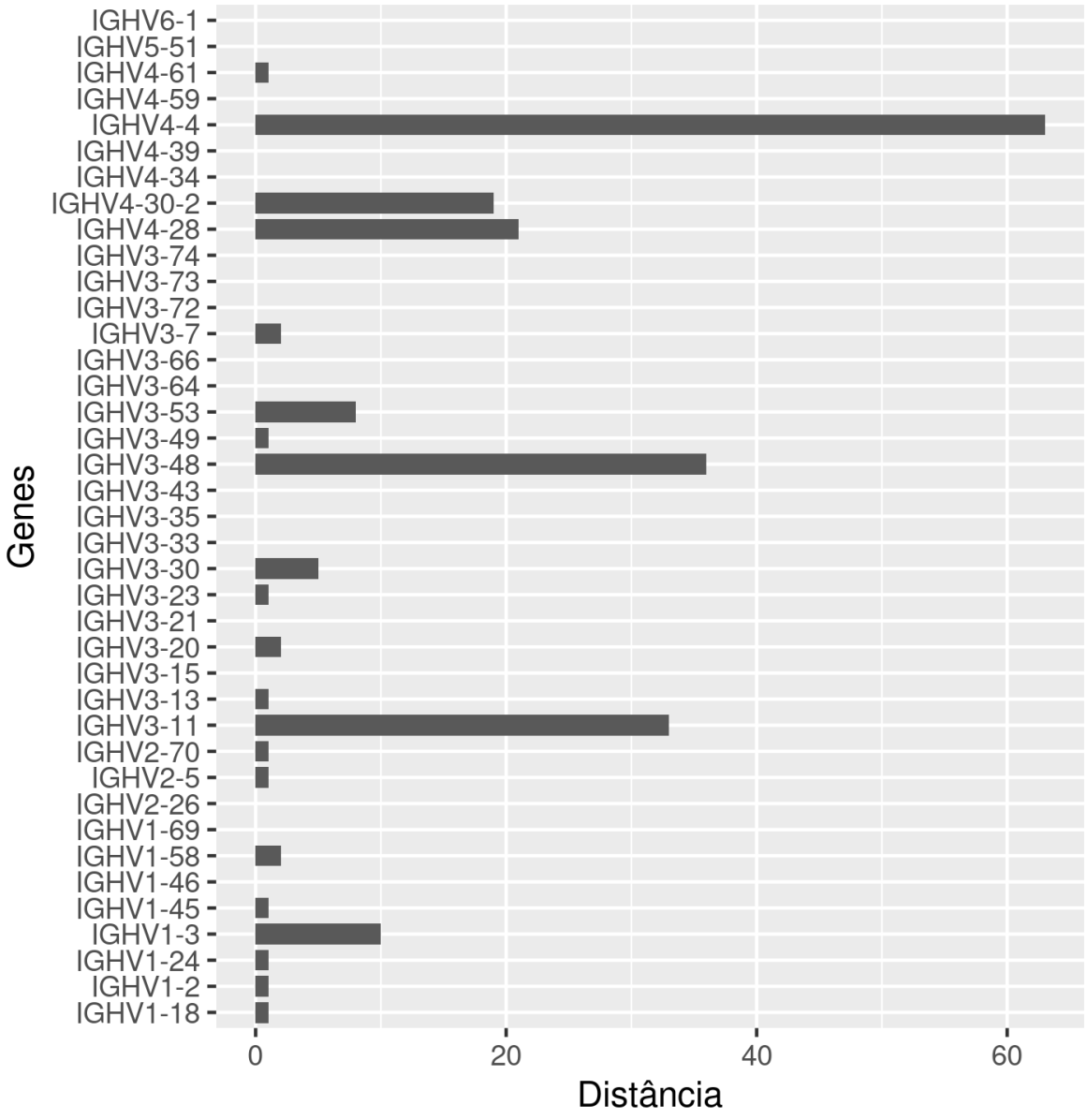
4.4 Comparação das distâncias entre os genes dos genomas de referência

A fim de verificar a distância entre as sequencias dos genes dos genomas de referência GRCh37 e GRCh38, foi desenvolvido um *script* na linguagem Python (APÊNDICE B), que dado uma sequência do gene de uma referência, realiza a busca da sequência do mesmo gene na outra referência. Os genes identificados em apenas uma versão do genoma não foram considerados na comparação e ainda a referência base na comparação foi a GRCh37 (a comparação oposta gera os mesmos valores). O *script* utiliza o algoritmo de *Levenshtein*, que

calcula o número mínimo de edições necessárias para que duas sequencias sejam consideradas iguais. A saída do *script* é retornada em formato .CSV.

De um total de 39 genes IGHV presentes nas duas versões dos genomas, foram identificados 21 genes com alterações, vale observar que os genes IGHV4-4, IGHV3-11, IGHV3-48 apresentaram as maiores diferenças, sendo respectivamente 63, 33 e 36. Essas e outras alterações podem ser observadas na Figura 9. Para os genes IGHD e IGHJ não foram identificadas diferenças entre as sequencias dos genomas de referência.

Figura 9 - Gráfico de distância entre os genes dos genomas de referência



Fonte: Elaborado pelo autor, 2023

5 CONSIDERAÇÕES FINAIS

Analisando as tabelas geradas de cada gene e comparando com as suas respectivas versões GRCh37 e GRCh38, ficou evidente as diferenças entre elas. O gene IGHV apresentou diferenças substanciais como a ausência de alguns genes e a diferença na quantidade de nucleotídeos entre as versões, já os genes IGHD e IGHJ apresentaram os mesmos genes nas duas versões do genoma de referência. Todos os tipos de gene apresentaram um deslocamento nas posições iniciais entre as duas versões.

Por fim, com relação a trabalhos futuros, recomenda-se realizar análises utilizando outras tecnologias de sequenciamento como *Long Reads* e comparar os resultados encontrados entre eles. Tais incrementos geram um grande valor aos pesquisadores que desenvolvem seus trabalhos na área de genômica de imunoglobulinas.

REFERÊNCIAS

AZEVEDO, GILSON. **Splicing alternativo e a diversidade biológica**, 2021. Disponivel em: https://blog.varsomics.com/splicing-alternativo-e-a-diversidade-biologica/

BORGES, Luiz Eduardo. **Python para desenvolvedores: aborda Python 3.3**. Novatec Editora, 2014.

CARBONNELLE P. **PYPL PopularitY of Programming Language**. 2023. Disponivel em: https://pypl.github.io/PYPL.html.

CHURCH, D. M. et al. **Modernizing reference genome assemblies**. PLoS Biology, v. 9, n. 7, p. 1–5, 2011.

CLARO, Daniela Barreiro; SOBRAL, João Bosco Mangueira. **Programação em JAVA**. Livro programando em Java 1ª edição, p. 12, 2008.

CRITCHLOW, T.; MUSICK, R.; SLEZAK, T. 2000. An Overview of Bioinformatics Research at Lawrence Livermore National Laboratory. Department of Energy by University of California Lawrence. Califórnia U.S.

Deitel, H. M., Deitel, P. J., & Deitel, P. H. (2015). **Java: Como programar** (10ª ed.). Pearson Education do Brasil.

FARIA, Pedro Duarte; PARGA, João Pedro Figueira Amorim. **INTRODUÇÃO À LINGUAGEM R**. 2020.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FÉLIX, J. M. 2002. **Genoma Funcional. Biotecnologia, Ciência & Desenvolvimento, Nº 24, janeiro a fevereiro.**

FIETTO, J. L. R.; MACIEL, T. E. F. **Sequenciando genomas**. In: MOREIRA, L.M. Ciencias genômicas: fundamentos e aplicações.Ribeirão Preto, SP: Sociedade Brasileira de Genética, 2015. v. 1, p. 27–64. Disponível em: <http://professor.pucgoias.edu.br/SiteDocente/admin/arquivosUpload/18497/material/Sequ%C3%AAnciamdo%20genomas.pdf>. Acesso em: 09 de maio de 2023.

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

GIL, Antônio Carlos. Como classificar as pesquisas. **Como elaborar projetos de pesquisa**, v. 4, n. 1, p. 44-45, 2002.

GINESTOUX, Chantal; LEFRANC, Marie-Paule. **Chromosomal localization: human (Homo sapiens) IGH**, 2001. Disponível em:

https://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=chromosomes&species=human&group=IGH. Acesso em 14 de maio de 2023.

GIUDICELLI, V.; CHAUME, D.; LEFRANC, M. **IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes Nucl**. Acids Res. Disponível em: https://www.imgt.org/genedb/resultPage.action;jsessionid=0855D32776F439EA50171BF28B41FB13?gene.id.species=Homo+sapiens&molComponent=IG&geneTypeLike=variable&allele.fcode=functional&cloneName=&locusLike=IGH&mainLocusLike=IGH+locus&cosLocusLike=any&groupLike=IGHV&subgroup=-1&geneLike=&selection=any. Acesso em 14 de maio de 2023.

GÓES, Andréa Carla de Souza; OLIVEIRA, Bruno Vinicius Ximenes de. **Projeto Genoma Humano: um retrato da construção do conhecimento científico sob a ótica da revista Ciência Hoje**. Ciência & Educação (Bauru), v. 20, p. 561-577, 2014.

GRIFFITHS, A.J.F.; GELBART, W.M.; MILLER, J.H.; LEWONTIN, R.C. **Genética Moderna**. Rio de Janeiro: Guanabara Koogan, 2001, 589p.

HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001.

HAN, J; KAMBER, M. **Data Mining – Concepts and Techniques**. Morgan Kaufmann Publishers, Inc, 2001.

JARGAS, Aurélio Marinho. **Shell script profissional**. Novatec Editora, 2008.

KUMAR, Kishore R.; COWLEY, Mark J.; DAVIS, Ryan L. **Next-generation sequencing and emerging technologie**s. In: Seminars in thrombosis and hemostasis. Thieme Medical Publishers, 2019. p. 661-673.

LEFRANC, Marie-Paule; LEFRANC, Gérard. **Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions**. Biomedicines, v. 8, n. 9, p. 319, 2020.

LESK, Arthur M. **Introdução à bioinformática**. Artmed, 2008.

LEFRANC, Marie-Paule; LEFRANC, Gérard. **The immunoglobulin factsbook. Academic press**, 2001.

LI, Heng et al. **The sequence alignment/map format and SAMtools**. bioinformatics, v. 25, n. 16, p. 2078-2079, 2009.

MARTINS, Fabio R. et al. **Discovery of 10,828 new putative human immunoglobulin heavy chain IGHV variants**. bioRxiv, p. 2021.01. 15.426262, 2021.

MATSUDA, Fumihiko et al. **The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus**. The Journal of experimental medicine, v. 188, n. 11, p. 2151-2162, 1998.

MAYER, F; ZEVIANI, W. **Pacotes R**. [s.d]. Disponivel em: http://cursos.leg.ufpr.br/prr/capPacR.html#motiva%C3%A7%C3%A3o

MCKINNEY, Wes. **Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython**. Novatec Editora, 2018.

MURPHY, Kenneth. **Imunobiologia de Janeway-8**. Artmed Editora, 2014.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Genbank Overview**, 2022. Disponível em: https://www.ncbi.nlm.nih.gov/genbank/. Acesso em: 24 de maio de 2023.

OLIVEIRA, Sofia Cisneiros Alves de. **Metodologia científica: tipos de pesquisa**. [S. l.], 4 nov. 2021. Disponível em: https://www.sanarmed.com/metodologia-cientifica-tipos-de-pesquisa-colunistas. Acesso em: 24 de maio de 2023.

RIBEIRO, U. **Shell do Linux para Iniciantes - Certificação Linux**, 2021. Disponível em: <https://www.certificacaolinux.com.br/shell-do-linux-para-iniciantes/#:~:text=O%20que%20%C3%A9%20o%20Shell>. Acesso em: 18 set. 2023.

SANTOS, Vanessa Sardinha dos. **Genes**; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/biologia/genes.htm. Acesso em 23 de maio de 2023.

SANTOS, Vanessa Sardinha dos. **Nucleotídeo**; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/biologia/nucleotideo.htm. Acesso em: 18 de abril de 2023.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. Elsevier, 2014.

WETTERSTRAND, KA. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)**. Disponível em: www.genome.gov/sequencingcostsdata. Acesso em 23 de maio de 2023.

ZAHA, Arnaldo; FERREIRA, Henrique Bunselmeyer; PASSAGLIA, Luciane MP. **Biologia Molecular Básica-5**. Artmed Editora, 2014.

APÊNDICE A – CÓDIGO DE CRIAÇÃO DAS TABELAS DOS GENES

```
1. import sys
2.
3. def processFile(pathFile):
4.
5.     fileIn = open(pathFile, "r")
6.
7.     #>IGHD7-27-NC_000014.8:106331761-106331771
8.     #CTAACTGGGGA
9.     for linha in fileIn:
10.
11.         if ">" in linha:
12.             vec = linha.split(":")
13.             vec2 = vec[1].split("-")
14.             gene = vec[0].strip().replace(">", "")
15.             gene = gene.split("-NC")[0]
16.             print(gene + ";" + vec2[0].strip() + ";" + vec2[1].strip())
17.              #print(linha.replace("\n", ""))
18.
19.     fileIn.close()
20.
21. #python3 createTableOfGene.py pathFileDataOfGenome > nameFileOutput
22. if __name__ == "__main__":
23.
24.     #pathFile = "/home/fabio/Desktop/tmp/seqGenesDRef-NCBI-IMGT-final-fmt-CR.txt"
25.     pathFile = sys.argv[1]
26.     processFile(pathFile)
```

APÊNDICE B – CÓDIGO DE COMPARAÇÃO DAS DISNTÂNCIAS ENTRE OS GENES DOS GENOMAS DE REFERÊNCIA

```
1. from Levenshtein import distance
2. import sys
3.
4. def verifyDiffSeq(file1, file2):
5.
6.  fileIn1 = open(file1, "r")
7.
8.  for linha in fileIn1:
9.
10.   if ">" in linha:
11.    linhaTmp = linha
12.    linhaComp = fileIn1.readline()
13.
14.    seqComp = getSeqComp(file2, linhaTmp.split("-NC")[0])
15.
16.    if len(linhaComp) > 0 and len(seqComp) > 0:
17.     a = distance(linhaComp.strip(), seqComp.strip())
18.
19.     print(linhaTmp.strip().split("-NC")[0] + ";" + str(a))
20.
21.  fileIn1.close()
22.
23. #>IGHV1-18-NC
24. def getSeqComp(pathFile2, nameGene):
25.  fileIn2 = open(pathFile2, "r")
26.  ret = ""
27.  for linhaFind in fileIn2:
28.   if ">" in linhaFind:
29.    comp = linhaFind.split("-NC")[0]
30.    if comp.strip() == nameGene.strip():
31.     #print(linhaFind.strip())
32.     ret = fileIn2.readline()
33.     break
34.
35.  fileIn2.close()
36.  return ret
37.
38. if __name__ == "__main__":
39.
40.    file1 = sys.argv[1]
41.    file2 = sys.argv[2]
42.
43.    verifyDiffSeq(file1, file2)
44.
```